

Clustering based on Dirichlet mixtures of attribute ensembles

Peter D. Hoff *

Technical Report no. 448

Department of Statistics

University of Washington

May 18, 2004

Abstract

We propose a model-based approach to identifying clusters of objects based on subsets of attributes, so that the attributes that distinguish a cluster from the rest of the population, called an attribute ensemble, may depend on the cluster being considered. The model is based on a Pólya urn cluster model, which is equivalent to a Dirichlet process mixture of multivariate normal distributions. This model-based approach allows for the incorporation of application-specific data features into the clustering scheme. For example, in an analysis of genetic CGH array data we account for spatial correlation of genetic abnormalities along the genome.

Some key words: nonparametric Bayes, unsupervised learning, subspace clustering, variable selection, COSA.

1 Introduction

In this paper we consider a model-based approach to clustering objects based on subsets of attributes. The data we consider consist of m -dimensional vectors of attributes \mathbf{y}_i measured on each member of a population of units $i = 1, \dots, n$. In a typical model-based cluster analysis, one tries to find a value $K < n$ such that the data are well approximated by a mixture of K multivariate normal distributions with means $\boldsymbol{\mu}_{(1)}, \dots, \boldsymbol{\mu}_{(K)}$ (see McLachlan and Basford (1988), or Fraley and Raftery (2002) for a review). Such procedures estimate each attribute mean separately for each cluster, typically with $\hat{\mu}_{(k),j} = \bar{y}_{(k),j}$, the mean of attribute j for observations in cluster k . In some

*Departments of Statistics, Biostatistics and the Center for Statistics and the Social Sciences, University of Washington, Seattle, Washington 98195-4322, U.S.A.. Email: hoff@stat.washington.edu. This research was supported by National Cancer Institute grant CA077607-04. The author thanks Peggy Porter's lab at the Fred Hutchinson Cancer Research Center for helpful discussions and the use of their data. The author also thanks Mary Emond, Li Hsu, Douglas Grove, Elena Erosheva and Werner Steutzle for helpful discussions.

cases this may result in overfitting: suppose the differences between two given clusters can be summarized by a difference in only a subset of the attribute means, with the subset depending on the pair of clusters being compared. For example, consider a bivariate population which is a mixture of three groups having means $(\mu_{(A1)}, \mu_{(B1)})$, $(\mu_{(A1)}, \mu_{(B2)})$ and $(\mu_{(A2)}, \mu_{(B2)})$. There is variation in both of the attributes, but two of the three cluster pairs differ at only one attribute.

A simple exploratory approach to finding clusters based on subsets of attributes is principle components analysis. However, such an analysis for the bivariate example described above could obscure the data. Figure 1 plots a situation in which the three clusters are identified along the first principal component and two along the second. The clustering along the second component puts the blue and green groups together, and separate from the red group, even though the both the blue and green groups are closest to the red group in each of the original attributes. In much higher dimensions the differences between the groups in the original coordinate attributes could become further obscured.

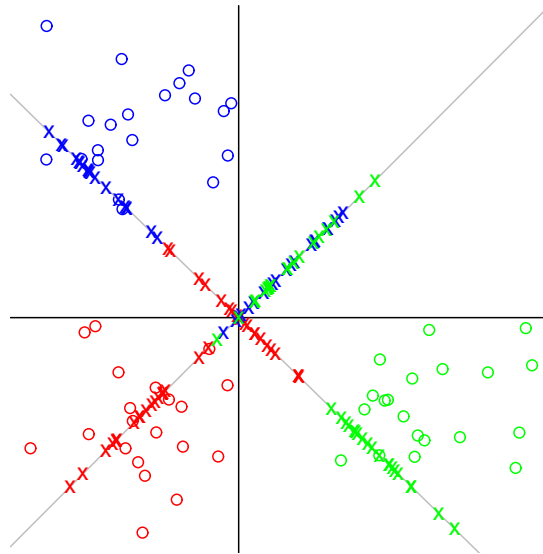


Figure 1: Looking for clusters along the principle components.

In data mining applications where the number of attributes is large, it is quite possible that only a small number of the attributes will provide a grouping of the observations, and that among these attributes, only some will differ between any two particular groups. This has lead Friedman and Meulman (2004) to develop the notion of “clustering on subsets of attributes.” One version of their approach iteratively generates a dissimilarity between each pair of objects based on weighted attribute differences, where the weights are object-specific. Their clustering criteria and computational approaches are largely driven by heuristics, and their methods do not provide an estimate of the number of clusters. A related group of methodologies are subspace clustering algorithms, which search for clusters in subspaces of the attributes (see Parsons, Haque and Liu 2004 for a

review). These too are driven by heuristic criteria and search algorithms. In contrast, Newton (2002) provides a model-based subspace clustering method for binary data arising from genomic abnormalities.

This paper presents a model-based clustering approach from which provides estimates of the number of clusters, cluster memberships and an identification of which attributes are likely to be defining each cluster. This approach combines the goals of principle components analysis (finding combinations of attributes that contribute to variation), variable selection (identifying which variables differ from a population mean), and clustering (finding groups having similar attribute values), but provides them in a unified statistical procedure. Additionally, this model based approach allows for an assessment of uncertainty in the clustering, and the incorporation of known features of the data into the clustering algorithm. For example, in an analysis of genetic array data we will search for clusters of tumor cells that have abnormalities at common genomic locations. These genetic abnormalities occur spatially along the genome, which is a feature we will incorporate into the clustering model.

The clustering approach is based on finding groups whose attribute means differ from one another. In the context of a statistical model, we are looking for a value K , a cluster membership function $c : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$ and K m -dimensional means $\boldsymbol{\mu}_{(1)}, \dots, \boldsymbol{\mu}_{(K)}$ such that $\sum_{k=1}^K \sum_{i:c(i)=k} (\mathbf{y}_i - \boldsymbol{\mu}_{(k)})^2$ is small, as in the usual clustering problem, but where we hope to write $\boldsymbol{\mu}_{(k)} = \boldsymbol{\mu} + \boldsymbol{\gamma}_{(k)}$ with $\boldsymbol{\gamma}_{(k)}$, the vector of “mean shifts” for group k , being zero at many entries. We define the *attribute ensemble* for group k as the set of attribute indices $\{j : \gamma_{(k),j} \neq 0\}$. Alternatively, we can parameterize the vector of mean shifts as $\boldsymbol{\gamma}_{(k)} = \mathbf{s}_{(k)} \times \boldsymbol{\delta}_{(k)}$, with $\boldsymbol{\delta}_{(k)} \in \mathbb{R}^m$, $\mathbf{s}_{(k)} \in \{0, 1\}^m$ and “ \times ” indicating element-wise multiplication. In what follows, for a vector $\boldsymbol{\gamma} = \{\gamma_1, \dots, \gamma_m\}$ we will sometimes denote $s_j = 1(\gamma_j \neq 0)$, the indicator that γ_j is non-zero, and $\delta_j = \gamma_j$ if $s_j = 1$.

To make it clearer what these parameters represent, consider the fabricated data in Figure 2. The three right-most panels of the figure represent a possible clustering of the left-most panel into three clusters with attribute ensembles $\{3, 4\}$, $\{2, 3\}$ and $\{4, 5\}$. For these data and clustering, the attributes defining the clusters depend on which cluster is being examined. Cluster 1 differs from the population mean at attributes 3 and 4, whereas cluster 2 differs from the population mean at attributes 2 and 3, and so on.

In the next section we outline the model, which is based on a Dirichlet process for the distribution of the mean shifts $\boldsymbol{\gamma}$. In Section 3 we discuss parameter estimation and model behavior, and follow this up with a small simulation study in Section 4. In section 5 we apply an extension of the model to spatially correlated genetic array data in which we try to identify groups of tumor cells having similar patterns of chromosomal abnormalities. A discussion follows in Section 6.

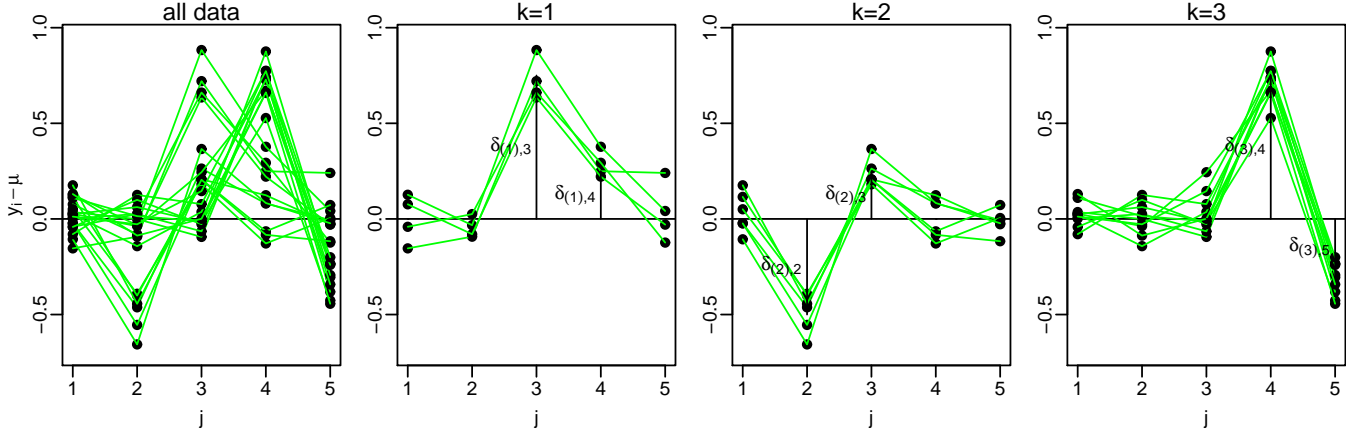


Figure 2: Some data and a possible clustering. The green lines connect the attribute values of a single unit.

