Bayesian Analysis (2011) **6**, Number 2, pp. 329–352

Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination

Christopher Yau* and Chris Holmes†

Abstract. We propose a hierarchical Bayesian nonparametric mixture model for clustering when some of the covariates are assumed to be of varying relevance to the clustering problem. This can be thought of as an issue in variable selection for unsupervised learning. We demonstrate that by defining a hierarchical population based nonparametric prior on the cluster locations scaled by the inverse covariance matrices of the likelihood we arrive at a 'sparsity prior' representation which admits a conditionally conjugate prior. This allows us to perform full Gibbs sampling to obtain posterior distributions over parameters of interest including an explicit measure of each covariate's relevance and a distribution over the number of potential clusters present in the data. This also allows for individual cluster specific variable selection. We demonstrate improved inference on a number of canonical problems.

Keywords: Bayesian mixture models, Bayesian nonparametric priors, variable selection, unsupervised learning

1 Introduction

As information becomes cheaper to capture there is increased interest in statistical approaches that can uncover structure in high-dimensional data sets. Such exploratory discovery driven analyses are common place in important fields such as genomics and e-commerce. One of the most widely used analysis tools in these situations is provided by clustering algorithms which seek to group together similar observations, where "similarity" is defined by some metric. Hierarchical, partition based and model based clustering are the three most common clustering methods, for an overview see Hastie et al. (2009). The general trend of decreasing cost associated with measurements means that often studies are not targeted at specific variables known *a priori* to be relevant to group discrimination and hence many of the gathered variables are likely to be irrelevant from a clustering perspective.

In this paper we investigate fully probabilistic Bayesian nonparametric mixture model priors designed for such situations. Under our prior we model the relevance of each covariate by a scale-standardised distribution on cluster locations in each covariate. We show that by placing a prior which encourages shrinkage on the locations towards a

*Department of Statistics, University of Oxford, Oxford, U.K., mailto:yau@stats.ox.ac.uk
†Department of Statistics & The Oxford-Man Institute, University of Oxford, Oxford, U.K., http://www.stats.ox.ac.uk/~cholmes/

 DOI:10.1214/11-BA612

common mean we can effectively prune out irrelevant dimensions and characterise the relative relevance of those remaining.

Variable selection has a substantial literature within supervised learning, for example see Hastie et al. (2009) and Claeskens and Hjort (2008) for an overview of various approaches. An important special case of variable selection in the Bayesian literature is provided by the so called Automatic Relevance Determination (ARD) first described in MacKay (1995) and Neal (1996) and popularised in nonlinear pattern recognition by Tipping (2001). ARD priors provide a continuous measure of the relevance of each variable in a supervised learning problem and it is in this spirit that we proceed. Somewhat surprisingly in unsupervised learning there has been far fewer papers on variable selection perhaps due to the objective measure being harder to quantify. However notable work in this field exists and we mention Friedman and Meulman (2004); Law et al. (2004) and Hoff (2006). In the machine learning community, Pan and Shen (2007) investigated Lasso type penalties in a maximum penalised likelihood approach whereas Dy and Broadly (2004) investigated a within to between cluster variance approach using a stepwise subset selection algorithm. In the Bayesian literature, we note the work of Kim et al. (2006) and Tadesse et al. (2005) who consider both parametric and nonparametric Bayesian mixture models with a hard variable selection point mass prior. Raftery and Dean (2006) consider relevance via a stepwise variable selection approach choosing both the number of variables, the number of clusters and the form of the covariance matrices according to the Bayesian Information Criterion; with extensions and comparisons provided in Maugis et al. (2009). In this paper we have sought to generalise some of these approaches to a fully Bayesian nonparametric approach with implicit sparsity priors on the relevant variables which allows for cluster specific variable selection. While necessitating Markov chain Monte Carlo our approach provides a continuous (probability) measure on covariate relevance.

In the next section we review model based clustering and the hierarchical prior we recommend. We then show how this can be extended to situations when the number of clusters is unknown. In Section 4 we present results on a number of simulated and real data sets. Section 5 provides a brief conclusion. All of the code used is written in MATLAB and is available at: https://sites.google.com/site/mixlasso/.

2 Model based clustering

We consider data $X \in \mathbb{R}^p$ and a data set $\{x_i\}_{i=1}^n$ of n observations which we suppose arises from a mixture model with $K \in \mathbb{N}$ components

$$\pi(x_i) = \sum_{k=1}^{K} w_k f(x_i; \theta_k),$$

where w_k are non-negative mixing weights, $\sum_k w_k = 1$, and $f(\cdot)$ are mixture densities with parameters θ_k. For example, for the Gaussian case we have $\theta_k = \{\mu_k, \Sigma_k\}$ for location parameters $\mu_k = \{\mu_{k1}, \ldots, \mu_{kp}\}$ and p-by-p variance-covariance matrix Σ_k. Throughout this paper we will restrict ourselves to the Gaussian mixture case though

the extension to, say Student mixtures, is straightforward using scaled-mixture representations (Andrews and Mallows 1974).

It is useful to define a latent indicator variable z_i for each datum that conditionally assigns x_i to one and only one component

$$\pi(x_i|z_i = k) = f(x_i; \theta_k)$$

and clearly $\pi(z_i = k) = w_k$. Bayesian modelling proceeds by characterizing uncertainty in parameter values via prior probability distributions. In our case this involves priors over the mixture weights, cluster locations, cluster covariances and perhaps also the number of clusters; whereas in non-Bayesian methods these quantities are estimated according to some loss function. Inference for mixture models has been extensively described in Titterington et al. (1985) and more recently in the Bayesian context by Frühwirth-Schnatter (2006) as well with extensions by Fraley and Raftery (2002); Richardson and Green (1997) and Jasra et al. (2005).

In this report we consider the case where some of the dimensions of $X \in \mathbb{R}^p$ are not relevant to the clustering problem, that is, where the mixtures heavily "overlap" in certain co-ordinate directions. It will help to illustrate our approach by first considering the simplest situation with $K = 2$ Gaussian mixtures and a common diagonal covariance matrix.

2.1 A simple example for K-component mixtures

To motivate our approach we shall first consider a two-component mixture model and assume also that the mixture components share a common diagonal covariance matrix $\Sigma_1 = \Sigma_2 = \Sigma = diag\{\sigma_1^2, \ldots, \sigma_p^2\}$; the more general case is considered later on but it is instructive to begin with this. Recall that our motivation is to explore variable selection models, or rather measures of covariate relevance. With this aim it is interesting to compare with the analogous problem of variable selection in discriminant analysis within supervised learning which involves models of $\pi(z_i|x_i)$ with both z_i and K known and data sets $\{z_i, x_i\}_{i=1}^n$. In one sense, model-based clustering is simply discriminant analysis whereby all of the z_i group labels are missing. With this in mind, as noted by Fisher (1936) the natural unit of measurement for separability of clusters is given by the standardised distance $((\mu_{1j} - \mu_{2j})/\sigma_j)^2$, which is a function of the L_2 distance between cluster centres standardised by the variance of the measurements around each cluster. From this criterion Fisher went on to derive what we now know of as Linear Discriminant Analysis (LDA), one of the most popular pattern recognition models used today. Taking Fisher's standpoint we can consider $(\mu_{1j} - \mu_{2j})/\sqrt{\sigma_j^2}$ as a *primitive* on which we wish to express prior beliefs. That is, for covariates which are irrelevant to the clustering, i.e. directions which are uninformative for $\pi(z_i|x_i)$, we expect $(\mu_{1j} - \mu_{2j})/\sqrt{\sigma_j^2}$ to be small (in absolute value) and vice versa for influential covariates.

