

Subset clustering of binary sequences, with an application to genomic abnormality data

Peter D. Hoff *

Technical Report no. 456

Department of Statistics

University of Washington

July 27, 2004

Abstract

This article develops a model-based approach to clustering multivariate binary data, in which the attributes that distinguish a cluster from the rest of the population may depend on the cluster being considered. The clustering approach is based on a multivariate Dirichlet process mixture model, which allows for the estimation of the number of clusters, the cluster memberships, and the cluster-specific parameters in a unified way. Such a clustering approach has applications in the analysis of genomic abnormality data, in which the development of different types of tumors may depend on the presence of certain abnormalities at subsets of locations along the genome. Additionally, such a mixture model provides a nonparametric estimation scheme for dependent sequences of binary data.

Some key words: genetic pathway, correlated binary data, nonparametric Bayes, unsupervised learning.

1 Introduction

In this article we consider a model-based approach to clustering multivariate binary data based on subsets of attributes. The data we consider are m -dimensional vectors \mathbf{y}_i representing measurements on m binary attributes for each member of a population of units $i = 1, \dots, n$. A simple cluster model for such data might be to assume that vectors of observations from units within a common cluster k are i.i.d. from a distribution $p(\mathbf{y}_i | \boldsymbol{\theta}_{(k)}) = \prod_{j=1}^m \theta_{(k),j}^{y_{i,j}} (1 - \theta_{(k),j})^{1-y_{i,j}}$, where the vector $\boldsymbol{\theta}_{(k)} = \{\theta_{(k),1}, \dots, \theta_{(k),m}\}$ represents the attribute rates for cluster k . A standard clustering

*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.

approach would be to estimate every attribute rate separately for each cluster. However in many applications only a subset of the attributes may be relevant in defining any given cluster, and the attributes that are relevant may vary from cluster to cluster. As a particular example, consider genomic abnormality data on a population of tumors, where $y_{i,j}$ is the indicator that tumor i has an abnormality at genomic location j . A model for the cellular evolution of cancer is that a cell lineage undergoes tumorigenesis if it accumulates genetic abnormalities at certain combinations of locations along the genome. Such combinations of abnormalities are sometimes called *pathways* to tumorigenesis. Abnormalities in a pathway, along with any other abnormalities the cell lineage accumulates, are passed on to the descendant cells that eventually make up a tumor. If the tumor cell genomes of several different, unrelated individuals all have abnormalities at a common set of locations, then this is some evidence that these locations play a role in tumorigenesis. Early data analysis methods identified abnormalities of interest based on having high marginal rates of occurrence (Brodeur et al. 1982; Newton, Wu and Reznikoff 1994). Recognizing that abnormalities in a common pathway will be correlated, Desper et al. (1999) looked beyond marginal rates by computing distances between abnormalities based on empirical correlations, and then applying a distance-based clustering algorithm. For a more detailed review of statistical methods for data of this type, see Newton (2002).

Since there are many different pathways to tumorigenesis, several may be represented in any given population of tumor cells from different individuals. Therefore, we might expect abnormality data $\mathbf{y}_1, \dots, \mathbf{y}_n$ to form clusters if some of the tumors share common pathways. Given a clustering of the data into K groups, if a cluster k consists of tumors sharing a common pathway then we might expect $\theta_{(k),j}$, the within-cluster rate of abnormality j , to be large if j is in the pathway or is near an abnormality in the pathway. On the other hand, if j is not in the pathway for cluster k or near a pathway abnormality, then we would expect $\theta_{(k),j}$ to be equal to some background abnormality rate θ_j . This motivates a parameterization of cluster-specific rates as $\boldsymbol{\theta}_{(k)} = \mathbf{r}_{(k)} \times \tilde{\boldsymbol{\theta}}_{(k)} + (1 - \mathbf{r}_{(k)}) \times \boldsymbol{\theta}$, where $\mathbf{r}_{(k)} \in \{0, 1\}^m$ is a binary vector indicating the “relevance” of each abnormality to cluster k and “ \times ” indicates element-wise multiplication. If $r_{(k),j} = 1$ then abnormality j is relevant for cluster k , and the cluster rate $\theta_{(k),j}$ for this attribute differs from the baseline rate θ_j . If $r_{(k),j} = 0$ then abnormality j is not relevant for this cluster and the cluster rate is the baseline rate. Presuming independence across these attributes within a cluster, this parameterization can be represented by the following model for a data vector \mathbf{y}_i of a unit i in cluster k :

$$\Pr(\mathbf{y}_i | \boldsymbol{\theta}, \mathbf{r}_{(k)}, \tilde{\boldsymbol{\theta}}_{(k)}) = \left\{ \prod_{j:r_{(k),j}=0} \theta_j^{y_{i,j}} (1 - \theta_j)^{1-y_{i,j}} \right\} \times \left\{ \prod_{j:r_{(k),j}=1} \tilde{\theta}_{(k),j}^{y_{i,j}} (1 - \tilde{\theta}_{(k),j})^{1-y_{i,j}} \right\}. \quad (1)$$