2 Modeling clusters with a Pólya urn scheme

The clustering approach is based on the following model and prior:

$$f \sim \text{Dirichlet}(\alpha, f_0) \quad (0)$$

$$\gamma_1, \dots, \gamma_n \sim \text{i.i.d. } f \quad (1)$$

$$\epsilon_1, \dots, \epsilon_n \sim \text{i.i.d. multivariate normal } (\mathbf{0}, \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\}) \quad (2)$$

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i. \quad (3)$$

The values $\gamma_1, \dots, \gamma_n$ are the “mean shifts” away from $\boldsymbol{\mu}$ described earlier, and f is the distribution of the mean shifts. Via a particular form for f_0 , this model provides the following:

- (a) the possible values of f consist of all discrete distributions on \mathbb{R}^m (f is modeled nonparametrically);
- (b) a sample of n mean shifts from f may have less than n unique values (the γ ’s “cluster”);
- (c) the attributes j for which $\gamma_{i,j} \neq 0$ may depend on cluster membership (clustering attributes may be cluster-specific).

The distribution f is estimated nonparametrically using the Dirichlet process prior. The result is called a Dirichlet process mixture model, as the distribution for \mathbf{y} can be written as the mixture $p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \int p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma})f(d\boldsymbol{\gamma})$ where the mixing measure f is a Dirichlet process. Such models have a history going back to Antoniak (1974), and have been put to practical use by MacEachern (1994), Escobar and West (1995,1998), MacEachern and Müller (1998), Neal (2000), Dahl (2003) and others.

Samples from a Dirichlet process are discrete, and so f will have support on a countable number of $\boldsymbol{\gamma}$ -values. This discreteness implies that a sample from f could have a number of ties, and thus

$$\text{Pólya Urn}(\alpha, f_0) \left\{ \begin{array}{l}
\gamma_{(1)} \left\{ \begin{array}{l} \mathbf{y}_1 = \boldsymbol{\mu} + \gamma_{(1)} + \epsilon_1 \\ \vdots \\ \mathbf{y}_{n_1} = \boldsymbol{\mu} + \gamma_{(1)} \times + \epsilon_{n_1} \end{array} \right. \\
\gamma_{(2)} \left\{ \begin{array}{l} \mathbf{y}_{n_1+1} = \boldsymbol{\mu} + \gamma_{(2)} + \epsilon_{n_1+1} \\ \vdots \\ \mathbf{y}_{n_1+n_2} = \boldsymbol{\mu} + \gamma_{(2)} + \epsilon_{n_1+n_2} \end{array} \right. \\
\vdots \\
\gamma_{(K)} \left\{ \begin{array}{l} \mathbf{y}_{n-n_K+1} = \boldsymbol{\mu} + \gamma_{(K)} + \epsilon_{n-n_K+1} \\ \vdots \\ \mathbf{y}_n = \boldsymbol{\mu} + \gamma_{(K)} + \epsilon_n \end{array} \right.
\end{array} \right.$$

Figure 3: The clustering scheme. Indices of the units are ordered according to cluster membership, and n_k denotes the number of members of cluster k .

form a clustering. That the Dirichlet process prior gives a simple and interpretable model for a clustering process can be seen via the Pólya urn representation of a sample from a Dirichlet process, which is described in Blackwell and MacQueen (1973): If $f \sim \text{Dir}(\alpha, f_0)$ and $\gamma_1, \dots, \gamma_n$ are i.i.d. samples from f , then the marginal joint distribution of the γ_i 's is equal in distribution to a sequence generated as follows:

1. sample $\gamma_1 \sim f_0$;
2. sample $\gamma_2 \sim \frac{\alpha}{\alpha+1}f_0 + \frac{1}{\alpha+1}1_{\gamma_1}(\cdot)$;
- \vdots
- n. sample $\gamma_n \sim \frac{\alpha}{\alpha+n-1}f_0 + \frac{n-1}{\alpha+n-1}\hat{f}_{n-1}$,

where $1_{\gamma_1}(\cdot)$ is a point-mass measure on γ_1 and \hat{f}_{n-1} is the empirical distribution of $\gamma_1, \dots, \gamma_{n-1}$. The above process is called a Pólya urn scheme with parameters α and f_0 . It is clear that, depending on α , the sample $\gamma_1, \dots, \gamma_n$ may have been generated by fewer than n draws from f_0 and thus have fewer than n unique values, achieving item (b) described above. We will denote the number of draws from f_0 as K , and the values of the draws as $\gamma_{(1)}, \dots, \gamma_{(K)}$. The function mapping the unit labels to the independent draws will be denoted $c : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$. As can be seen, the parameter α determines the prior distribution on K , whereas f_0 determines the distribution on the values of $\gamma_{(1)}, \dots, \gamma_{(K)}$. A diagram describing this hierarchy is given in Figure 3. Rewriting in terms of the Pólya urn representation, our full probability model for all of the unknown quantities

can thus be described by replacing lines (0) and (1) with

$$\gamma_1, \dots, \gamma_n \sim \text{Pólya urn}(\alpha, f_0) \quad (1')$$

To model property (c) described above, we choose f_0 so that a sample γ from f_0 is equal in distribution to the random variable defined by $\mathbf{s} \times \boldsymbol{\delta}$ where $\mathbf{s} \in \{0, 1\}^m$, $\boldsymbol{\delta} \in \mathbb{R}^m$ and “ \times ” represents elementwise multiplication. For example, we could model each component γ_j as being equal in distribution to $s_j \times \delta_j$ where $s_j \sim \text{binary}(\frac{e^\theta}{1+e^\theta})$ and $\delta_j \sim \text{normal}(0, \tau^2)$. In this case, θ represents the log-odds that a cluster has a given attribute in its ensemble, and thus has a mean shift away from the population mean at that attribute. The parameter τ^2 represents the average squared magnitude of the mean shifts. This type of distribution has been suggested as a prior on regression parameters by Mitchell and Beauchamp (1988) as a means of providing a type of Bayesian variable selection. This differs somewhat from our use of this type of distribution, since even if $s_{(k),j} = 0$, attribute j could still be in the ensemble of another cluster.

3 Model estimation and behavior

We first consider modeling f_0 so that the components of $\gamma \sim f_0$ are independent, so

$$f_0(\boldsymbol{\gamma}) = \prod_{j=1}^m \left\{ \left[\frac{e^{\theta_j}}{1+e^{\theta_j}} g(\gamma_j | \psi_j) \right]^{(\gamma_j \neq 0)} \times \left[\frac{1}{1+e^{\theta_j}} \right]^{(\gamma_j = 0)} \right\}. \quad (4)$$

where $g(\gamma_j | \psi_j)$ is a continuous density. Each term in the above product is then a density with respect to the measure $\nu(A) = \text{Leb}(A) + 1(0 \in A)$, i.e. Lebesgue measure plus a point-mass at zero. This is less complicated than it sounds - one can simply view each γ_j as being equal in distribution to a binary random variable s_j multiplied by a sample δ_j from $g(\cdot | \psi_j)$. For simplicity, we will work with the variables $\mathbf{s} = \{s_1, \dots, s_m\}$ and $\boldsymbol{\delta} = \{\delta_1, \dots, \delta_m\}$ when the distinction does not matter. The parameters $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_m\}$ and $\boldsymbol{\psi} = \{\psi_1, \dots, \psi_m\}$ determine the probabilities of mean shifts within a cluster and the distributions of those mean shifts, respectively. For computational generality we allow these to vary across attributes in this section, but later we will suggest a simplified model.

We first examine the probability of the data within a cluster conditional on the clustering $c(\cdot)$ but marginal over the values of the mean shifts $\boldsymbol{\gamma}_{(1)}, \dots, \boldsymbol{\gamma}_{(K)}$. This probability calculation offers insight into the types of clusters that the model will identify, as well as assist us in parameter estimation. This joint density of observations from a common cluster k is obtained by integrating over the unknown $\boldsymbol{\gamma}$ for that cluster, or equivalently, over the possible values of \mathbf{s} and $\boldsymbol{\delta}$.

$$p(\{\mathbf{y}_i : c(i) = k\} | c(\cdot), \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{\mathbf{s} \in \{0,1\}^m} f_0(\mathbf{s} | \boldsymbol{\theta}) \int \left\{ \prod_{i:c(i)=k} p(\mathbf{y}_i | \mathbf{s}, \boldsymbol{\delta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \right\} f_0(\boldsymbol{\delta} | \boldsymbol{\psi}) d\boldsymbol{\delta}.$$

Due to the independence across attributes, this is equal to

$$\prod_{j=1}^m \sum_{s_j=0}^1 \frac{e^{\theta_j s_j}}{1 + e^{\theta_j}} \int \left\{ \prod_{i:c(i)=k} p(y_{i,j}|s_j, \delta_j, \mu_j, \sigma_j^2) \right\} g(\delta_j|\psi_j) d\delta_j.$$

assuming we can do the integral over δ_j and obtain $p(\{y_{i,j} : c(i) = k\}|s_j = 1, \mu_j, \sigma_j^2, \psi_j)$, let

$$\hat{\theta}_j(k) = \log \frac{p(\{y_{i,j} : c(i) = k\}|s_j = 1, \mu_j, \sigma_j^2, \psi_j)}{p(\{y_{i,j} : c(i) = k\}|s_j = 0, \mu_j, \sigma_j^2)}. \quad (5)$$

The value of $\hat{\theta}_j(k)$ can be thought of as the adjustment to the log odds of there being a mean shift in attribute j for members of cluster k , having observed data from that cluster and given values of μ_j, σ_j^2 and ψ_j . Alternatively, $\hat{\theta}_j(k)$ is a log Bayes factor for evaluating $H : E(y_{i,j}) \neq \mu_j$ versus H^c for data in group k .