So we aim to provide a prior which encodes these beliefs, which lead us to the

following K-component hierarchical Bayesian Gaussian mixture model,

$$x_i|z_i = k, \mu, \Phi \sim \mathrm{N}(\mu_k, \Phi), i = 1, \ldots, n, \tag{1}$$

$$\Phi = diag(\sigma_1^2, \ldots, \sigma_p^2), \tag{2}$$

$$\mu_{1,j}, \ldots, \mu_{K,j}|\lambda_j, \Phi \sim \prod_{k=1}^{K} \mathrm{N}(\theta_j, \sigma_j^2\lambda_j), j = 1, \ldots, p, \tag{3}$$

$$\theta_j \sim \mathrm{N}(0, \infty), j = 1, \ldots, p, \tag{4}$$

$$\lambda_j \sim \mathrm{Ga}(c, d), j = 1, \ldots, p, \tag{5}$$

$$\sigma_j^2 \sim \mathrm{IG}(r, s), j = 1, \ldots, p, \tag{6}$$

$$d \sim \mathrm{Ga}(g, h), \tag{7}$$

where $\mathrm{N}(\cdot)$, $\mathrm{Ga}(\cdot)$ and $\mathrm{IG}(\cdot)$ denote the Normal, Gamma and Inverse-Gamma distributions respectively. Note that we have not yet defined the prior distribution on the mixture weights w - we shall discuss this later. The important point is in (3) where we see that the cluster locations in each dimension, $\{\mu_{1,j}, \ldots, \mu_{K,j}\}$, follow an exchangeable distribution which shrinks towards a common mean θ_j in each direction. It can then be seen that this hierarchy induces the following conditional distributions on the scale-standardised locations in each covariate

$$(\mu_{sj} - \mu_{tj})/\sqrt{2\sigma_j^2} \sim \mathrm{N}(0, \lambda_j)$$

and moreover by taking, $\lambda_j \sim \mathrm{Ga}(c = 1, d = \eta)$ this then induces the marginal density

$$(\mu_{sj} - \mu_{tj})/\sqrt{2\sigma_j^2} \sim \mathrm{DE}(\eta, \theta_j)$$

where $\mathrm{DE}(\eta, \theta)$ denotes the double-exponential distribution with shape parameter η centred at θ_j. It is well known that the double exponential density, due to its relatively high kurtosis, has the tendency to shrink small values to zero while performing little shrinkage on larger values. There is also a direct correspondence between the Bayesian MAP estimate using the double-exponential prior and the LASSO penalized likelihood approach, which is known to induce sparse solutions. This setting of independent priors on the prior variances of lower-level parameters is also reminiscent of the ARD methods used in regression.

Hence, the above hierarchical model implies a marginal prior on the primitive $(\mu_{sj} - \mu_{tj})/\sqrt{2\sigma_j^2}$ which shrinks the cluster means towards a common location in a way that encourages sparsity. This was exactly our intention. Such a model should then have the tendency to prune out "irrelevant" variables by shrinking the group means $\{\mu_{1j}, \ldots, \mu_{Kj}\}$ towards a common value θ_j. Moreover by examining the marginal posterior distribution of $\pi(\lambda_j|x)$ we should gain an understanding of the relative relevance of each covariate.

An important point is that the shrinkage parameters λ_j's do not enter into the likelihood of the data. That is, x is conditionally independent of $\{\lambda_1, \ldots, \lambda_p\}$ given $\{\mu, \Phi\}$. This feature allows for automatic relevance determination in the covariates as shown next.

2.2 Relationship to existing methods

One key difference between our proposal and previous reported Bayesian methods is that we do not specify the variable selection model within the likelihood. For instance, in Tadesse et al. (2005) they define the model via,

$$\pi(x_i|z_i = k) = \mathrm{N}(x_{iI(\gamma)}|\mu_{kI(\gamma)}, \Sigma_{I(\gamma)})\mathrm{N}(x_{iI(\gamma^c)}|\mu_{kI(\gamma^c)}, \Sigma_{I(\gamma^c)})$$

where $I(\gamma)$ indexes those variables currently selected in the model and $I(\gamma^c)$ the complementary set which are not included. This model implies independence between the relevant and irrelevant variables, which is perhaps often difficult to justify.

In the approach of Raftery and Dean (2006) they factorise the joint density as

$$\pi(x_i|z_i = k) = \pi(x_{iI(\gamma)}|z_i)\pi(x_{iI(\gamma^c)}|x_{iI(\gamma)})\pi(x_{iI(\gamma^{cc})}|x_{iI(\gamma^c)}, x_{iI(\gamma)}) \tag{8}$$

where γ^c are a set of potential variables to possibly be included within the model and γ^{cc} are variables excluded from consideration of inclusion. The advantage of (8) is that you no longer have to assume independence between the variables. However one is then left with the task of modelling each factor of the joint distribution on the rhs of (8).

Here we leave the likelihood unaltered but consider the posterior marginal distribution of $\pi(\lambda_j|x)$ as a measure on the relevance of the jth covariate. In fact it quantifies the distribution of the variance parameter of the scale-standardised distribution of cluster centres

$$(\mu_{ij} - \mu_{kj})/\sqrt{2\sigma_j^2} \sim N(0, \lambda_j),$$

which we believe is an intuitive measure of the relevance. This is akin to a Bayesian ANOVA on the standardised cluster locations.

2.3 Non-diagonal, non-identical covariance matrices

The assumption of a common, diagonal covariance structure limits the practicality of the model. In order to incorporate cluster-specific, full covariance structure, we can incorporate an additional layer into the hierarchy as follows,

$$x_i|z_i = k, \mu, \Sigma \sim \mathrm{N}(\mu_k, \Sigma_k), i = 1, \ldots, n, \tag{9}$$

$$\Sigma_k \sim \mathrm{IW}(\gamma, \Phi), k = 1, \ldots, K, \tag{10}$$

$$\Phi = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}, \tag{11}$$

where $\mathrm{IW}(\cdot)$ denotes the Inverse-Wishart distribution and other layers in the hierarchy are similar to those in Section 2.2. The key feature of this modification is that the primitive on which we are expressing prior beliefs is preserved, $(\mu_{sj} - \mu_{tj})/\sqrt{2\sigma_j^2}$, where the scaling is by σ_j^2 which now describes an average variance across clusters in that variable. However, the likelihood now involves a full cluster-specific, covariance matrix, whose prior distribution is an Inverse-Wishart distribution with a scale matrix given by the variable-specific variances $\{\sigma_1^2, \ldots, \sigma_p^2\}$.

2.4 Inference on the number of clusters

We have so far assumed that the number of distinct clusters and hence the number components in the mixture model (K) is known. This is not typically the case and the determination of the number of clusters is often also of interest. In Raftery and Dean (2006) and Maugis et al. (2009) this problem is tackled as a model selection problem using the Bayesian Information Criterion (BIC) as a measure to choose between finite mixture models involving differing number of components. An alternative approach to fitting mixture models with an unknown number of components uses stochastic processes that permit the creation and destruction of mixture components such as the reversible-jump Markov Chain Monte Carlo by Richardson and Green (1997) or the continuous-time Markov birth-death processes by Stephens (2000). Our chosen method is closer to the Bayesian nonparametric approach of Tadesse et al. (2005) who define mixture distributions with a countably infinite number of components implemented using a Dirichlet process prior (Antoniak 1974; Ferguson 1973) on the mixture proportions.

A number of Monte Carlo samplers are available for efficient posterior simulation with Dirichlet Process Mixture models, which can be roughly divided into two categories: marginal methods in which the Dirichlet Process is analytically integrated out (Escobar 1994; Escobar and West 1995; MacEachern 1998; Neal 2000; Jain and Neal 2000; Green and Richardson 2001) and conditional methods utilising the stick-breaking representation due to Sethuraman (1994) in which the parameters of the Dirichlet Process prior are explicitly imputed (Ishwaran and James 2001; Walker 2007; Papaspiliopoulos and Roberts 2008; Kalli et al. 2011).

We impose the following hierarchical structure on the latent allocation variables z and mixture weights w which utilises the auxiliary variable construction proposed by Walker (2007)

$$\pi(z_i = k|w, u_i) \propto \mathbb{I}(k : w_k > u_i), i = 1, \ldots, n, \tag{12}$$

$$u_i \sim \mathrm{U}(0, 1), \tag{13}$$

$$v_k \sim \mathrm{Be}(1, \alpha), \tag{14}$$

$$w_1 = v_1, w_k = v_1 \prod_{j=1}^{k-1} (1 - v_j), k > 1, \tag{15}$$

$$\alpha \sim \mathrm{Ga}(a, b). \tag{16}$$

where $\mathrm{Be}(\cdot)$ and $\mathrm{U}(\cdot)$ denote the Beta and Uniform distributions respectively. Equations (14-16) specify the Dirichlet Process prior on the mixture weights w with concentration parameter α. Walker (2007) noted that the density

$$f_{w,\mu,\Sigma}(x) = \sum_{k=1}^{\infty} w_k \mathrm{N}(x; \mu_k, \Sigma_k),$$

can be written as the marginal of the joint density,

$$f_{w,\mu,\Sigma}(x, u) = \sum_{k=1}^{\infty} \mathbb{I}(w_k > u) \mathrm{N}(x; \mu_k, \Sigma_k).$$

Crucially, given the auxiliary variable u, the set $A_u = \{k : w_k > u\}$ is finite, hence the likelihood can be written as a summation over a finite number of terms

$$f_{w,\mu,\Sigma}(x|u) = \frac{1}{\sum_{k\in A_u} w_k} \sum_{k\in A_u} \mathrm{N}(x; \mu_k, \Sigma_k).$$

Consequently, for posterior updating of the latent allocation variable z_i, given the auxiliary variable u_i it is only necessary to consider a finite number of allocations rather than the infinite number that would be required. Walker (2007) showed that in order to update all allocation variables, it is only necessary to simulate K^* mixture components (v_k, μ_k, Σ_k) such that $\sum_{k=1}^{K^*} w_k > 1 - \min_i\{u_i\}$ (simulating any additional component parameters from the prior if necessary).