This model was considered by Patrikainen and Mannila (2004), who developed an estimation method for a fixed number of clusters K based on an EM-like algorithm and iterative χ^2 -significance tests. An approach more specific to the genomic abnormality data mentioned above has been taken by Newton (2002), who derived a statistical model for binary genomic abnormality data based on a

causal model relating the genomic abnormalities to tumorigenesis. His model is very parsimonious, presuming that no two clusters have overlapping sets of relevant abnormalities and that the background rates are constant across abnormalities. The first restriction seems generally undesirable, and the second may or may not be appropriate depending on the application: Some genomic regions are not necessary for the survival of certain types of cells, while others regions are, regardless of their involvement in pathways. As a result, a population of genetically unstable tumor cells might have background abnormality rates that vary across genomic locations, including locations that are not involved in pathways to tumorigenesis.

This article develops a general methodology for clustering binary data on subsets of attributes, based on a Dirichlet process mixture distribution for the parameters in model (1). This approach is similar to a model-based clustering method for continuous data developed in Hoff (2004). For general types of data, the identification of clusters based on subsets of attributes has been called “subspace” clustering, for which there are several non-model-based approaches: see for example Friedman and Meulman (2004), or Parsons, Haque and Liu (2004) for a review. Unlike other methods, the binary clustering approach presented in this paper allows for estimation of both the number of clusters K , the cluster memberships, and the relevant attributes in a unified way. Additionally, the model-based approach allows for an assessment of uncertainty in the estimated clustering and the incorporation of any available prior information into the data analysis procedure. For example, in an analysis of genomic abnormality data we will use a conditional prior distribution on $\tilde{\theta}$ to focus the search for clusters by assuming that the abnormality rates at relevant locations are higher than the background rates.

In the next section we develop a clustering approach from a Dirichlet process mixture model for binary sequences. A Markov chain Monte Carlo algorithm for parameter estimation is presented in Section 3, and the behavior of the cluster model is discussed in Section 4. It is seen that the ability of the model to identify accurate clusters depends on the number of relevant attributes within a cluster and the magnitude of the differences between the within-cluster and background rates. In Section 5 we apply the method to the analysis of genomic abnormality data from a set of 116 renal cell carcinomas, and compare the results to those of Newton (2002). A discussion follows in Section 6.

2 Mixture modeling of binary sequences

We derive our subset cluster model from a mixture model for binary sequences. We assume the vectors $\mathbf{y}_i \in \{0, 1\}^m$, $i = 1, \dots, n$, are independent samples from some population distribution parameterized in terms of a mixture of independent binary distributions:

$$p(\mathbf{y}_i | \boldsymbol{\theta}) = \int \left\{ \prod_{j=1}^m \theta_{i,j}^{y_{i,j}} (1 - \theta_{i,j})^{1-y_{i,j}} \right\} q(d\boldsymbol{\theta}_i | \boldsymbol{\theta}), \quad (2)$$

where q is a mixing distribution over $\boldsymbol{\theta}_i$ -values, conditional on a baseline set of rates $\boldsymbol{\theta}$. There is no loss of generality in modeling binary sequences with a mixture of independent sequences, as every distribution has such a representation: For example, a trivial representation of a given $p(\cdot|\boldsymbol{\theta})$ is obtained via $q(\boldsymbol{\theta}_i|\boldsymbol{\theta}) = \sum_{\mathbf{y} \in \{0,1\}^m} p(\mathbf{y}|\boldsymbol{\theta}) \delta_{\mathbf{y}}(\boldsymbol{\theta}_i)$, where $\delta_{\mathbf{y}}(\boldsymbol{\theta}_i)$ is the indicator that $\boldsymbol{\theta}_i = \mathbf{y}$.