Since the denominator in (5) is equal to $\prod_{i:c(i)=k} \text{normal}(y_{i,j} : \mu_j, \sigma_j^2)$, the joint probability density of all the data is then

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n | c(\cdot), \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\psi}) = \left\{ \prod_{k=1}^K \prod_{j=1}^m \frac{1 + e^{\theta_j + \hat{\theta}_j(k)}}{1 + e^{\theta_j}} \right\} \times \left\{ \prod_{i=1}^n \prod_{j=1}^m \text{normal}(y_{i,j} : \mu_j, \sigma_j^2) \right\}. \quad (6)$$

The second term on the right does not depend on the clustering or the parameters describing the distribution of mean shifts. This marginal distribution (6) provides a mechanism by which we can make Bayesian inference on the unknown parameters and gather insight into how the model chooses clusters and selects attributes. These items are explored in the next two subsections.

3.1 Parameter estimation via MCMC

One can make inference on $\{c(\cdot), \alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\psi}\}$ by constructing a relatively straightforward Markov chain which converges to the posterior distribution $p(c(\cdot), \alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{y}_1, \dots, \mathbf{y}_n)$. We suggest an algorithm based on Gibbs sampling of the cluster membership function and a Metropolis-Hastings algorithm for the other parameters, both done marginally over the γ -values. Alternatively, with conjugate models one can iteratively use the Gibbs sampler to generate values of cluster-specific mean shifts $\gamma_{(1)}, \dots, \gamma_{(K)}$ as well as the parameters $\boldsymbol{\theta}, \boldsymbol{\psi}$ defining their distribution. These approaches are the “standard” estimation techniques for Dirichlet process mixture models, and are discussed by MacEachern (1994) for mixtures of univariate normals, and in general by Neal (2000).

Given a current state of $\{c(\cdot), \alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\psi}\}$, the algorithm iteratively samples new states for each quantity as follows:

1. For $i \in \{1, \dots, n\}$ in random order, sample $c(i)$ from its full conditional distribution;
2. Sample α from its full conditional distribution;

3. (a) For $j = 1, \dots, m$, generate values of $\{\mu_j, \sigma_j^2, \theta_j, \psi_j\}^*$ from a proposal distribution, and accept with the appropriate probability.
3. (b) Alternatively or in tandem with 3.(a), use a Gibbs sampling procedure:
 - i. For $k \in \{1, \dots, K\}$, sample $\gamma_{(k)}$ from its full conditional distribution.
 - ii. For $j \in \{1, \dots, m\}$, sample $\{\mu_j, \sigma_j^2, \theta_j, \psi_j\}$ from its full conditional distribution.

The above steps are outlined in more detail below.

Sampling $c(\cdot)$: Unconditional on the observed data, the conditional distribution of $c(i)$ given the other values of $c(\cdot)$ and α is computed as follows: Let K be the number of unique values of $\{c(i') : i' \neq i\}$, and relabel these values as $1, \dots, K$ if unit i is currently in its own cluster. The conditional distribution of $c(i)$ is

$$\Pr(c(i) = k | c(i'), i' \neq i) \propto \begin{cases} n_{k,-i} & \text{if } k < K + 1 \\ \alpha & \text{if } k = K + 1 \end{cases}$$

where $n_{k,-i}$ is the number of units in cluster k not including unit i . In other words, unit i is placed into an existing cluster with probability proportional to the cluster's size, and is placed in a new cluster with probability proportional to α . Conditional on the data and the other parameters these probabilities are reweighted as

$$\Pr(c(i) = k | c(i'), i' \neq i, \alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{y}_1, \dots, \mathbf{y}_n) \propto \begin{cases} n_{k,-i} \times w_k & \text{if } k < K + 1 \\ \alpha \times w_{K+1} & \text{if } k = K + 1 \end{cases} \quad (7)$$

where the weights are given by

$$w_k = \prod_{j=1}^m \frac{1 + \exp\{\theta_j + \hat{\theta}_j^{+i}(k)\}}{1 + \exp\{\theta_j + \hat{\theta}_j^{-i}(k)\}} \quad \text{if } k < K + 1$$

$$w_{K+1} = \prod_{j=1}^m \frac{1 + \exp\{\theta_j + \hat{\theta}_j^{+i}(K)\}}{1 + \exp\theta_j}$$

and $\hat{\theta}_j^{+i}(k)$, $\hat{\theta}_j^{-i}(k)$ are calculated as in (5) but including and excluding \mathbf{y}_i in cluster k for the marginal probability calculation, respectively. Each weight w_k represents the relative probability of the data under $c(i) = k$.

With each resampling the number of clusters could increase by one, decrease by one, or remain unchanged, allowing the Markov chain to move around the space of clusters. If the mixing of such a Markov chain is too slow, it may be desirable to move around the space of clusters by proposing a change at more than just one value of $c(\cdot)$ at a time, for example, by splitting or merging clusters as suggested by Jain and Neal (2004) or Dahl (2003).

Sampling α : Fixed values of α and n provide a prior predictive distribution for the number of clusters K . To obtain different types of priors for K one can put a prior on α and include it as an unknown parameter in the MCMC scheme. As shown in Antoniak (1974), the distribution of K as a function of α is proportional to $\alpha^K \Gamma(\alpha) / \Gamma(\alpha + n)$. This can be highly skewed in $\alpha \in \mathbb{R}^+$ depending on K . Since K varies over the MCMC sampling procedure, coming up with a fixed proposal distribution for a Metropolis-Hastings update is problematic. Escobar and West (1998) provide a sampling approach based on data augmentation if the prior for α is a gamma distribution. Alternatively we reparameterize in terms of $\pi = \frac{\alpha}{\alpha+1} \in (0, 1)$, which represents the probability that a given pair of units will be in different clusters. Changing variables, we have

$$p(\pi|K) \propto p(\pi) \times \left(\frac{\pi}{1-\pi} \right)^K \frac{\Gamma[\pi/(1-\pi)]}{\Gamma[\pi/(1-\pi) + n]}$$

Sampling from $p(\pi|K)$ can be achieved by sampling from a grid on $(0, 1)$.

Sampling $\mu, \sigma^2, \theta, \psi$: Under the independence assumptions of (4), the full conditional distribution of $\omega_j = \{\mu_j, \sigma_j^2, \theta_j, \psi_j\}$ depends only on the observations from attribute j . It is relatively straightforward to update these parameters for each $j = 1, \dots, m$ using a Metropolis-Hastings step or a Gibbs step, in the case of conjugate models:

- Metropolis-Hastings: Sample ω_j^* from a proposal distribution $J(\omega_j^*|\omega_j)$, and accept the proposed value with probability

$$1 \wedge \left\{ \prod_{k=1}^K \frac{1 + e^{\theta_j^* + \hat{\theta}_j^*(k)}}{1 + e^{\theta_j^*}} \frac{1 + e^{\theta_j}}{1 + e^{\theta_j + \hat{\theta}_j(k)}} \times \prod_{i=1}^n \frac{p(y_{i,j}|\mu_j^*, \sigma_j^{2*})}{p(y_{i,j}|\mu_j, \sigma_j^2)} \times \frac{J(\omega_j|\omega_j^*)}{J(\omega_j^*|\omega_j)} \right\},$$

where $\hat{\theta}_j^*(k)$ is computed as in (5) under the proposed values of the parameters.

- Gibbs: If certain conjugate families are used, then conditional samples of $\omega = \{\mu, \sigma^2, \theta, \psi\}$ may be obtained by first sampling values of $\gamma_{(1)}, \dots, \gamma_{(K)}$ from their full conditional distribution, and then sampling ω from its conjugate full conditional. From the description of the Pólya urn sampling scheme in Section 2, one can see that the within-cluster values $\{\gamma_{(1)}, \dots, \gamma_{(K)}\}$ are a priori i.i.d. from $f_0(\gamma|\theta, \psi)$. As discussed in Neal (2000), conditional on the data, the values of $\gamma_{(k)}$ for each cluster k can be sampled from a distribution proportional to $f_0(\gamma|\theta, \psi) p(\{\mathbf{y}_i : c(i) = k\}|\gamma, \mu, \sigma^2)$. For example, if the distribution of the non-zero mean shifts is normal, a value of $\gamma_{(k),j}$ can be sampled from its full conditional by first sampling $s_{(k),j}$ from a Bernoulli distribution with log-odds $\theta_j + \hat{\theta}_j(k)$, and if $s_{(k),j} = 1$, then sampling $\delta_{(k),j}$ from the appropriate normal distribution and setting $\gamma_{(k),j} = s_{(k),j} \times \delta_{(k),j}$. Since the values $\{\gamma_{(1)}, \dots, \gamma_{(K)}\}$ represent i.i.d. samples from $f_0(\gamma|\theta, \psi)$, the conditional distribution of $\{\theta_j, \psi_j\}$ is computed via (4) and is given by

$$p(\theta_j, \psi_j|\gamma_{(1)}, \dots, \gamma_{(K)}) \propto p(\theta_j, \psi_j) \times \frac{\exp\{\theta_j \sum_{k=1}^K s_{(k),j}\}}{(1 + \exp\{\theta_j\})^K} \times \prod_{k:s_{(k),j}=1} g(\delta_{(k),j}|\psi_j).$$

Sampling from the full conditional of $\{\theta_j, \psi_j\}$ is straightforward assuming $p(\theta_j, \psi_j)$ is conjugate.