2.5 Conditional distributions and Gibbs sampling

One immediate advantage of this particular hierarchical model is that the conditional distributions all have standard form and hence allow for a full Gibbs sampling strategy. Fairly simple inspection of the hierarchical model reveals the following conditional distributions:

$$\pi(z_i = k|x_i, u_i, w, \mu, \Sigma) = \frac{\mathrm{N}(x_i; \mu_k, \Sigma_k)}{\sum_{j:w_j>u_i} \mathrm{N}(x_i; \mu_j, \Sigma_j)}, \tag{17}$$

$$u_i|z_i = k, w \sim \mathrm{U}(0, w_k), \tag{18}$$

$$v_k|\alpha, z \sim \mathrm{Be}\left(1 + n_k, \alpha + n - \sum_{j=1}^{k-1} n_j\right), \tag{19}$$

$$\mu_k|\cdot \sim \mathrm{N}\left(V_k(\Sigma_k^{-1}\overline{x}_k + \Phi^{-1}\Lambda^{-1}\theta), V_k\right), k = 1, \ldots, K, \tag{20}$$

$$\Sigma_k|\cdot \sim \mathrm{IW}\left(\gamma + n_k, \Phi + \sum_{i:z_i=k} (x_i - \mu_k)(x_i - \mu_k)'\right), k = 1, \ldots, K, \tag{21}$$

where $n_k = \sum_{i=1}^n \mathbb{I}(z_i = k)$, $\overline{x}_k = \sum_{i:z_i=k} x_i$ and $V_k = (n_k\Sigma_k^{-1} + \Phi^{-1}\Lambda^{-1})^{-1}$.

$$\lambda_j|\cdot \sim \mathrm{GIG}\left(c - \frac{1}{2}K', \frac{1}{2\sigma_j^2}\sum_{k:n_k>0}(\mu_{jk} - \theta_j)^2, 2d\right), \tag{22}$$

$$\sigma_j^2|\cdot \sim \mathrm{GIG}\left(r + \frac{1}{2}K'(\gamma - 1), 2t + \frac{1}{\lambda_j}\sum_{k:n_k>0}(\mu_{jk} - \theta_j)^2, \sum_{k:n_k>0}[\Sigma_k^{-1}]_{jj}\right), \tag{23}$$

$$d|\cdot \sim \mathrm{Ga}\left(g + p, h + \sum_{j=1}^p \lambda_j\right), \tag{24}$$

where $K' = \sum_{k=1}^{K} \mathbb{I}(n_k > 0)$ is the number of mixture components to which data has been assigned and GIG$(\cdot)$ is the Generalized Inverse Gaussian distribution of the form,

$$\text{GIG}(x; a, b, c) \propto x^{a-1} \exp\left(-\frac{1}{2}\left[\frac{b}{x} + cx\right]\right).$$

We update the concentration parameter α using the mixtures of Gamma method of Escobar and West (1995).

Finally, we introduce two label-switching moves, as given in Papaspiliopoulos and Roberts (2008), since the augmentation of w in our model makes the components in the infinite mixture weakly identifiable. The first label-switching move proposes to switch the labels of two randomly chosen components i and j with acceptance probability $\min(1, (w_j/w_i)^{n_i - n_j})$ whilst the second move proposes to swap the labels of two adjacent components j and $j+1$ (and their corresponding stick-breaking variables v_j and v_{j+1}) with acceptance probability $\min(1, (1-v_{j+1})^{n_j}/(1-v_j)^{n_{j+1}})$. These moves are complementary with the former having high acceptance probability of switching labels of components with similar weights whilst the latter has high probability of switching very unequal clusters.

Note, that in the updates of the parameters (λ, σ^2, d), a summation over all mixture components is required. In our model, there are a countably infinite number of components, however, finite computation is possible since we only need to sum over the mixture components that are associated with data as the effect of the non-realised mixture components is integrated out. For instance,

$$\pi(\lambda_j, \{\mu_{j,k}\}_{k \notin A_k} | d, \sigma_j^2, \{\mu_{j,k}\}_{k \in A_k}) \propto \pi(\lambda_j | d) \prod_{k \in A_k} f(\mu_{jk}; \theta_j, \sigma_j^2 \lambda_j) \prod_{k \notin A_k} f(\mu_{jk}; \theta_j, \sigma_j^2 \lambda_j) \tag{25}$$

where the set $A_k = \{k : n_k > 0\}$ and $f(x; \mu, \sigma^2)$ is the probability density function of the Normal distribution of the random variable x with mean μ and variance σ^2. Integrating out the μ's from mixture components to which no data is associated,

$$\pi(\lambda_j | d, \sigma_j^2, \{\mu_{j,k}\}_{k \in A_k}) \propto \pi(\lambda_j | d) \prod_{k \in A_k} f(\mu_{jk}; \theta_j, \sigma_j^2 \lambda_j) \prod_{k \notin A_k} \int f(\mu_{jk}; \theta_j, \sigma_j^2 \lambda_j) d\mu_{jk}, \tag{26}$$

since $\int f(\mu_{jk}; \theta_j, \sigma_j^2 \lambda_j) d\mu_{jk} = 1$ then

$$\pi(\lambda_j | d, \sigma_j^2, \{\mu_{j,k}\}_{k \in A_k}) \propto \pi(\lambda_j | d) \prod_{k \in A_k} f(\mu_{jk}; \theta_j, \sigma_j^2 \lambda_j). \tag{27}$$

2.6 Cluster-specific variable relevance determination

An extension of our model may be obtained by allowing each cluster to possess its own set of variances λ_{kj} for the jth variable's relevance in the kth cluster. This construction

allows us to identify if a particular variable segregates one cluster from the others; similar motivation for cluster specific variable selection can be found in Friedman and Meulman (2004). The hierarchical model is specified as follows:

$$\mu_{kj}|\lambda_{kj}, \Phi \sim \prod_{k=1}^{K} \mathrm{N}(\theta_j, \sigma_j^2 \lambda_{kj}), j = 1, \ldots, p, \tag{28}$$

$$\lambda_{kj} \sim \mathrm{Ga}(1, d_j), j = 1, \ldots, p, \tag{29}$$

$$d_j \sim \mathrm{IG}(g, h), j = 1, \ldots, p. \tag{30}$$

This leads to a prior distribution on the primitive $(\mu_{sj} - \mu_{tj})/\sqrt{2\sigma_j^2}$ of the form $\mathrm{N}(\theta_j, \lambda_{sj} + \lambda_{tj})$.

3 Examples

In all the example we applied our method to the datasets obtaining 11,000 MCMC samples and discarding the first 1,000 as burn-in and thinning by every 10 samples. For the larger simulated datasets, average run-times for each data set was approximately 45-60 minutes per dataset using a MATLAB-based implementation. Run-times vary due to variations in the number of realised mixture components during the MCMC run. We initialised using label assignments from a 50-component Gaussian Mixture Model fitted using a standard maximum likelihood-based expectation-maximization approach.

3.1 Simulation Study

Data. Eighty simulated data sets were generated based on the seven scenarios generated by Maugis et al. (2009) and an eighth scenario which is a replica of scenario 6 but with non-identical noise covariances. Each dataset consists of 2,000 data points $x_i^{(1,2)}$ from a mixture of four Gaussian distributions $\mathrm{N}(\mu_k, I_2)$ with $\mu_1 = (-2, -2)', \mu_2 = (-2, 2)', \mu_3 = -\mu_2$ and $\mu_4 = -\mu_1$ with mixture proportions $w = (0.3, 0.2, 0.3, 0.2)$. Eight irrelevant variables were appended and simulated according to $x_i^{(3,\ldots,10)} = \beta' x_i^{(1,2)} + \epsilon_i$, $\epsilon_i \sim \mathrm{N}(0, \Sigma_{z_i})$ (see Table 1 for details).