Restricting q to be a discrete distribution allows mixture modeling to be viewed as a clustering procedure. As is well known, $\mathbf{y}_1, \dots, \mathbf{y}_n \sim \text{i.i.d. } p$ is equivalent to sampling $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n \sim \text{i.i.d. } q$ and then sampling $\mathbf{y}_i \sim \text{binary}(\boldsymbol{\theta}_i)$. If q is discrete then there is some probability that $\boldsymbol{\theta}_{i_1} = \boldsymbol{\theta}_{i_2}$, in which case we say units i_1 and i_2 are in the same ‘‘cluster’’. A nonparametric Bayesian approach to the clustering procedure would define a prior distribution on q such that the support of the prior is the space of all discrete probability distributions on $\boldsymbol{\theta}$. The simplest such prior for q is the Dirichlet process prior, which is parameterized by a clustering parameter $\alpha \in \mathbb{R}^+$ and a baseline probability distribution q_0 . In this case, the model (2) is called a Dirichlet process mixture model (Antoniak 1974; MacEachern 1994). That this mixture model gives an interpretable model for a clustering process can be seen via the Pólya urn representation of a sample from a Dirichlet process, as described in Blackwell and MacQueen (1973). In our context, if $q \sim \text{Dir}(\alpha, q_0)$ and $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ are i.i.d. samples from q , then unconditional on q the joint distribution of the $\boldsymbol{\theta}_i$ ’s is equal to that of an exchangeable sequence generated as follows:

1. sample $\boldsymbol{\theta}_1 \sim q_0$;
2. sample $\boldsymbol{\theta}_2 \sim \frac{\alpha}{\alpha+1}q_0 + \frac{1}{\alpha+1}\delta_{\boldsymbol{\theta}_1}$;
- ⋮
- n . sample $\boldsymbol{\theta}_n \sim \frac{\alpha}{\alpha+n-1}q_0 + \frac{n-1}{\alpha+n-1}\hat{q}_{n-1}$,

where \hat{q}_{n-1} is the empirical distribution of $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n-1}$. The above process is called a Pólya urn scheme with parameters α and q_0 . As can be seen, the sample $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ may have been generated by fewer than n draws from q_0 , and so the $\boldsymbol{\theta}$ -values form ‘‘clusters.’’ We denote the number of draws from q_0 as K , and the values of the draws as $\boldsymbol{\theta}_{(1)}, \dots, \boldsymbol{\theta}_{(K)}$. We define the cluster membership function c to be the function mapping the unit labels $\{1, \dots, n\}$ to the independent draws $\{1, \dots, K\}$. From the description of the Pólya urn scheme above, it is clear that the parameter α determines the distribution of the number of clusters K , whereas q_0 determines the distribution of the cluster-specific rates $\boldsymbol{\theta}_{(1)}, \dots, \boldsymbol{\theta}_{(K)}$.

To allow for the possibility that differences between clusters are based on subsets of attributes, we choose a particular form for q_0 so that a sample $\boldsymbol{\theta}_{(k)}$ takes the form as described by (1), i.e. $\boldsymbol{\theta}_{(k)} = \mathbf{r}_{(k)} \times \tilde{\boldsymbol{\theta}}_{(k)} + (\mathbf{1} - \mathbf{r}_{(k)}) \times \boldsymbol{\theta}$ for some vectors $\tilde{\boldsymbol{\theta}}_{(k)} \in [0, 1]^m$ and $\mathbf{r}_{(k)} \in \{0, 1\}^m$. Conditional on a baseline rate vector $\boldsymbol{\theta}$, we parameterize q_0 as

$$q_0(\boldsymbol{\theta}_{(k)}|\boldsymbol{\theta}) = \prod_{j=1}^m \left\{ \left[\frac{e^{\lambda_j}}{1 + e^{\lambda_j}} g_j(\boldsymbol{\theta}_{(k),j}) \right]^{(\boldsymbol{\theta}_{(k),j} \neq \boldsymbol{\theta}_j)} \times \left[\frac{1}{1 + e^{\lambda_j}} \right]^{(\boldsymbol{\theta}_{(k),j} = \boldsymbol{\theta}_j)} \right\},$$

where each g_j is a probability density on $[0, 1]$. Each term in the above product is a density with respect to a measure $\nu_j(A) = \text{Leb}(A) + \delta_{\theta_j}(A)$, i.e. Lebesgue measure plus a point-mass at θ_j . To put it more simply, each $\theta_{(k),j}$ can be obtained by sampling a binary random variable $r_j \sim \text{binary}\{e^{\lambda_j}/(1 + e^{\lambda_j})\}$ and a continuous random variable $\tilde{\theta}_j \in [0, 1]$ from g_j and setting $\theta_{(k),j} = r_j\tilde{\theta}_j + (1 - r_j)\theta_j$. An alternative representation of the same model is where q_0 is the equivalent distribution on pairs of vectors $\{\mathbf{r}, \tilde{\boldsymbol{\theta}}\} \in \{0, 1\}^m \times [0, 1]^m$, so that

$$q_0(\mathbf{r}, \tilde{\boldsymbol{\theta}}) = \prod_{j=1}^m \left(\frac{e^{\lambda_j r_j}}{1 + e^{\lambda_j}} g_j(\tilde{\theta}_j) \right). \quad (3)$$