To sample from the conditional distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ given the clustering $c(\cdot)$ and $\gamma_{(1)}, \dots, \gamma_{(K)}$ we can compute $\xi_{i,j} = y_{i,j} - \gamma_{i,j}$, where $\gamma_i = \gamma_{(c(i))}$. The residual $\xi_{i,j}$ has conditionally a normal(μ_j, σ_j^2) distribution. Assuming conjugate priors for these parameters, we have

$$\mu_j \sim \text{normal}(\hat{\mu}_j, \hat{\sigma}_j^2), \text{ where } \hat{\sigma}_j^2 = (n/\sigma_j^2 + 1/v_j)^{-1}, \hat{\mu}_j = \hat{\sigma}_j^2(\sum_{i=1}^n \xi_{i,j}/\sigma_j^2 + m_j/v_j), \text{ and } m_j, v_j \text{ are the prior mean and variance of } \mu_j.$$

$$1/\sigma_j^2 \sim \text{gamma}[(\nu_0 + n)/2, (\nu_0\sigma_0^2 + \sum_{i=1}^n (\xi_{i,j} - \mu_{i,j})^2)/2], \text{ where } \nu_0, \sigma_0^2 \text{ are hyperparameters.}$$

3.2 Clustering behavior

We now consider evaluating what kinds of clusters the model selects. The conditional distribution of the clustering given the data and other parameters is proportional to

$$p(c(\cdot)|\mathbf{y}_1, \dots, \mathbf{y}_n, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\psi}, \alpha) \propto p(c(\cdot)|\alpha) \times p(\mathbf{y}_1, \dots, \mathbf{y}_n|c(\cdot), \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\psi})$$

The first term is the Pólya urn model for the clustering and is given by

$$p(c(\cdot)|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^K \prod_{k=1}^K \Gamma(n_k)$$

We see that α controls the number of clusters, with large α resulting in many clusters. The product of gamma functions counters this to some extent, by down-weighting the probability of having many small clusters, as $\Gamma(n_1 + n_2) \geq \Gamma(n_1)\Gamma(n_2)$.

The second term, or “likelihood” is given by (6), and as a function of $c(\cdot)$ is proportional to

$$\prod_{k=1}^K \prod_{j=1}^m \frac{1 + \exp\{\theta_j + \hat{\theta}_j(k)\}}{1 + \exp\{\theta_j\}}$$

The k, j th term in the product is an increasing function of $\hat{\theta}_j(k)$. Taking logs and using a Taylor series expansion of $\log(1 + e^{\theta_j + \hat{\theta}_j(k)})$ about $\hat{\theta}_j(k) = 0$, we have

$$\log p(\mathbf{y}_1, \dots, \mathbf{y}_n|c(\cdot), \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\tau}^2) \approx a + \sum_{k=1}^K \sum_{j=1}^m \frac{e^{\theta_j}}{1 + e^{\theta_j}} \hat{\theta}_j(k)$$

where a does not depend on $c(\cdot)$. Thus the data put weight on clusterings for which $\hat{\theta}_j(k)$ is large across attributes, each weighted by $\frac{e^{\theta_j}}{1 + e^{\theta_j}}$. Recall that the conditional log-odds that $s_{(k),j} = 1$ is $\theta_j + \hat{\theta}_j(k)$, so $\hat{\theta}_j(k)$ represents the effect of the data on the log-odds of there being a mean shift in

attribute j for group k . In the case where $g(\gamma_j|\psi_j)$ is normal with mean zero and variance τ_j^2 , and writing $\xi_{i,j} = y_{i,j} - \mu_j$, we have

$$\hat{\theta}_j(k) = \frac{1}{2} \left\{ \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2/n_k} \frac{n_k}{\sigma_j^2} [\bar{\xi}_j(k)]^2 + \log \frac{\sigma_j^2/n_k}{\sigma_j^2/n_k + \tau_j^2} \right\}$$

where $\bar{\xi}_j(k) = \sum_{i:c(i)=k} \xi_{i,j}/n_k$. Note that $\theta_j(k)$ is increasing in $[\bar{\xi}_j(k)]^2$, which is a measure of the within-group covariance of attribute j :

$$\begin{aligned} [\bar{\xi}_j(k)]^2 &= \frac{1}{n_k} \left[\frac{\sum_i \xi_{i,j}^2}{n_k} + (n_k - 1) \frac{\sum_{i1 \neq i2} \xi_{i1} \xi_{i2}}{n_k(n_k - 1)} \right] \\ &= \frac{1}{n_k} \text{vâr}(\xi_{i,j}) + \frac{n_k - 1}{n_k} \text{côv}(\xi_{i1,j}, \xi_{i2,j}) \end{aligned}$$

From a model-based point of view, $E([\bar{\xi}_j(k)]^2 | s_j) = \sigma_j^2/n_k + s_j \tau_j^2$, and so $[\bar{\xi}_j(k)]^2$ is an empirical approximation to $\sigma_j^2/n_k + \tau_j^2$ if cluster k has a mean shift at attribute j .

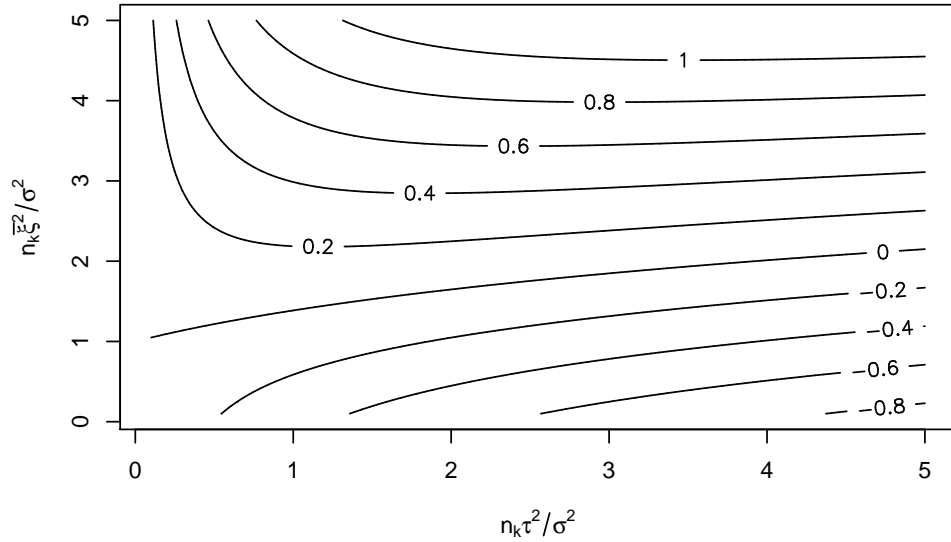


Figure 4: $\hat{\theta}_j(k)$ as a function of $n_k \tau_j^2 / \sigma_j^2$ and $n_k [\bar{\xi}_j(k)]^2 / \sigma_j^2$.

Figure 4 plots the dependence of $\hat{\theta}_j(k)$ on $n_k \tau_j^2 / \sigma_j^2$ and $n_k [\bar{\xi}_j(k)]^2 / \sigma_j^2$. Note that the impact of the data $[\bar{\xi}_j(k)]^2$ on $\hat{\theta}_j(k)$ increases as $n_k \tau_j^2 / \sigma_j^2$ increases. This is what we would expect - if the sampling variability σ_j^2/n_k of a mean attribute value within a cluster is large compared to the size of an expected mean shift τ_j^2 , then our data is not very informative and the odds of a mean shift are not altered. At the opposite extreme mean shifts should be easy to detect if they exist, and so our data should be quite informative about $s_{(k),j}$.

The values of the $\hat{\theta}_j(k)$'s also play a role in how the MCMC algorithm updates the values of $c(i)$, or equivalently, how the model would categorize a new unit based on its data \mathbf{y}_{n+1} . From (7), we see that given a clustering of the data into K groups, the probability that unit $n + 1$ is placed in cluster k is proportional to $n_k w_k$, and the probability it is placed in its own cluster is αw_{K+1} . Examining w_k further, we note that the contribution from attribute j is

$$w_{k,j} = \frac{1 + \exp\{\theta_j + \hat{\theta}_j^+(k)\}}{1 + \exp\{\theta_j + \hat{\theta}_j^-(k)\}},$$

where the $+$ and $-$ in the superscripts indicate $\hat{\theta}_j(k)$ being computed with and without \mathbf{y}_{n+1} respectively. Approximating both the log of the numerator and denominator with a first order Taylor series expansion gives

$$\log w_{k,j} \approx [\hat{\theta}_j^+(k) - \hat{\theta}_j^-(k)] \times \frac{e^{\theta_j}}{1 + e^{\theta_j}}$$

The log relative probability of the new unit being categorized into cluster k is then approximately

$$\log p(c(n+1) = k | \mathbf{y}_1, \dots, \mathbf{y}_{n+1}) \approx a + \log n_k + \sum_{j=1}^m \left\{ \frac{e^{\theta_j}}{1 + e^{\theta_j}} [\hat{\theta}_j^+(k) - \hat{\theta}_j^-(k)] \right\}$$

where we set $n_{K+1} = \alpha$ and $\hat{\theta}_j^-(K+1) = 0$, and a is a normalizing constant. Thus we see that the model puts higher probability on categories k for which the addition of the unit results in increases in the $\hat{\theta}_j(k)$'s, weighted by the θ_j 's. The model also favors putting a unit into larger clusters over smaller clusters, or placing a unit by itself if α is large.

3.3 A default methodology

The types of clusters identified by the above modeling strategy are determined in part by the model f_0 for the mean shifts, which in turn depends on the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. Ideally, one will have a good idea of what types of clusters to look for and will be able to specify prior values or distributions for some or all of these parameters. This will be the case for the example in Section 6, where we have a reasonable amount of information about the data-generating mechanism.