Results. Figure 1 shows that, for each dataset, our method was able to identify the number of clusters in the data by assigning non-negligible mixture weights to four components. Posterior means for the mixture weights for each of the eighty datasets were close to the actual simulation values of (0.3, 0.3, 0.2, 0.2). The relevant clustering variables were also identified in all scenarios with these having significantly larger values of λ than the irrelevant non-clustering variables as anticipated.

We note that although our method does not explicitly model the relationships between variables it is possible to distinguish between *independent* clustering variables that are directly related to the clustering and *related* clustering variables which are independent of the clustering given the *independent* clustering variables. This is achieved by the examining the noise variances σ^2 associated with each clustering variable. In

Scenario	Regression Coefficients	Covariance
1.	$\beta = 0_8$	$\Sigma = I_8$
2.	$\beta = 0_8$	$\Sigma = 0.5I_8$
3.	$\beta = ([2,0]', 0_7)$	$\Sigma = I_8$
4.	$\beta = ([0.5,0]', [0,1]', 0_6)$	$\Sigma = I_8$
5.	$\beta = (\beta_1, 0_4)$	$\Sigma = \text{diag}(I_2, 0.5I_2, I_4)$
6.	$\beta = (\beta_1, \beta_2, 0_2)$	$\Sigma = \text{diag}(I_2, 0.5I_4, I_2)$
7.	$\beta = (\beta_1, \beta_2, \beta_3)$	$\Sigma = \text{diag}(I_2, 0.5I_4, I_2)$
8.	$\beta = (\beta_1, \beta_2, \beta_3)$	$\Sigma_1 = \text{diag}(I_2, 0.5I_4, I_2)$ $\Sigma_2 = \text{diag}(I_2, 0.15I_4, I_2)$ $\Sigma_3 = \text{diag}(I_2, 0.35I_4, I_2)$ $\Sigma_4 = \text{diag}(I_2, 0.25I_4, I_2)$

Table 1: Simulation models. Scenarios 1-7 are derived from Maugis et al. (2009) whilst Scenario 8 is a replica of Scenario 7 but using non-identical noise covariance matrices for each mixture component. The regression parameters used were $\beta_1 = ((0.5,0)', (0,1)', (2,0)', (0,3)')$, $\beta_2 = ((2,0.5)', (0.5,1)')$, $\beta_3 = ((2,0)', (0,3)')$. The notation 0_n denotes an n element row of zeroes.

scenarios 3-8, the related variables have comparatively higher variance than the independent clustering variables. Furthermore, it is interesting to note that by examining the posterior mean correlation matrix of the cluster with the highest posterior mean mixture weight, the correlations between variables can also be observed.

As a consequence we are able to qualitatively identify the two independent clustering variables and our overall analysis compares favourably with that in the original study by Maugis et al. (2009) and is much improved compared to the method of Raftery and Dean (2006) which tends to select both relevant and irrelevant clustering variables. We note that for this simulated example the data generating process follows the structure of the model proposed in Maugis et al. (2009), that is, they have the true model contained under their prior.

3.2 Cluster-specific variable relevance determination

Data. We simulated $1,000$ p-dimensional normally distributed random vectors for each of the three simulation scenarios described in Table 2 according to, $x_i|z_i = k, \mu, \Sigma \sim \text{N}(\mu_k, \Sigma_k)$. The cluster assignments z_i were drawn from a multinomial distribution with parameters w and the covariance matrix for each cluster was randomly drawn from an Inverse-Wishart distribution $\Sigma_k \sim \text{IW}(p+1, I_p)$ where I_p denotes the $p \times p$ identity matrix.

Scenario (a) depicts a case where the first three variables are relevant to all three clusters in the data. In scenario (b) each of the first three variables is responsible for differentiating one cluster from the others whilst scenario (c) uses 4 clustering variables where variables 1-3 are responsible for one cluster each and variable 4 is associated with

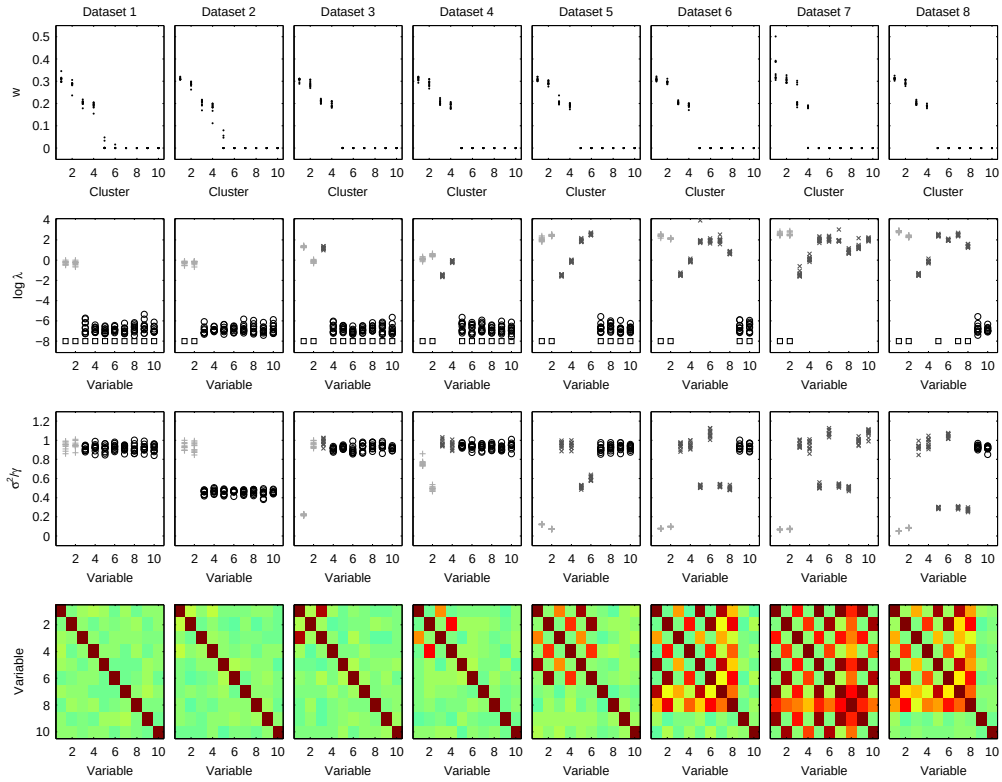

Figure 1: Variable selection for eight simulated scenarios. Top row - posterior mean estimates of mixture weights for ten datasets generated for each of the eight simulation scenarios. In all instances our method was able to estimate the correct number of clusters (4) and the mixture weights (0.3, 0.3, 0.2, 0.2). Second row - Posterior mean estimates of the $\log \lambda$ for the clustering variables ($+$) and the related variables ($\times$) are high whilst the irrelevant clustering variables ($\circ$) have low values indicating that the means in this coordinate direction, for all clusters, are being shrunk towards a common value. Variables selected using the method of Raftery and Dean (2006) are shown using asterisks ($\square$). Third row - posterior mean estimates of the quantity, σ_j^2/γ_j, which describes the variance associated with measurements of the j-th variable. For the clustering variables, the variance associated with the directly relevant clustering variables is found to be lower than that associated with related clustering variables. Bottom row - heatmaps showing example correlation matrices for cluster 1 estimated from each scenario. The correlation structure correctly identifies the relationships between variables used in the simulation of the data.

two clusters. Figure 2 depicts data for Scenario (c) where we can clearly see that one cluster can be clearly segregated from the others using variables 1-3 whilst variable 4 separates two clusters from the others.

Scenario	(a)	(b)	(c)
p	10	10	10
K	3	3	5
w	$[1/3, 1/3, 1/3]$	$[1/3, 1/3, 1/3]$	$[1/5, 1/5, 1/5, 1/5, 1/5]$
μ	$\mu_1 = [-3, 3, -3, 0_7]$ $\mu_2 = [0, 0, 0, 0_7]$ $\mu_3 = [3, -3, 3, 0_7]$	$\mu_1 = [3, 0_9]$ $\mu_2 = [0, 3, 0_8]$ $\mu_3 = [0, 0, 3, 0_7]$	$\mu_1 = [-3, 0_9]$ $\mu_2 = [0, 3, 0_8]$ $\mu_3 = [0, 0, -3, 0_7]$ $\mu_4 = [0, 0, 0, 3, 0_6]$ $\mu_5 = [0, 0, 0, -3, 0_6]$

Table 2: Cluster-specific variable importance simulation parameters.