In this case, the clustering model can be described as:

$$\{\mathbf{r}_1, \tilde{\boldsymbol{\theta}}_1\}, \dots, \{\mathbf{r}_n, \tilde{\boldsymbol{\theta}}_n\} \sim \text{Pólya urn}(\alpha, q_0) \quad (4)$$

$$\boldsymbol{\theta}_i = \mathbf{r}_i \times \tilde{\boldsymbol{\theta}}_i + (1 - \mathbf{r}_i) \times \boldsymbol{\theta} \quad (5)$$

$$p(\mathbf{y}_i | \boldsymbol{\theta}_i) = \prod_{j=1}^m \theta_{i,j}^{y_{i,j}} (1 - \theta_{i,j})^{1 - y_{i,j}} \quad (6)$$

Again, recall that the Pólya urn scheme potentially produces ties among the latent variables $\{\mathbf{r}_1, \tilde{\boldsymbol{\theta}}_1\}, \dots, \{\mathbf{r}_n, \tilde{\boldsymbol{\theta}}_n\}$ which in turn induces ties among $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$. Additionally, the values of $\theta_{i,j}$ for $j = 1, \dots, m$ will equal the baseline rates for some subset of attributes $\{j : r_{i,j} = 0\}$, and differ from the baseline rates at attributes $\{j : r_{i,j} = 1\}$. In this way the model and prior allow for the possibility that the units are clustered based on subsets of the attributes.

Useful for understanding the above model is the the log-likelihood, or log-probability of the data given the clustering c and the baseline parameter $\boldsymbol{\theta}$. This can be derived by first examining the probability of the data in a given cluster k having cluster-specific latent variables \mathbf{r} and $\tilde{\boldsymbol{\theta}}$:

$$\Pr(\{\mathbf{y}_i : c(i) = k\} | \mathbf{r}_{(k)} = \mathbf{r}, \tilde{\boldsymbol{\theta}}_{(k)} = \tilde{\boldsymbol{\theta}}) = \left\{ \prod_{j:r_j=0} \theta_j^{n_{k,j,1}} (1 - \theta_j)^{n_{k,j,0}} \right\} \times \left\{ \prod_{j:r_j=1} \tilde{\theta}_j^{n_{k,j,1}} (1 - \tilde{\theta}_j)^{n_{k,j,0}} \right\}$$

where $n_{j,k,1} = \sum_{i:c(i)=k} y_{i,j}$ and $n_{j,k,0} = \sum_{i:c(i)=k} (1 - y_{i,j})$. Here we are implicitly conditioning on c and $\boldsymbol{\theta}$. Using model (3) for the within-cluster latent variables \mathbf{r} and $\tilde{\boldsymbol{\theta}}$, the marginal distribution for the data from attribute j in cluster k can be obtained by summing/integrating over the latent variables:

$$p(\{y_{i,j} : c(i) = k\}) = \sum_{r_j=0}^1 \frac{e^{\lambda_j r_j}}{1 + e^{\lambda_j}} \int p(\{y_{i,j} : c(i) = k\} | r_j, \tilde{\theta}_j) g_j(\tilde{\theta}_j) d\tilde{\theta}_j.$$

Integrating over $\tilde{\theta}_j$ for both $r_j = 1$ and $r_j = 0$, let

$$\hat{\lambda}_j(k) = \log \frac{p(\{y_{i,j} : c(i) = k\} | r_{(k),j} = 1)}{p(\{y_{i,j} : c(i) = k\} | r_{(k),j} = 0)}. \quad (7)$$

We then have

$$p(\{y_{i,j} : c(i) = k\}) = \frac{1 + e^{\lambda_j + \hat{\lambda}_j(k)}}{1 + e^{\lambda_j}} \times p(\{y_{i,j} : c(i) = k\} | r_{(k),j} = 0).$$

Taking the product over all clusters k and attributes j gives

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n | c, \boldsymbol{\theta}) = \left\{ \prod_{k=1}^K \prod_{j=1}^m \frac{1 + e^{\lambda_j + \hat{\lambda}_j(k)}}{1 + e^{\lambda_j}} \right\} \times \left\{ \prod_{j=1}^m \prod_{i=1}^n \theta_j^{y_{i,j}} (1 - \theta_j)^{1 - y_{i,j}} \right\} \quad (8)$$

Note that the second term on the right-hand side does not involve the clustering, and is just the probability of the data under the baseline rates.