In situations where this is not the case, using diffuse priors and separately estimating θ_j and ψ_j for each attribute could lead to identifiability issues and slow convergence of the Markov chain. For example, the lack of large between-cluster variation in the mean of attribute j could be represented by either a large negative θ_j or a value of ψ_j making δ_j near zero with high probability. To avoid such issues we suggest the following simplified version of model (4):

$$f_0(\boldsymbol{\gamma}) = \prod_{j=1}^m \left\{ \left[\frac{e^{\theta}}{1 + e^{\theta}} \text{normal}(\gamma_j : 0, \eta\sigma_j^2) \right]^{(\gamma_j \neq 0)} \times \left[\frac{1}{1 + e^{\theta}} \right]^{(\gamma_j = 0)} \right\},$$

i.e. mean shifts across attributes and clusters are independent and γ_j is equal in distribution to a binary($\frac{e^{\theta}}{1+e^{\theta}}$) random variable multiplied by a normal($0, \eta\sigma_j^2$) random variable. This is the same as

the model discussed in 3.2 but with $\theta_j = \theta$ and $\tau_j^2/\sigma_j^2 = \eta$ fixed across attributes. In this reduced model, $e^\theta/(1+e^\theta)$ represents the fraction of attributes that will have a mean shift within a cluster, and η represents the average squared magnitude of the mean shifts of an attribute, relative to its variance.

Diffuse conjugate priors in this case are uniform on (0,1) for $e^\theta/(1+e^\theta)$ and inverse- χ_1^2 for η . For the clustering parameter α , a uniform prior on $\alpha/(\alpha+1)$ results in a prior predictive distribution function for K that is monotonically decreasing from $K=1$ but has a reasonably heavy tail. Finally, for default priors on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ we suggest ‘‘unit information’’ priors (Kass and Wasserman, 1995) whereby a priori we have $\mu_j \sim \text{normal}(\bar{y}_{\cdot,j}, s_j^2)$ and $\sigma_j^2 \sim \text{inverse gamma}(1/2, s_j^2/2)$. Although based on the data, such priors are proper and are probably less informative than any actual elicited priors a researcher would provide.

4 Simulation study

We further examine the behavior of the above clustering approach with a small simulation study. Datasets were generated as follows:

1. for $k = 1, \dots, K$, generate $\boldsymbol{\gamma}_{(k)} = \mathbf{s}_{(k)} \times \boldsymbol{\delta}_{(k)}$ via
 - $s_{(k),1} \dots, s_{(k),m} \sim \text{i.i.d. binomial}(e^\theta/(1+e^\theta))$, and
 - $\delta_{(k),1} \dots, \delta_{(k),m} \sim \text{i.i.d. normal}(0, \tau^2)$.
2. for $i = 1, \dots, n$, sample $c(i)$ uniformly from $\{1, \dots, K\}$ and set $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_{(c(i))}$.
3. for $i = 1, \dots, n$ sample $\mathbf{y}_i \sim \text{multivariate normal}(\boldsymbol{\gamma}_i, \mathbf{I})$.

Four datasets were generated using the above procedure, with $\{n, K\} \in \{50, 200\} \times \{5, 10\}$ and $m = 50$, $\theta = -1$, $\tau^2 = 1$ for each dataset. Two additional datasets were generated with $K = 1$ and $n \in \{50, 200\}$ with $\boldsymbol{\gamma}_{(1)} = \mathbf{0}$, i.e. no mean shift. To increase comparability across sample sizes, for $K \in \{5, 10\}$ the $n = 50$ and $n = 200$ datasets were generated with the same values of $\boldsymbol{\gamma}_{(1)}, \dots, \boldsymbol{\gamma}_{(K)}$. Furthermore, each $n = 50$ dataset was a subset of the corresponding $n = 200$ dataset. Data from the case $n = 50, K = 5$ are plotted in Figure 5.

For cluster estimation we used the default model and prior distributions described in Section 3.3 above. Two Markov chains of length 20,000 each were run for each dataset, one starting with all n units in the same cluster, the other with all n units in separate clusters. Convergence of the two chains to a common region of the parameter space was rapid, typically occurring in the first few thousand scans. At each scan of the Markov chain we computed $\log p(\mathbf{y}|c(), \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \theta, \eta) + \log p(c()|\alpha) + \log p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \theta, \eta, \alpha)$, which is proportional to the log posterior probability density of the parameters and clustering. The maximum a posteriori (MAP) estimate of the parameters was taken to be the set of parameters that maximized this quantity over all samples from the Markov

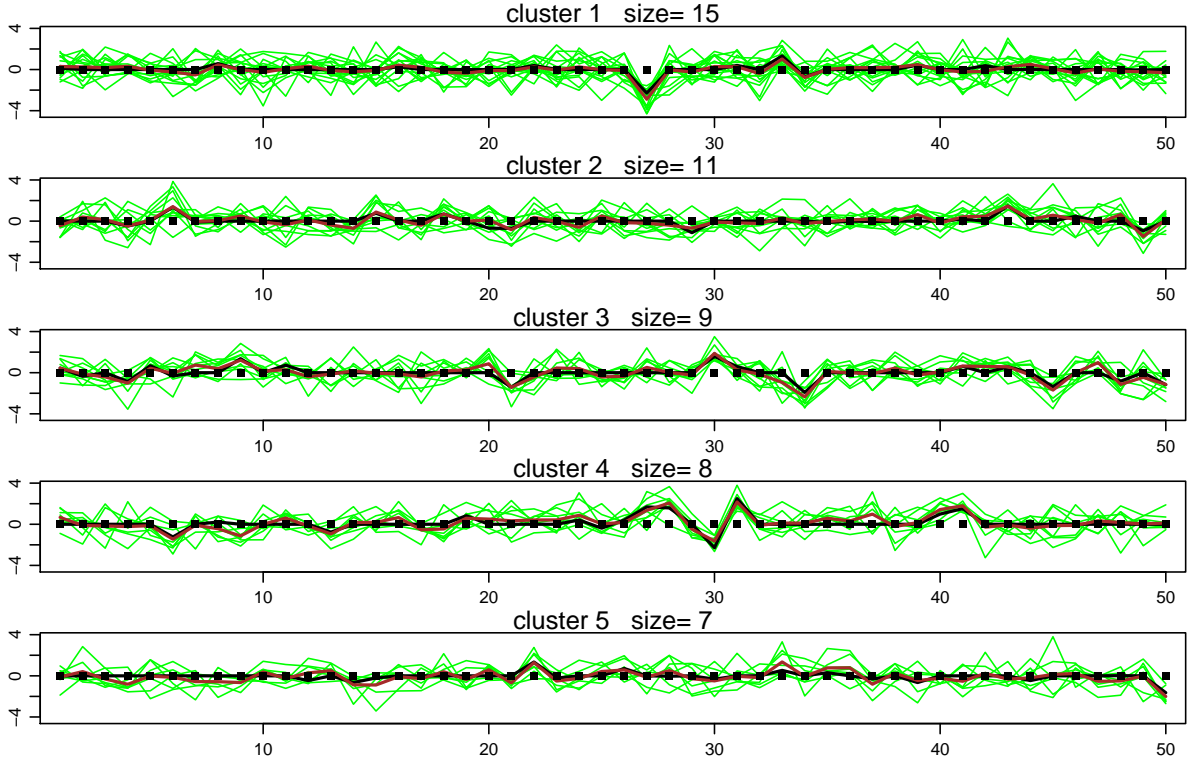


Figure 5: Data for the case $n = 50, K = 5$. Green lines are attribute profiles of the 50 objects, black lines give the value of γ in each cluster, and brown lines are the empirical means within a cluster.

chains. Such an estimate has been used before in Bayesian approaches to clustering problems - see for example Newton (2002) and Cheeseman and Stutz (1996). An alternative approach is to record the function $c()$ for each scan or subset of scans and then identify the most common clustering function. This latter approach may run into problems if m is very large or the posterior distribution over $c()$ is diffuse relative to the number of posterior samples obtained, in which case it may be possible that no clustering function $c()$ is sampled more than once.

The performance of these MAP estimates is detailed in Table 1. For each estimate of $\hat{c}()$ of $c()$, the purity of putative cluster k is given by

$$\max_{1 \leq k' \leq K} |\{i : c(i) = k' \text{ and } \hat{c}(i) = k\}| / |\{i : \hat{c}(i) = k\}|,$$

which is the maximum fraction of objects in a putative cluster having the same true cluster membership. Note that a cluster with just one object obtains the maximum purity level. In order to compare clusterings with different numbers of clusters, we report the purity of a clustering as the unweighted average purity of clusters having at least two members.

The proposed estimation procedure did reasonably well: The MAP estimates did not identify any clusters when there were none, and identified most of the clusters correctly in the other cases.

truth		(a)	(b)	(c)		
n	K	\hat{K}	purity	\hat{K}_m	purity $_m$	purity $_{\text{COSA}}$
50	1	1	-	1	-	-
50	5	4	0.79	11	0.76	0.47
50	10	10	0.83	32	0.37	0.31
200	1	1	-	1	-	-
200	5	5	0.94	6	0.92	0.40
200	10	10	0.93	10	0.87	0.27

Table 1: Performance of the posterior mode clustering for the six datasets: Column (a) is using the Dirichlet mixture of attribute ensembles, and (b) is using a standard Dirichlet mixture model. The results in column (c) are from average linkage dendrograms based on COSA distances, with clusters derived from on runt pruning (Steutzle 2003) with the true value of K assumed.

As we expect, the performance of the method improves as the sample size increases. The worst performing situation for the proposed method was when $n = 50, K = 5$. The data for this case are plotted in Figure 5, and the resulting clustering in Figure 6. The estimated clustering essentially misses cluster 5 of Figure 5, and places objects from this group into cluster 1 in Figure 6. Still, most of the other cluster features are estimated well.

The method compares favorably to other clustering approaches. Two other approaches of comparative interest are a Dirichlet process mixture model where all attribute means are different between clusters (essentially $\theta = \infty$), and the COSA algorithm of Friedman and Meulman (2004). The former is a submodel of the one proposed here, and will allow us to assess any advantage in allowing mean shifts to be identically zero. The latter method is very different algorithmically from the one in this paper, but has a similar clustering goal.