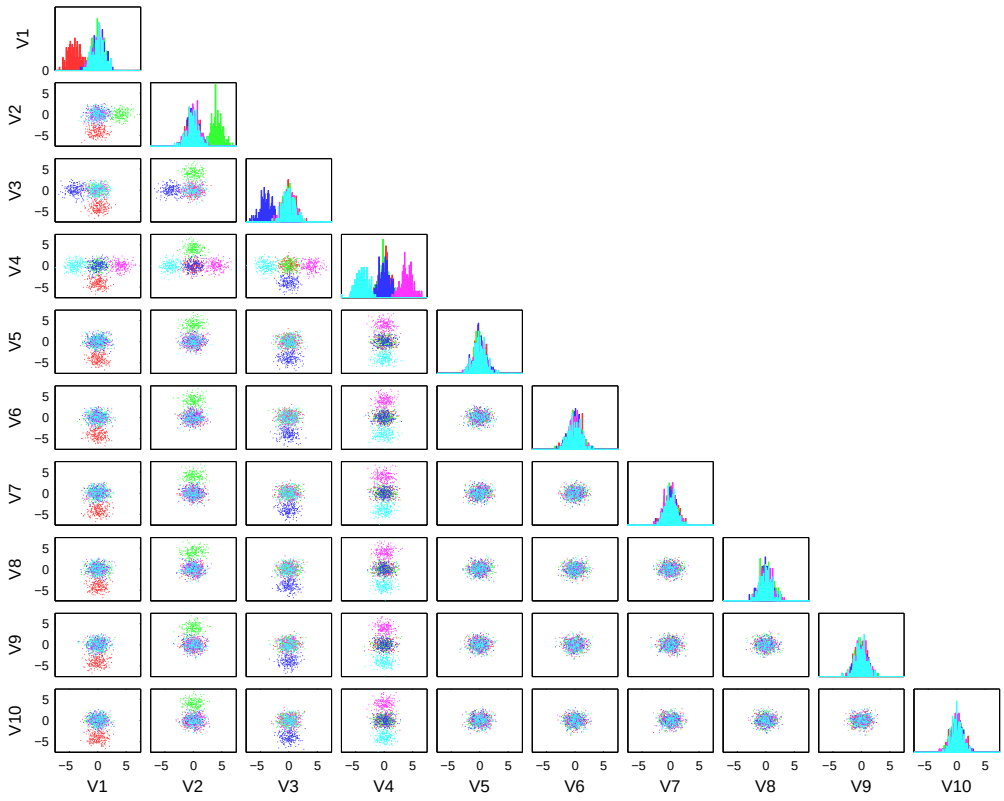

Figure 2: A simulated dataset for cluster-specific variable selection.

Results. Figure 3 shows the analysis of these three datasets using the cluster-specific variable relevance determination variant of our model. In all three scenarios, the posterior estimates of the mixture weights indicate that the number of clusters was correctly

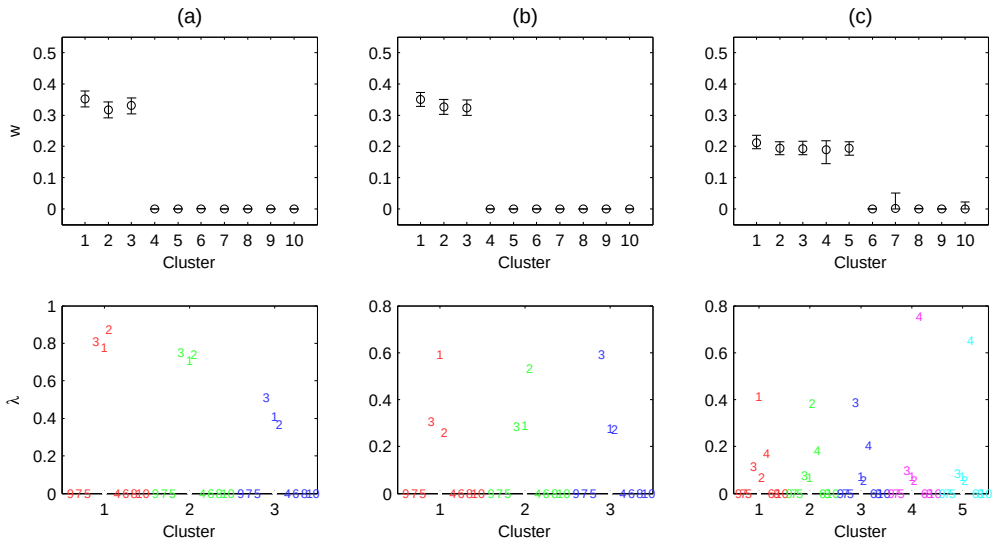

Figure 3: Cluster-specific variable selection on three simulated datasets. Top row - Posterior means (with 5% and 95% quantiles) for the mixture weights w. The correct number of clusters (3, 3 and 5) are identified by our method in each instance. Bottom row - Posterior means of λ are shown for each cluster and the numbers indicate the corresponding variable.

identified with 3, 3 and 5 clusters having significant weights respectively. An examination of the variances λ_{jk} for each cluster shows that our cluster-specific variable relevance determination method is able to identify that, in scenario (a), the clustering variables 1-3 all play a part in the identification of the three clusters. In scenario (b), whilst all of the variables 1-3 are shown to be useful for clustering, compared to variables 4-10, each cluster has one λ value that shows greater importance over the other two. For scenario (c), the variables 1-3 are each associated with clusters 1-3 respectively, whilst the fourth variable is important for clustering of all five classes but particularly associated with the fourth and fifth clusters.

3.3 Real Data Examples

Data. We analysed the *Leptograpus* Crabs and Iris datasets. The Crabs data contains four classes with fifty data points in each class and measurements for five variables. The Iris data has a three-class structure with fifty data points in each class and consists of measurements for four variables. For both datasets, we also appended three additional irrelevant variables simulated from a standard normal distribution.

Results. For the Crabs data, we computed the principal components on the standardised data and clustered in the principal component space. Clustering in the original

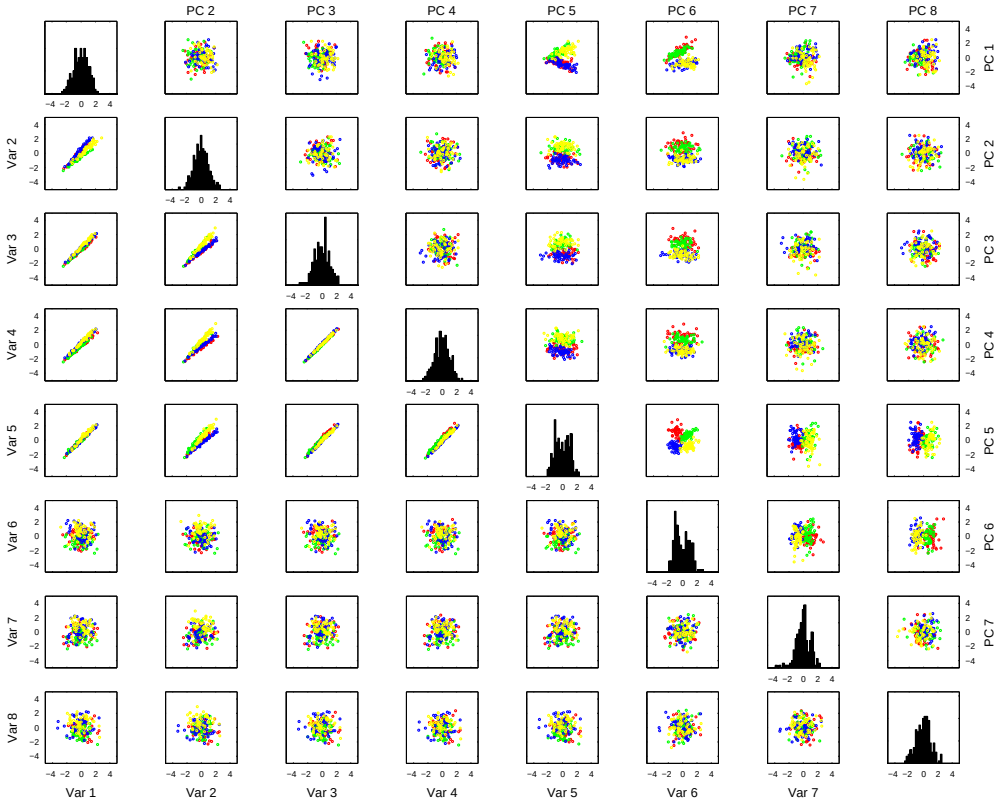

Figure 4: Crabs data (including three irrelevant variables) in standardized original and principal component coordinate space.

coordinate space led to high sensitivity to initialisation as the variables are highly correlated and the data lie in an extremely thin, elongated region in the original coordinate space. This near-collinearity is revealed by principal component analysis which estimates the variances as 4.8048, 1.2764, 1.0541, 0.6583, 0.1488, 0.0449, 0.0110 and 0.0017 of each of the principal components (see Figure 4) and the first principal component lies in the direction $[0.4509, 0.4276, 0.4521, 0.4500, 0.4502, 0.0265, -0.0208, -0.0550]'$. The principal components 2-4 correspond to variation in the added irrelevant variables.