In the case where g_j is a beta(\tilde{a}_j, \tilde{b}_j) density, we have

$$\begin{aligned} \hat{\lambda}_j(k) &= \log \int \tilde{\theta}^{n_{k,j,1}} (1 - \tilde{\theta})^{n_{k,j,0}} g_j(\tilde{\theta}) d\tilde{\theta} - n_{k,j,1} \log \theta_j + n_{k,j,0} \log(1 - \theta_j) \\ &= \log \frac{\text{Beta}(\tilde{a}_j + n_{k,j,1}, \tilde{b}_j + n_{k,j,0})}{\text{Beta}(\tilde{a}_j, \tilde{b}_j)} - \log \theta_j^{n_{k,j,1}} (1 - \theta_j)^{n_{k,j,0}}. \end{aligned}$$

This is a log Bayes factor testing $H : p(y_j = 1) = \theta_j$ versus H^c for the data in cluster k , and represents the evidence in the data that attribute j is relevant for cluster k . Quantities (7) and (8) are useful for parameter estimation as discussed in the next section, and also for understanding the behavior of the model as discussed in Section 4.

3 Parameter estimation

The parameters in the likelihood (8) to be estimated include the clustering c and the baseline rates $\boldsymbol{\theta}$. Posterior inference for these parameters can be made by constructing approximate samples from their joint posterior distribution via a Markov chain Monte Carlo algorithm. In general we will also want to include α in our Markov chain: A fixed value of α gives a prior predictive distribution for the number of clusters K that is relatively concentrated around a single value. A wider variety of prior predictive distributions for K can be obtained by putting a prior on α and including it as an unknown parameter in the MCMC scheme. For example, a uniform distribution on $\alpha/(\alpha + 1)$ induces a prior predictive distribution on K that generally has a maximum at $K = 1$ but a heavy tail out to $K = n$.

We suggest the following MCMC algorithm which iteratively resamples values of these parameters conditional on each other and the data. Given values $\{c^s, \boldsymbol{\theta}^s, \alpha^s\}$ at scan s of the Markov chain, $\{c^{s+1}, \boldsymbol{\theta}^{s+1}, \alpha^{s+1}\}$ are generated by

1. sampling c^{s+1} via a combination of Gibbs and Metropolis-Hastings steps;
2. sampling relevance vectors $\mathbf{r}_{(1)}, \dots, \mathbf{r}_{(K)}$ from their full conditional distributions and then sampling $\boldsymbol{\theta}^{s+1}$ from its conditional distribution given $\mathbf{r}_{(1)}, \dots, \mathbf{r}_{(K)}$, the clustering and the data;
3. sampling α^{s+1} from its full conditional distribution.

Each of these steps is outlined in more detail below.

Sampling c : As discussed in MacEachern (1994) and Neal (2000), one standard approach to MCMC sampling of c is to iteratively sample $c(i)$ conditional on $\{c(i'), i' \neq i\}$, the data, and the values of the other parameters, for each $i = 1, \dots, n$. Letting K be the number of unique values of $\{c(i') : i' \neq i\}$ and n_k the number of units in cluster k , not including unit i , the conditional distribution of $c(i)$ is given by

$$\Pr(c(i) = k | \{c(i') : i' \neq i\}, \boldsymbol{\theta}, \mathbf{y}_1, \dots, \mathbf{y}_n) \propto \begin{cases} n_k \times w_k & \text{if } k < K + 1 \\ \alpha \times w_{K+1} & \text{if } k = K + 1 \end{cases}$$

where the weights are given by

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

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

and $\hat{\lambda}_j^{+i}(k)$, $\hat{\lambda}_j^{-i}(k)$ are calculated as in (7) 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 $c(i) = k$, and α and the n_k 's are artifacts of the Pólya urn prior that reduce the probability of having too many small, singleton clusters.

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. However, with this Gibbs sampling approach new clusters are initially formed with only a single member, and mixing of such a procedure can be slow. If this is the case it may be desirable to also move around the space of clusters by proposing changes at more than just one value of c at a time, for example, by splitting or merging clusters as suggested by Jain and Neal (2004) and Dahl (2003). The general procedure of these authors is as follows:

1. Sample a pair of units $\{i_1, i_2\}$ uniformly from $\{1, \dots, n\}$;
2. If $c(i_1) = k = c(i_2)$, propose a clustering c^* by initially putting i_1 and i_2 into their own clusters, and then reallocating the remaining units of cluster k to these two clusters;
3. If $c(i_1) = k_1 \neq k_2 = c(i_2)$, propose a clustering c^* by putting all units from clusters k_1 and k_2 into a common cluster;
4. Accept the proposed clustering c^* with the appropriate acceptance probability.