As shown in Table 1, the standard Dirichlet process mixture model generally overestimates the number of clusters for these simulated data. This is perhaps a result of the fact that any clustering for this model will tend to reduce the within-cluster variance σ_j^2 for *all* attributes j , which in turn could lead to overidentification of mean differences. To compare to the COSA algorithm, targeted COSA distances were computed between the n objects (with targets set to the 5 and 95% quantiles of the attribute distributions). Average linkage dendrograms were computed and K clusters were formed using runt pruning as described in Steutzle (2003). Note that this algorithm was provided with the true number of clusters. Even so, this approach based on COSA distances produced clusters having little to do with the original clustering. This is also apparent from the dendrogram plots based on the COSA distances (not shown). Other experiments with COSA suggest that it has difficulty identifying mean differences unless they are much larger than the noise variance, unlike the data generated here ($\sigma^2 = \tau^2 = 1$). Finally, model based clustering (Fraley and Raftery 2002)

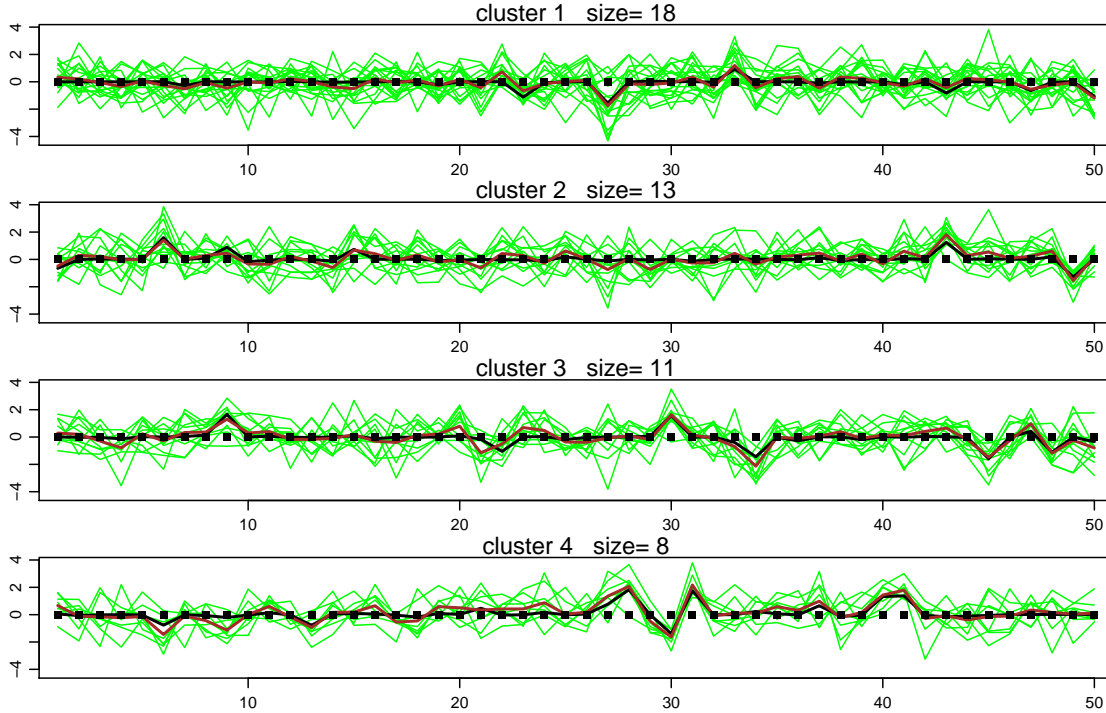


Figure 6: Estimated clustering for the case $n = 50, K = 5$.

was implemented on each dataset using the R-package `mclust`. The value of K maximizing the BIC was $K = 1$ for each of the six datasets, and so this method did not identify any of the clusters.

5 A model extension for chromosomal deletion and insertion

A model for the cellular evolution of cancer within an individual is that an accumulation of genetic abnormalities in certain chromosomal regions of a cell lineage eventually results in tumorigenesis. In particular, abnormalities may take the form of chromosomal gain or loss. A normal cell will have two copies of each chromosome, whereas cells having undergone errors in duplication may have lost or replicated certain sections of chromosomes, potentially resulting in one copy of chromosomal material at a given location (deletion) or more than two copies (insertion). If a cell lineage undergoes tumorigenesis, then such copy number changes will be passed on to the descendant cells that eventually make up a tumor. If several tumor cells from different, unrelated individuals all have the same types of deletion or insertion events at a combination of locations, then some of these locations may have a causal role in tumorigenesis. For this reason, it is of interest to determine the extent to which any tumors have similar patterns of response. This reasoning has been applied to several studies, including Hemminki (1997), Roylance et al. (1999), and is discussed in Gray and Collins (2000). Newton (2002) uses this idea to develop a model for binary genomic aberration data used to identify causal mechanisms of tumorigenesis.

In a study using comparative genomic hybridization (CGH) array data from 44 breast cancer tumors (detailed in Loo et al. 2004), $y_{i,j}$ is the log base 2 relative hybridization level of DNA from tumor i at genomic location j compared to the hybridization level of a normal cell's DNA at that location. Large negative or positive values of $y_{i,j}$ suggest deletion or insertion abnormalities, respectively, for tumor i at genetic location j . We analyze data from chromosomes 1, 6, 16, and 17, these being chromosomes of interest to the researchers who provided the data. Hybridization levels are measured at 345, 183, 150, and 133 locations on the four chromosomes respectively, for a total of $m = 811$ observations for each of the $n = 44$ tumors. Our data analysis goal is to determine if any of the 44 tumors have similar patterns of response along the four chromosomes.

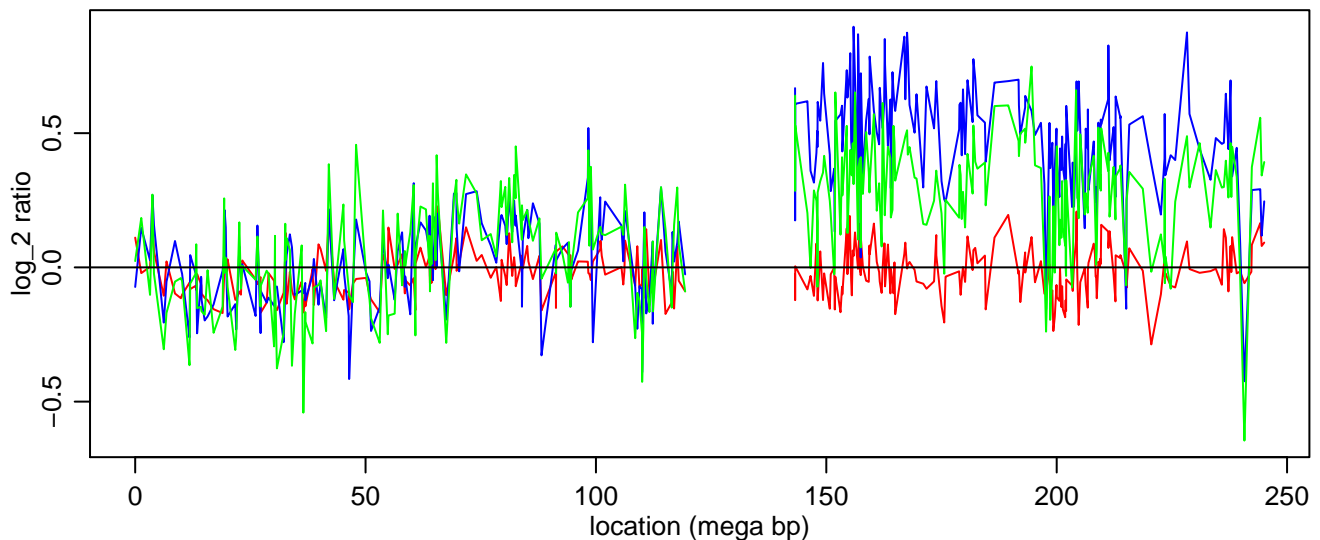


Figure 7: Hybridization ratio data from chromosome 1 for three sampled tumors. The gap is the centromere.

Figure 7 shows hybridization data from chromosome 1 for three tumors. The figure points out several features of the data:

- location specific means : Even though the lack of a genomic alteration should give a mean response near zero, the data suggest that responses at a single location can be consistently higher or lower than zero across tumors, suggesting that $\mu \neq \mathbf{0}$ for these data. These consistent deviations from zero are partly due to the fact that the mechanism used to measure hybridization at a given location (a location-specific bacterial artificial chromosome array, or BAC) is common across tumors at that location.
- clustering: The data from the three tumors show a similar pattern at the beginning of chromosome 1. In contrast, on the second part of the chromosome, two of the three show an amplified signal, and one shows little deviation from zero. This suggests there may be clus-

tering of the patterns of hybridization, and that the only certain subsets of the locations might be involved in the clustering.

- spatial correlation: Deletion and insertion events typically affect entire regions of a chromosome, thus the presence of a mean shift in the data at one location is likely to accompany a mean shift at nearby locations. This can be seen in the figure, where two of the three tumors plotted show an increase in response along most of the second-half of chromosome 1.

5.1 Modeling spatial genetic events

We retain the basic model outlined by (0) - (3) to analyze these data, but we develop a specific form for f_0 to account for the special features of these data described above and below. Below we describe a model for $\mathbf{s} \in \{0, 1\}^m$ and $\boldsymbol{\delta} \in \mathbb{R}^m$ and then model $\boldsymbol{\gamma}$ as equal in distribution to $\mathbf{s} \times \boldsymbol{\delta}$.