Figure 5 shows that, when clustering using the principal components, the four clusters were found to have significant weight. The principal components 5 and 6 were found to be of particular importance to clustering, but no principal component was particularly associated with any single cluster. If we plot the data in these coordinates only the four clusters can be observed. If we assigned each data point to the cluster with maximum posterior probability then we obtain 15 classification errors. Note that since we are using the principal components, the variable selection is performed not on

the original variables but on a linear combination of those variables.

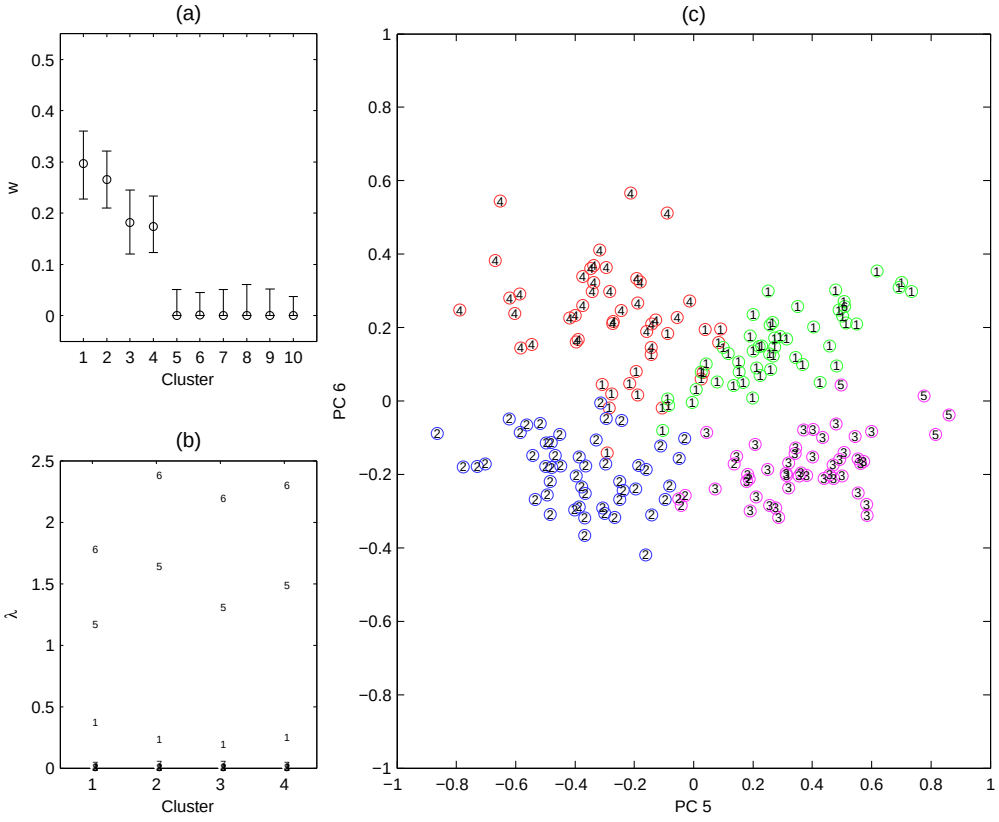

Figure 5: Cluster-specific variable selection on the Crabs dataset. (a) Posterior means (with 5% and 95% quantiles) for the mixture weights w, (b) Posterior estimates of the cluster-specific variance parameters λ and (c) the data with the true classification shown in colored circles and the classification given by our method by numbers.

For the Iris data, our method identified the three underlying classes and indicated that variables 3 (petal length) and 4 (petal width) to be of particular importance. An examination of the data in the space spanned by these variables shows that the three classes can be distinguished based on these two variables alone.

3.4 Cancer Data Analysis

We examined a cancer dataset containing Illumina SNP-CGH microarray measurements for 80 patients with B-cell chronic lymphocytic leukaemia (B-CLL) (Figure 7(a)) (Knight et al. (2011)). Each array is comprised of 1,022,726 SNP-CGH probe measurements that measure relative genome-wide DNA copy number abundance. Patterns

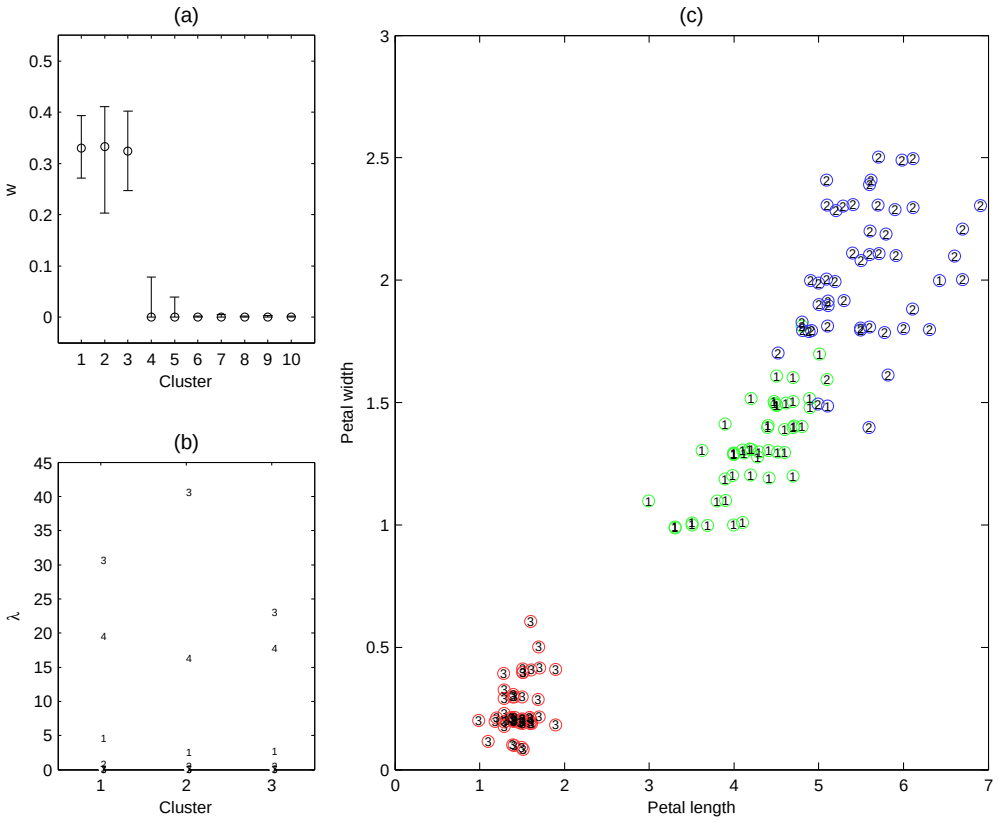

Figure 6: Cluster-specific variable selection on the Iris dataset. (a) Posterior means (with 5% and 95% quantiles) for the mixture weights w, (b) Posterior estimates of the cluster-specific variance parameters λ and (c) the data with the true classification shown in colored circles and the classification given by our method by numbers.

of genome-wide gain and loss of DNA copy number can provide important prognostic markers for disease type and severity and it is our interest to find distinct genetic sub-types of B-CLL associated with certain clinical outcomes. We applied our clustering method, assuming cluster-specific but diagonal covariances, and used cluster-specific variable relevance determination to identify potential genetic sub-types of B-CLL in our sample dataset. Due to the considerable dimensionality of the problem, we reduced to a summarised set of 9,953 summary measurements by taking averages over 100 probe windows.