Computing the acceptance probability is a relatively straightforward calculation that involves computing part of the likelihood (8). Details of these split-merge approaches can be found in the above-mentioned articles.

Sampling θ : If conjugate beta(a_j, b_j) priors are used for each θ_j , sampling θ can be achieved by first sampling relevance vectors $\mathbf{r}_{(1)}, \dots, \mathbf{r}_{(K)}$ from their full conditional distributions, then sampling θ conditional on these vectors:

1. For $k \in \{1, \dots, K\}$ and $j \in \{1, \dots, m\}$, sample $r_{(k),j} \sim \text{binary}\left(\frac{\exp\{\lambda_j + \hat{\lambda}_j(k)\}}{1 + \exp\{\lambda_j + \hat{\lambda}_j(k)\}}\right)$.
2. For $j = 1, \dots, m$, sample $\theta_j \sim \text{beta}(a_j + \sum_{k=1}^K \{n_{k,j,1} \times (1 - r_{(k),j})\}, b_j + \sum_{k=1}^K \{n_{k,j,0} \times (1 - r_{(k),j})\})$

In step 2 above, the sums determining the parameters of the beta distribution for a given j are counts of the number of 1's and 0's in the dataset that were sampled from the baseline parameter θ_j , presuming the values of $r_{(1),j}, \dots, r_{(K),j}$.

Sampling α : As shown in Antoniak (1974), the distribution of K as a function of α is proportional to $\alpha^K \Gamma(\alpha) / \Gamma(\alpha + n)$, which can be highly skewed in $\alpha \in \mathbb{R}^+$ depending on K . Since we allow K to vary over the MCMC sampling procedure, a fixed proposal distribution for α in Metropolis-Hastings updates may lead to poor mixing. Alternatively, Escobar and West (1998) provide a Gibbs sampling approach based on data augmentation if the prior for α is a gamma distribution. For arbitrary priors, a different approach is to reparameterize in terms of $\pi = \frac{\alpha}{\alpha+1} \in (0, 1)$, which represents the probability that a given pair of units are 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 $(0, 1)$.

4 Model behavior

Some insight into the ability of the model to correctly classify objects can be gained by examining the conditional probability that $c(i) = k$, given the other cluster memberships, the data, and the values of θ and α . The unnormalized probability $n_k w_k$ that $c(i) = k$ can be approximated by taking logs and using a Taylor series expansion of $\log(1 + e^{\lambda_j + \hat{\lambda}_j(k)})$ about $\hat{\lambda}_j(k) = 0$. This gives $\log n_k w_k \approx \log n_k + \sum_k \sum_j \frac{e^{\lambda_j}}{1 + e^{\lambda_j}} \left\{ \hat{\lambda}_j^{+i}(k) - \hat{\lambda}_j^{-i}(k) \right\}$. The second term in the sum simplifies as follows:

$$\begin{aligned} \hat{\lambda}_j^{+i}(k) - \hat{\lambda}_j^{-i}(k) &= \log \frac{\text{Beta}(\tilde{a}_j + n_{k,j,1} + y_{i,j}, \tilde{b}_j + n_{k,j,0} + 1 - y_{i,j})}{\text{Beta}(\tilde{a}_j + n_{k,j,1}, \tilde{b}_j + n_{k,j,0})} - \log \theta_j^{y_{i,j}} (1 - \theta_j)^{1 - y_{i,j}} \\ &= \log \left(\frac{\tilde{a}_j + n_{k,j,1}}{\tilde{a}_j + \tilde{b}_j + n_k} \right)^{y_{i,j}} \left(\frac{\tilde{b}_j + n_{k,j,0}}{\tilde{a}_j + \tilde{b}_j + n_k} \right)^{1 - y_{i,j}} - \log \theta_j^{y_{i,j}} (1 - \theta_j)^{1 - y_{i,j}} \\ &= \log \left(\frac{\hat{\theta}_{(k),j}}{\theta_j} \right)^{y_{i,j}} \left(\frac{1 - \hat{\theta}_{(k),j}}{1 - \theta_j} \right)^{1 - y_{i,j}} \end{aligned}$$

where $\hat{\theta}_{(k),j} = (\tilde{a}_j + n_{k,j,1})/(\tilde{a}_j + \tilde{b}_j + n_k)$ is the Bayes estimate of $p(y_j = 1)$ in cluster k . Thus the procedure places an object i into a cluster with probability proportional to how well that cluster's empirical probabilities fit the vector \mathbf{y}_i relative to the baseline probabilities $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_m\}$ and weighted by the probabilities of attribute relevance. Note that if $\hat{\theta}_{(k),j} \approx \theta_j$, then the contribution of the data from attribute j to the probability that $c(i) = k$ is approximately zero. This makes sense: If cluster k 's rate for attribute j is close to the background rate, then j is probably not relevant for this cluster and so attribute j shouldn't play a large role in determining whether or not $c(i) = k$.