- Insertion and deletion events: Insertion and deletion events result in integer copy number changes in DNA, suggesting a mixture model for the magnitude of mean shifts. We use a mixture of three normal distributions, initially interpreting the mixture components as representing a copy number decrease of 1, a copy number increase of 1, and “everything else.” The model for $\boldsymbol{\delta}$ is then

$$f_{0\boldsymbol{\delta}}(\boldsymbol{\delta}) = \prod_{j=1}^m \left\{ \sum_{l=1}^L \lambda_l \times \text{normal}(\delta_j : \delta_{0,l}, \tau_l^2) \right\}$$

with $L = 3$. In this model, $\boldsymbol{\delta}_0 = \{\delta_{0,1}, \dots, \delta_{0,L}\}$ is initially interpreted as the average mean shift for each of the different copy number changes, and $\boldsymbol{\tau}^2 = \{\tau_1^2, \dots, \tau_L^2\}$ the variances.

- Spatial dependence of events: As described above, genetic events such as deletion or insertion can affect large regions of the genome, so we expect the presence or absence of mean shifts to be a spatially correlated process. We represent this statistically by modeling the presence of a mean shift with a Markov sequence:

$$f_{0\mathbf{s}}(\mathbf{s}) = \kappa(\boldsymbol{\theta}_1 \mathbf{1}, \boldsymbol{\theta}_2)^{-1} \exp\left\{ \theta_1 \sum_{j=1}^n s_j + \sum_{j=2}^m (\theta_{2,j} - \theta_1) s_j s_{j-1} \right\}$$

The parameters θ_1 and $\theta_{2,j}$ are the conditional log-odds that $s_{i,j} = 1$, a mean shift is present at j , given $s_{j-1} = 0$ and 1 respectively. We parameterize $\{\theta_{2,j}, j = 1, \dots, m\}$ in terms of two unknown parameters (a, b) and the known genetic distance $d_{j,j-1}$ (in mega-base pairs) between consecutive locations, so that $\theta_{2,j} = ae^{bd_{j,j-1}}$.

Our model for f_0 thus includes the following parameters to be estimated from the data: $\boldsymbol{\psi} = \{\boldsymbol{\lambda}, \boldsymbol{\delta}_0, \boldsymbol{\tau}^2\}$ for the magnitude of the mean shifts and $\boldsymbol{\theta} = \{\theta_1, a, b\}$ modeling the presence or absence of mean shifts.

5.2 Parameter estimation

Parameter estimation in this more complicated model is very similar to that described in 3.1. As before, a useful quantity for estimation is the marginal probability of the data in a cluster given the parameters and clustering. This is given by

$$p(\{\mathbf{y}_i : c(i) = k\} | c(\cdot), \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}, \boldsymbol{\psi}) = \left\{ \prod_{k=1}^K \frac{\kappa(\theta_1 \mathbf{1} + \hat{\boldsymbol{\theta}}(k), \boldsymbol{\theta}_2)}{\kappa(\theta_1 \mathbf{1}, \boldsymbol{\theta}_2)} \right\} \times \left\{ \prod_{j=1}^m \prod_{i:c(i)=k} \text{normal}(y_{i,j} : \mu_j, \sigma_j^2) \right\},$$

where $\hat{\boldsymbol{\theta}}(k) = \{\hat{\theta}_1(k), \dots, \hat{\theta}_m(k)\}$ and $\hat{\theta}_j(k)$ is as defined by (5). In the case of the L -component mixture model, we have

$$\hat{\theta}_j(k) = \log\left(\sum_{l=1}^L \lambda_l \exp\{\hat{\theta}_{j,l}(k)\}\right)$$

where $\hat{\theta}_{j,l}(k)$ are computed as in (5) under the L different normal models. Gibbs sampling of the cluster memberships proceeds as in Section 3.1 but with the following modification to the weights w_1, \dots, w_K :

$$\begin{aligned} w_k &= \frac{\kappa(\theta_1 \mathbf{1} + \hat{\boldsymbol{\theta}}^{+i}(k), \boldsymbol{\theta}_2)}{\kappa(\theta_1 \mathbf{1} + \hat{\boldsymbol{\theta}}^{-i}(k), \boldsymbol{\theta}_2)} \quad \text{if } k < K + 1, \\ w_K &= \frac{\kappa(\theta_1 \mathbf{1} + \hat{\boldsymbol{\theta}}^{+i}(k), \boldsymbol{\theta}_2)}{\kappa(\theta_1 \mathbf{1}, \boldsymbol{\theta}_2)} \end{aligned}$$

Finally, the parameters $\{\alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\psi}, \boldsymbol{\theta}\}$ are updated in the Markov chain by using a combination of the Metropolis-Hastings algorithm for a and b and the Gibbs sampler for $\alpha, \theta_1, \boldsymbol{\psi}, \boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$.

5.3 Prior distributions

The researchers who provided these data also performed a number of experiments which provide somewhat informative priors for this analysis. For some locations, hybridization measurements were taken under zero copy number changes, giving estimates of $\boldsymbol{\sigma}^2$. In ten samples of normal versus normal data (zero copy number change) along the 345 locations of chromosome 1, the variance of the responses was about .005. We represent this with an inverse-gamma (4, 2/100) distribution which is relatively diffuse but has a mode of .0067 and a mean of .0033. Experiments were also run under known copy number changes of -1 and +1, motivating normal prior distributions for $\delta_{0,1}$ and $\delta_{0,2}$ with means of $\{-.34, .38\}$ and variances of $\{0.025, .025\}$. For $\delta_{0,3}$, nominally representing all larger copy number changes, we used a normal (1, .010) prior. We used independent inverse-gamma(2,2/100) distributions for the variances of the mean shifts $\tau_1^2, \tau_2^2, \tau_3^2$. The prior for each μ_j was taken to be normal with mean zero and variance s_j^2 , similar to a unit information prior. The priors for the parameters $e^{\theta_1}/(1 + e^{\theta_1})$ and $\boldsymbol{\lambda}$ were taken to be uniform on the 2 and 3 dimensional simplexes, and the priors for a and b were diffuse. The prior for $\alpha/(\alpha + 1)$ was taken to be beta(1,1), which gives a prior predictive distribution for K that is monotonically decreasing from $K = 1$ with

$\Pr(K \leq 4) \approx .5$, but is heavy tailed so that there is some a priori probability that no tumors have a common response pattern ($K = 44$).

5.4 Posterior Inference

Two Markov chains as described above were run for 50,000 scans each, one starting with all tumors in the same cluster, and the other with each tumor in its own cluster. Output was saved every 10 scans. The Markov chains seemed to be in similar regions of the parameter space after about 10,000 scans. Several other shorter Markov chains, run from different starting values, also converged to the same region. Samples from the two long chains were tightly concentrated around an estimated posterior modal configuration having 26 “clusters”, 21 of which contained only a single tumor. The remaining 5 clusters were mostly small, containing 2 or 3 tumors, except for one cluster which contained 12 tumors. The data from this largest cluster are shown in Figure 8. Although it may

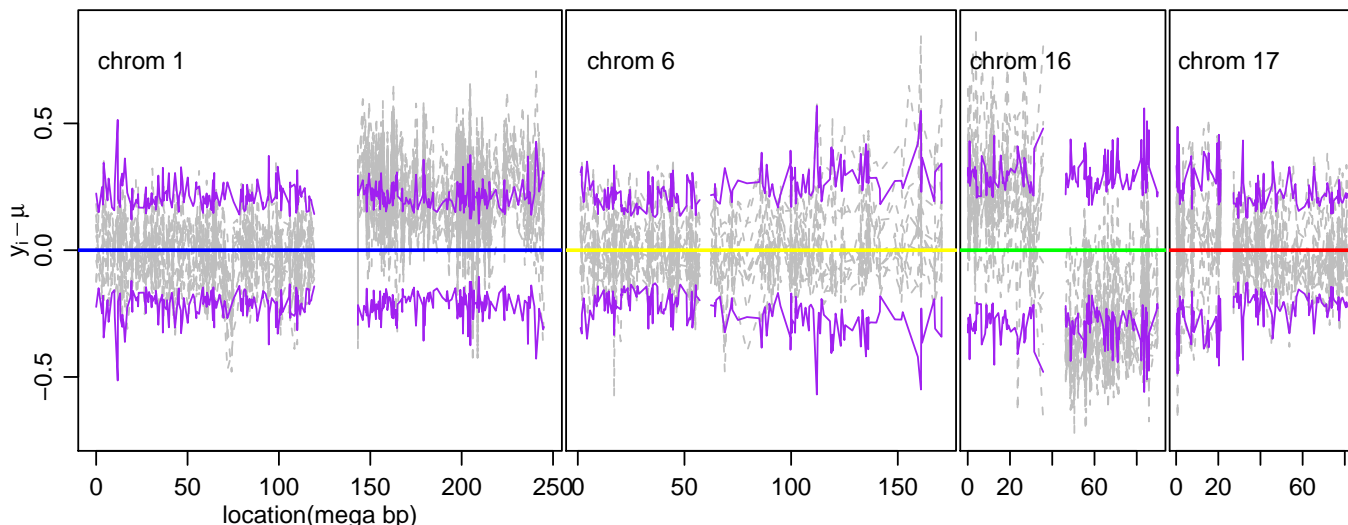


Figure 8: Raw data from an identified cluster of 12 tumors. The purple lines are $\pm 2 \times \sqrt{\hat{\sigma}^2}$.

be hard to discern from the plot, each tumor in this group of twelve has a response pattern which is high for most of the q-arm of chromosome 1, high for most of the p-arm of chromosome 16, and low for most the q-arm of chromosome 16. As shown in Figure 9, the centered mean $\bar{y}_{\cdot,j} - \hat{\mu}_j$ in these regions is generally well outside of the nominal confidence band of $2 \times \sqrt{\hat{\sigma}_j^2/12}$. The black line gives the expected value of the mean shift γ for this group, conditional on the data, the clustering, and MAP estimates of the other parameters. As we would expect, this conditional estimate approximates $\bar{y}_{\cdot,j} - \hat{\mu}_j$ when this quantity is large, and is zero (or close to it) otherwise.