Figure 7(b) shows ten clusters identified by our method that have non-negligible mixture weights. Although the significance of a number of these clusters must be verified for clinical significance, three of these clusters correspond to three well-established mutation types involving deletions of the long arm on chromosome 11 (cluster 2), trisomy

12 (an additional chromosome 12) (cluster 6) and loss of the short arm of chromosome 17 (cluster 3) (Döhner et al. (2000)). A fourth cluster (8) indicates potential significance of genomic aberrations of chromosomes 4, 16, 17, 18, 19 and 20 although it is beyond the remit of this study to validate the potential prognostic value of this genetic subtype. Figure 7(c) shows that our cluster-specific variable determination automatically highlights the relevant genetic regions for each of these sub-groups.

4 Discussion

We have derived a Bayesian hierarchical nonparametric mixture model approach applicable to situations where there is *a priori* uncertainty as to the relevance of the measured variables to the clustering problem. The hierarchical structure provides a flexible framework in which to build prior beliefs on primitives which capture the notion of sparsity; namely standardised differences of cluster locations. Such models can accommodate cluster specific variable relevance, while the nonparametric prior allows us to treat the number of mixtures as unknown.

Extensions to component distributions, other than the Normal or Student-t, for inference in problems involving mixed data types may be possible but are beyond the scope of this paper. In the spirit of the model presented here, the extension would rely on it being possible to parameterise the component distributions in terms of a location and then to provide a hierarchical structure to impose a shrinkage-type prior on the difference in locations between clusters. The details of the implementation would depend on the choice of component distributions being used and alternative data models may lack the conjugacy properties that have enabled us to use efficient Gibbs sampling strategies for computational inference here.

One point noted by the referees as a principle assumption of our model is that clusters are defined on differences in location. Although this scenario covers many applications there may be circumstances in which clusters have a common location but differ in shape. Figure 8(a) shows a simulated ten-dimensional dataset generated from a three-component Normal mixture model (equal weights) in which only the first two variables are relevant to the clustering. Two components share a common location but have different covariances whilst the third is well-separated. Data for the remaining variables were generated from a unit normal. Applying our method to data of this type, the variance parameter λ on the means μ indicates that the first variable is of relevance to the clustering as expected but no indication of the importance of the second variable is provided. However, inspection of the posterior distribution of the scale variables σ^2 does reveal some further information about the importance of the second variable since the σ^2 scale differs for the second clustering variable compared to the non-clustering variables. One could envisage the use of a more explicit hierarchical model structure that performs shrinkage (toward a common value) on the variance-covariance components of variables irrelevant to the cluster structure. This will be an interesting avenue for future investigation.

A further area for development is the modelling of the individual component den-

sities particularly in a high-dimensional context. In our cancer analysis example, for computational reasons, we restricted the covariances to be of diagonal form but more robust covariance modelling could be developed using dimension reduction techniques. As noted by one referee, model misspecification could lead to the appearance of spurious clusters and the differences in means between these spurious clusters could produce misleading evidence for variable importance. It is of benefit here that our measure of variable importance is continuous and allows us to qualitatively rank variables of importance. In our cancer example, it is clear that across the genome there are loci which do have values of λ which suggest some importance but which are likely due to over-clustering. However, this was not overly problematic since the differences in means associated with the clusters and loci of greatest interest (the known CLL sub-types) were much greater than any variations in means between clusters at irrelevant loci. Nonetheless, in more noisy datasets, a more flexible covariance model may be required.

Finally, a general issue that we have not directly addressed is the issue of designing well-mixing MCMC samplers that are able to fully explore the posterior parameter space. In this paper, we have taken advantage of the conjugacy within our model in order to find conditional distributions from which a Gibbs Sampler for posterior inference can be derived. Alternative samplers for Dirichlet process mixture models that employ more flexible split-merge moves have been previously proposed by Jain and Neal (2000) and Green and Richardson (2001) and these can also be integrated with tempering techniques as in Kim et al. (2006) in order to avoid the local modes in which Gibbs Samplers have the propensity to become trapped. However, it should be noted that in actual applications, the computational burden of more elaborate sampling techniques may be prohibitive when applied to very large, high-dimensional data sets and a more practical strategy would be to forego the full exploration of the parameter space and to focus only on a high probability region with a Gibbs sampler in conjunction with good initialisation. One of the techniques we have used in our analysis is to initialise from a heavily over fitted initial mixture model with many more components than we might expect to exist. Hence the emphasis is then on pruning out and merging clusters rather than cluster creation which is difficult without explicit split moves.

References

Andrews, D. and Mallows, C. (1974). "Scale mixtures of normal distributions." *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1): 99–102. 331

Antoniak, C. (1974). "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems." *The Annals of Statistics*, 2: 1152–1174. 334

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press. 330

Döhner, H., Stilgenbauer, S., Benner, A., Leupolt, E., Kröber, A., Bullinger, L., Döhner, K., Bentz, M., and Lichter, P. (2000). "Genomic aberrations and survival in chronic lymphocytic leukemia." *New England Journal of Medicine*, 343(26): 1910–1916. 345

Dy, J. and Broadly, C. (2004). "Feature selection for unsupervised learning." *Journal of Machine Learning Research*, 5: 845–889. 330

Escobar, M. (1994). "Estimating normal means with a Dirichlet process prior." *Journal of the American Statistical Association*, 89(425): 268–277. 334

Escobar, M. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90(430): 577–88. 334, 336

Ferguson, T. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1(2): 209–230. 334

Fisher, R. (1936). "The use of multiple measurements in taxonomic problems." *Annals of Eugenics*, 7: 179–188. 331

Fraley, C. and Raftery, A. (2002). "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American Statistical Association*, 97(458): 611–631. 331

Friedman, J. and Meulman, J. (2004). "Clustering objects on subsets of attributes (with Discussion)." *Journal of the Royal Statistical Society, Series B*, 66: 815–849. 330, 337

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. New York: Springer. 331

Green, P. and Richardson, S. (2001). "Modelling Heterogeneity With and Without the Dirichlet Process." *Scandinavian Journal of Statistics*, 28(2): 355–375. 334, 346

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition)*. Springer. 329, 330

Hoff, P. (2006). "Model-based subspace clustering." *Bayesian Analysis*, 1(2): 321–344. 330

Ishwaran, H. and James, L. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96: 161–73. 334

Jain, S. and Neal, R. (2000). "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model." *Journal of Computational and Graphical Statistics*, 13: 158–82. 334, 346

Jasra, A., Holmes, C., and Stephens, D. (2005). "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling." *Statistical Science*, 20(1): 50–67. 331

Kalli, M., Griffin, J., and Walker, S. (2011). "Slice sampling mixture models." *Statistics and Computing*, 1: 93–105. 334

Kim, S., Tadesse, M., and Vannucci, M. (2006). "Variable selection in clustering via Dirichlet process mixture models." *Biometrika*, 93(4): 877–893. 330, 346

Knight, S. J. L., Yau, C., Timbs, A., Sadighi-Akha, E., Dreau, H., Burns, A., Oscier, D., Pettitt, A., Holmes, C., Taylor, J., Cazier, J.-B., and Schuh, A. (2011). "A genome-wide array-based sequential analysis quantifies the proportion of sub-clones carrying genomic changes in B-cell chronic lymphocytic leukaemia and reveals the complexity of clonal dynamics." *Submitted to Leukemia*. 343

Law, M., Figueiredo, M., and Jain, A. (2004). "Simultaneous feature selection and clustering using mixture models." *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26(9): 1154–1166. 330

MacEachern, S. (1998). "Estimating mixture of Dirichlet process models." *Journal of Computational and Graphical Statistics*, 7(2): 223–238. 334

MacKay, D. (1995). "Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks." *Network: Computation in Neural Systems*, 6(3): 469–505. 330

Maugis, C., Celeux, G., and Martin-Magniette, M. (2009). "Variable Selection for Clustering with Gaussian Mixture Models." *Biometrics*, 65: 701–709. 330, 334, 337, 338

Neal, R. (2000). "Markov Chain Sampling: Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9: 249–265. 334

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag. 330

Pan, W. and Shen, X. (2007). "Penalized model-based Clustering with Application to Variable Selection." *Journal of Machine Learning Research*, 8: 1145–1164. 330