Now suppose an object i actually belongs in cluster k . The log relative probability of $c(i) = k$ versus $c(i) = l$ is

$$\log n_k w_k - \log n_l w_l \approx \log \frac{n_k}{n_l} + \sum_{j=1}^m \frac{e^{\lambda_j}}{1 + e^{\lambda_j}} \log \left(\frac{\hat{\theta}_{(k),j}}{\hat{\theta}_{(l),j}} \right)^{y_{i,j}} \left(\frac{1 - \hat{\theta}_{(k),j}}{1 - \hat{\theta}_{(l),j}} \right)^{1 - y_{i,j}}.$$

Assuming $\hat{\theta}_{(k),j} \approx \theta_{(k),j}$ and $\hat{\theta}_{(l),j} \approx \theta_{(l),j}$, the expected value of this log ratio over \mathbf{y}_i is

$$E[\log n_k w_k - \log n_l w_l] \approx \log \frac{n_k}{n_l} + \sum_{j=1}^m \frac{e^{\lambda_j}}{1 + e^{\lambda_j}} I(\theta_{(k),j} : \theta_{(l),j})$$

where $I(\theta_1 : \theta_2) = \theta_1 \log \frac{\theta_1}{\theta_2} + (1 - \theta_1) \log \frac{1 - \theta_1}{1 - \theta_2}$ is the information divergence (Kullback, 1959) and is interpreted as the expected amount of information in $y_{i,j} \sim \text{binary}(\theta_{(k),j})$ for discriminating against $H : y_{i,j} \sim \text{binary}(\theta_{(l),j})$. Note that if attribute j is not relevant for either cluster k or l then $\theta_{(k),j} = \theta_{(l),j} = \theta_j$ and $I(\theta_{(k),j} : \theta_{(l),j}) = 0$. Since the log probability that $c(i) = k$ is approximately the sum of these divergences, this suggests that the ability of the procedure to correctly classify objects into clusters is related to (a) the variability of the rates of relevant attributes, and (b) the number of relevant attributes. We investigate this claim empirically with a small simulation study. For sample sizes of both $n = 50$ and $n = 100$, we created a clustering c_0 of the units into five groups of sizes (14, 9, 8, 9, 10) and (22, 16, 18, 17, 27) respectively (this was done by uniform random allocation). Additionally, a vector of background rates $\boldsymbol{\theta}$ was generated by independently sampling $\theta_j \sim \text{beta}(1,1)$, $j = 1, \dots, m = 50$. Data were generated for each n and several values of λ and (\tilde{a}, \tilde{b}) as follows:

1. For each cluster $k \in \{1, \dots, 5\}$ and attribute $j \in \{1, \dots, 50\}$;
 - (a) sample $\tilde{\theta}_{(k),j} \sim \text{beta}(\tilde{a}, \tilde{b})$;
 - (b) sample $r_{(k),j} \sim \text{binary}(e^\lambda / (1 + e^\lambda))$;
 - (c) set $\theta_{(k),j} = r_{(k),j} \tilde{\theta}_{(k),j} + (1 - r_{(k),j}) \theta_j$.
2. For $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, 50\}$, sample $y_{i,j} \sim \text{binary}(\theta_{(c_0(i)),j})$

This was done for each combination of $\lambda \in \{-2, -1, 0\}$ and $(\tilde{a}, \tilde{b}) \in \{(1, 1), (1/2, 1/2), (1/4, 1/4)\}$, resulting in 9 simulated datasets each for $n = 50$ and $n = 100$. One additional dataset was

$n = 50$		(\tilde{a}, \tilde{b})			$n = 100$		(\tilde{a}, \tilde{b})		
		1	1/2	1/4			1	1/2	1/4
λ	-2	.19(1)	.19(1)	.51(4)	λ	-2	.20(1)	.41(4)	.86(5)
	-1	.28(4)	.72(4)	.85(5)		-1	.43(5)	.87(6)	.96(5)
	0	.60(6)	1.00(5)	1.00(5)		0	1.00(5)	.93(5)	1.00(5)

Table 1: Results of the simulation study: Numbers are the Jacard index of similarity between \hat{c} and c_0 . The value of \hat{K} for each \hat{c} is given in parentheses.