Having such a large number of clusters is not surprising for these data. Tumor cells are highly genetically unstable, and so it is not unlikely that any given pair of tumor cells have different patterns of genomic abnormalities. The Pólya urn cluster model seems appropriate in this regard, as it allows each object to be in its own group, and so cluster-specific parameters will not be driven

by a few outlying observations.

As a side note, the researchers have categorized each tumor as estrogen receptor-negative (ER-negative) or ER-positive. ER-positive tumors tend to be less aggressive than ER-negative tumors, and respond less better to treatment. Twenty-nine of the 44 tumors in this dataset are ER-positive, whereas 11 of the 12 tumors in the identified group are ER-positive, suggesting that the response pattern in this group might identify a subclass of ER-positive tumors.

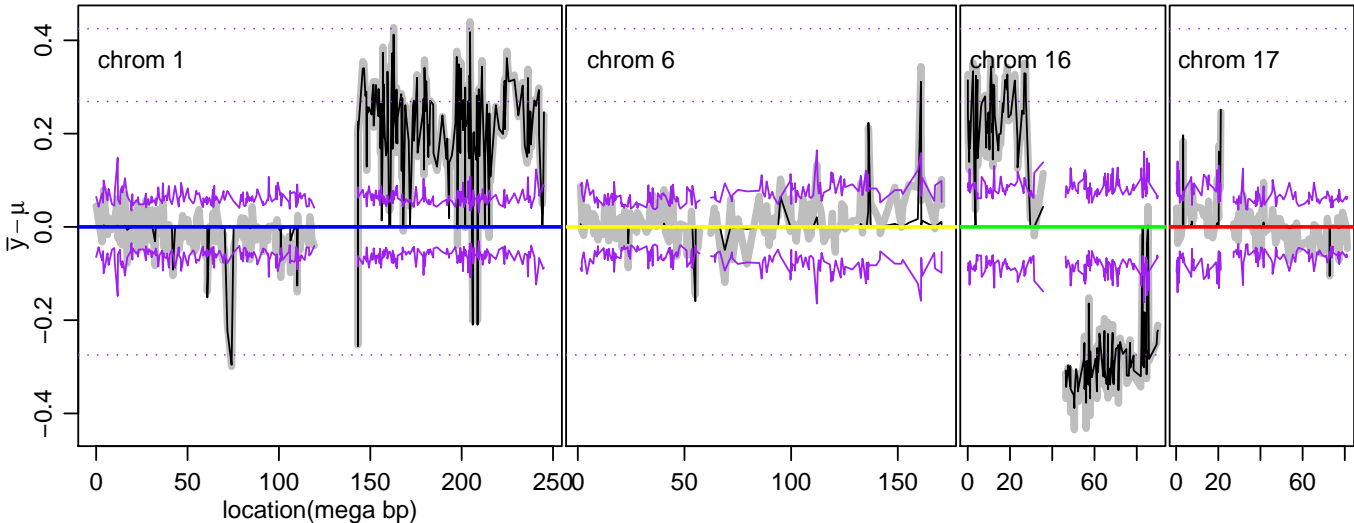


Figure 9: Mean response within the large cluster: The thick gray line is the mean of $y_i - \hat{\mu}$ over the 12 tumors. The purple lines are $\pm 2 \times \sqrt{\hat{\sigma}^2/12}$. The black line is the conditional expectation of the mean shift γ for this group, conditional on the data, the clustering and the MAP estimates of the other parameters.

6 Discussion

We have presented a model-based method of finding clusters based on subsets of attributes. The types of clusterings this method provides include cases where all attributes differ among clusters and where a fixed subset of attributes all differ among clusters, and so this approach is more general than a variable selection procedure. The method has several features shared by all Dirichlet process mixture models: The procedure generates a posterior distribution that potentially puts an object in a cluster by itself if its attribute pattern is not minimally similar to those of other objects. Depending on the application, this could be a desirable feature: Outlying observations can be identified and will not greatly influence the features of other clusters.

The basic method presented here for identifying clusters of attribute ensembles is extendable to more general data analysis situations. For example, heteroscedasticity can be accommodated by modeling data in cluster k with mean $\mu_{(k)} = \mu + s_{(k)} \times \delta_{(k)}$ and log-variance $\log \sigma^2 + s_{(k)} \times \zeta_{(k)}$,

which allows for a change in mean and variance at each attribute that defines a cluster. Non-normal data can be modeled with exponential family distributions, with $\boldsymbol{\mu}_{(k)} = \boldsymbol{\mu} + \mathbf{s}_{(k)} \times \boldsymbol{\delta}_{(k)}$ being used to model the canonical parameters.

The main disadvantage of such a model based approach is the amount of computation time required to obtain parameter estimates and measures of uncertainty. An analysis of 1000 attributes for 100 objects may take days to run on a standard unix-based PC using the computer code implemented in this paper (a combination of R and C). Although better code could perhaps decrease computing time, the method would still be infeasible as a quick exploratory data analysis tool for very large datasets. However the method could be useful in tandem with other, quicker, clustering procedures. For example, one could use a fast heuristic procedure to identify a potential clustering, and then use the model based procedure to see if the potential clustering is a strong or weak mode, if there is another mode nearby, or to identify ensembles of attributes relevant to the clustering.

References

- Antoniak, C. E. (1974), “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems,” *Ann. Statist.*, 2, 1152–1174.
- Blackwell, D. and MacQueen, J. B. (1973), “Ferguson distributions via Pólya urn schemes,” *Ann. Statist.*, 1, 353–355.
- Cheeseman, P. and Stutz, J. (1996), “Bayesian Classification (AutoClass): Theory and Results,” in *Advances in Knowledge Discovery and Data Mining*, eds. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, pp. 153–180, Cambridge MA: AAAI/MIT Press.
- Dahl, D. B. (2003a), “An improved merge-split sampler for conjugate Dirichlet process mixture model,” Technical report no. 1086, Department of Statistics, University of Wisconsin-Madison.
- Dahl, D. B. (2003b), “Modeling differential gene expression using a Dirichlet Process mixture model,” in *Proceedings of the American Statistical Association, Bayesian Statistical Sciences Section*, American Statistical Association, Alexandria, VA.
- Escobar, M. D. and West, M. (1998), “Computing nonparametric hierarchical models,” in *Practical nonparametric and semiparametric Bayesian statistics*, vol. 133 of *Lecture Notes in Statist.*, pp. 1–22, Springer, New York.
- Fraley, C. and Raftery, A. E. (2002), “Model-based clustering, discriminant analysis, and density estimation,” *J. Amer. Statist. Assoc.*, 97, 611–631.
- Friedman, J. H. and Meulman, J. J. (2004), “Clustering objects on subsets of attributes,” *Journal of the Royal Statistical Society*, to appear.

- Gray, J. and Collins, C. (2000), “Genome changes and gene expression in human solid tumors,” *Carcinogenesis*, 21, 443–452.
- Hemminki, A., Tomlinson, I., Markie, D., Järvinen, H., Sistonen, P., Björkqvist, A.-M., Knuutila, S., Reijo, S., Bodmer, W., Shibata, D., de la Chapelle, A., and Aaltonen, L. (1997), “Localization of a susceptibility locus for Peutz-Jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis,” *Nature Genetics*, 15, 87–90.
- Jain, S. and Neal, R. M. (2004), “A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model,” *Journal of Computational and Graphical Statistics*, 13, 158–182.
- Kass, R. E. and Wasserman, L. (1995), “A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion,” *J. Amer. Statist. Assoc.*, 90, 928–934.
- Loo, L., Grove, D., Neal, C., Williams, E., Cousens, L., Schubert, E., Holcomb, I., Massa, H., Glogovac, J., Li, C., Malone, K., Daling, J., Delrow, J., Trask, B., Hsu, L., and Porter, P. (2004), “Array CGH analysis of genomic alterations in breast cancer sub-types,” *Submitted*.
- MacEachern, S. N. (1994), “Estimating normal means with a conjugate style Dirichlet process prior,” *Comm. Statist. Simulation Comput.*, 23, 727–741.
- MacEachern, S. N. and Müller, P. (1998), “Estimating mixture of Dirichlet process models,” *Journal of Computational and Graphical Statistics*, 7, 223–238.
- McLachlan, G. J. and Basford, K. E. (1988), *Mixture models*, vol. 84 of *Statistics: Textbooks and Monographs*, Marcel Dekker Inc., New York, Inference and applications to clustering.
- Mitchell, T. J. and Beauchamp, J. J. (1988), “Bayesian variable selection in linear regression,” *J. Amer. Statist. Assoc.*, 83, 1023–1036, With comments by James Berger and C. L. Mallows and with a reply by the authors.
- Neal, R. M. (2000), “Markov chain sampling methods for Dirichlet process mixture models,” *J. Comput. Graph. Statist.*, 9, 249–265.
- Newton, M. A. (2002), “Discovering combinations of genomic aberrations associated with cancer,” *J. Amer. Statist. Assoc.*, 97, 931–942.
- Newton, M. A., Yang, H., Gorman, P. A., Tomlinson, I., and Roylance, R. R. (2003), “A statistical approach to modeling genomic aberrations in cancer cells,” in *Bayesian statistics, 7 (Tenerife, 2002)*, pp. 293–305, Oxford Univ. Press, New York, With a discussion by Scott C. Schmidler and a reply by the authors.

Parsons, L., Haque, E., and Liu, H. (2004), “Evaluating Subspace Clustering Algorithms,” in *Workshop on Clustering High Dimensional Data and its Applications, SIAM International Conference on Data Mining (SDM 2004)*, pp. 48–56.

Roylance, R., Gorman, P., Harris, W., Liebmann, R., Barnes, D., Hanby, A., and Sheer, D. (1999), “Comparative Genomic Hybridization of Breast Tumors Stratified by Histological Grade Reveals New Insights into the Biological Progression of Breast Cancer,” *Cancer Research*, 59, 1433–1436.