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo for Dirichlet process hierarchical models." *Biometrika*, 95: 169–186. 334, 336

Raftery, A. and Dean, N. (2006). "Variable selection for model-based clustering." *Journal of the American Statistical Association*, 101(473): 168–178. 330, 333, 334, 338, 339

Richardson, S. and Green, P. (1997). "On Bayesian analysis of mixtures with an unknown number of components." *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(4): 731–792. 331, 334

Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, 4: 639–50. 334

Stephens, M. (2000). "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods." *The Annals of Statistics*, 28(1): 40–74. 334

Tadesse, M., Sha, N., and Vannucci, M. (2005). "Bayesian Variable Selection in Clustering High-Dimensional Data." *Journal of the American Statistical Association*, 100(470): 602–618. 330, 333, 334

Tipping, M. (2001). "Sparse Bayesian Learning and the Relevance Vector Machine." *Journal of Machine Learning Research*, 1: 211–244. 330

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York, Wiley. 331

Walker, S. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics-Simulation and Computation*, 36(1-3): 45–54. 334, 335

Acknowledgments

CY is funded by a UK Medical Research Council Special Training Fellowship in Bioinformatics (Reference No. G0701810). CH is funded by a Programme Leaders award from the Medical Research Council. We thank Sam Knight and Anna Schuh for access to the Chronic Lymphocytic Leukemia data and the Associate Editor and anonymous referee for their comments.

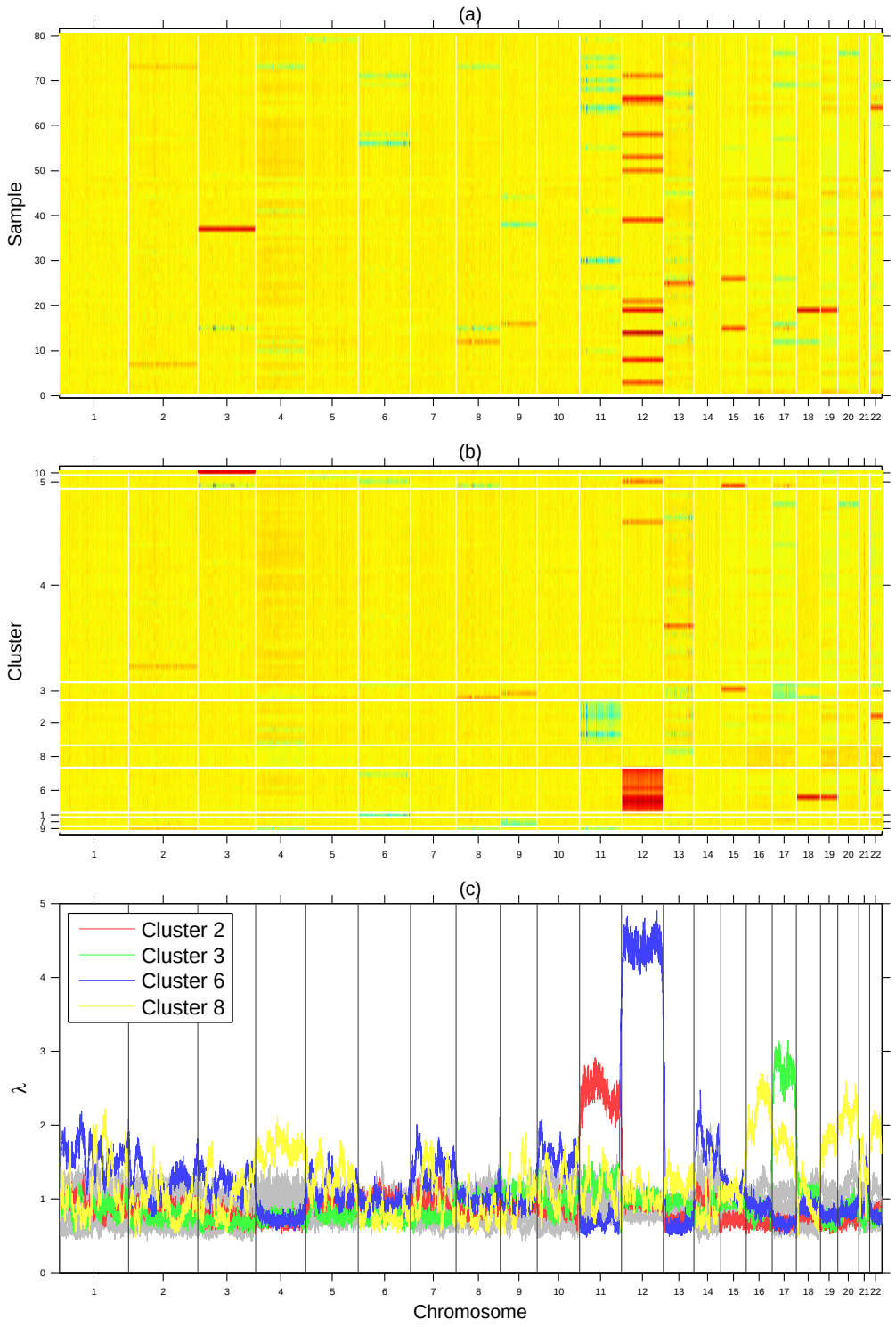

Figure 7: Clustering of chronic lymphoblastoid leukemia data. (a) Original data (red/green - low/high signal intensity), (b) clustered data and (c) variable relevance for clusters 2, 3, 6, and 8. The clusters 2, 3 and 6 correspond to known B-CLL genetic sub-types.

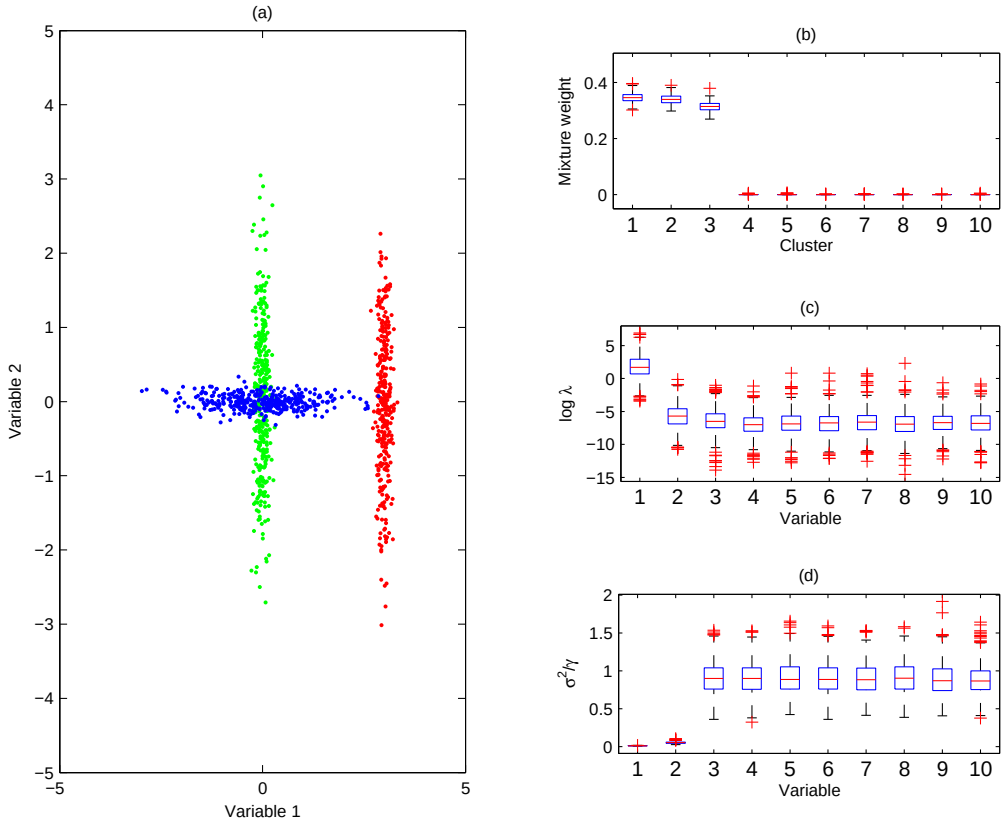

Figure 8: Mixture models based on cluster covariances. (a) Data (from the first two dimensions) of three components with common mean but differing covariances. (b) Posterior distribution of the mixture weights estimated from the data. (c) Posterior distribution of the variance relevance parameters $\log \lambda$. (d) Posterior distribution of the scale parameters σ^2.