$n = 50$		(\tilde{a}, \tilde{b})			$n = 100$		(\tilde{a}, \tilde{b})		
		1	1/2	1/4			1	1/2	1/4
λ	-2	.19(1)	.19(1)	.19(1)	λ	-2	.20(1)	.20(1)	.38(3)
	-1	.21(2)	.53(3)	.80(5)		-1	.30(3)	.59(4)	.96(5)
	0	.63(5)	.93(5)	1.00(5)		0	.68(4)	.93(5)	1.00(5)

Table 2: Results of the simulation study using a standard Dirichlet mixture model, clustering on all attributes.

constructed for each of the two sample sizes with one “cluster”, i.e. $K = 1$ and a multivariate mean of $\boldsymbol{\theta}$ for each of the n vectors. Note that the expected number of relevant attributes per cluster is increasing with increasing λ , and the variability of the rates of the relevant attributes is increasing with decreasing (\tilde{a}, \tilde{b}) .

For estimation of each simulated dataset, λ was set to zero, (\tilde{a}, \tilde{b}) was set to $(1, 1)$, and the prior on $\alpha/(\alpha + 1)$ was uniform over $[0, 1]$. The prior on $\boldsymbol{\theta}$ for each dataset was taken to be similar to a “unit information” prior (Kass and Wasserman 1995), where $\theta_j \sim \text{beta}(\bar{y}_{\cdot,j} + 1, (1 - \bar{y}_{\cdot,j}) + 1)$ which gives a relatively uninformative prior that is centered around the empirical mean. Alternatively we could have used a uniform prior on each θ_j , but the simulated θ_j ’s were sampled from a uniform distribution, so use of such a prior could potentially give overly favorable results.

Analysis of each dataset was made using the Markov chain Monte Carlo algorithm as described in Section 3, using a combination of Gibbs steps and Dahl’s (2003) split-merge procedure for resampling c . The starting value of c for each chain was $c(i) = i$, that is, each object was initially placed in its own cluster. The values of $\boldsymbol{\theta}$ were initially set to the empirical means, and α was initially set to 1. Ten thousand scans were run for each dataset, and the estimated clustering \hat{c} was taken to be the sample maximum a posteriori (MAP) estimate, that is, the value c for which $\{c, \alpha, \boldsymbol{\theta}\}$ maximized $\log p(\mathbf{y}_1, \dots, \mathbf{y}_n | c, \alpha, \boldsymbol{\theta}) + \log p(c, \alpha, \boldsymbol{\theta})$ over all scans of the Markov chain. We compared each MAP estimate \hat{c} to the clustering c_0 which generated the dataset using the Jacard index (Milligan, Soon and Sokal 1983), which is given by $j(\hat{c}, c_0) = N_{s,s}/N_{s|s}$, where $N_{s,s}$ is the number of pairs of objects in the same cluster under both clusterings, and $N_{s|s}$ is the number

of pairs in the same cluster under at least one clustering. This index, along with the number of clusters \hat{K} under \hat{c} , is given in Table 1 for each simulated dataset. As expected, the ability of the procedure to correctly cluster the data increases with increasing number of relevant attributes (increasing λ) and increasing extremity of the means of relevant attributes (decreasing a, b). For example, when $\lambda = -2$ and $a = 1$, on average only 12% ($e^{-2}/(1 + e^{-2})$), or 6 of the 50 attributes are relevant per cluster, and the values of $\tilde{\theta}_j$ associated with this small number of relevant attributes are not too different from the background rates. As λ increases and (\tilde{a}, \tilde{b}) decrease, the number of relevant attributes increases, as does the “purity” of the within-cluster data for a relevant attribute (the number of $y_{i,j}$ ’s that are all zero or all one). As a result our ability to identify the clusters increases. Finally, we note that the MAP estimates for the cases with $K = 1$ correctly identified the clustering for both sample sizes, i.e. $\hat{c}(i) = 1 \forall i \in \{1, \dots, n\}$, and so the procedure did not identify any clusters where there were none.

A relevant model to compare to the subset clustering approach is a Dirichlet process mixture model assuming all attributes are relevant to a cluster, i.e. the “standard” clustering approach in which all attribute rates are fit separately for each cluster. This model is essentially a submodel of the one developed in this paper with $\lambda = \infty$. Results for this model applied to the 18 simulated datasets are presented in Table 2. The subset clustering model outperformed or equaled this submodel for all simulated datasets except one ($n = 50, \lambda = 0, \tilde{a} = \tilde{b} = 1$), in which case results were similar. As would be expected, the subset clustering approach can give a substantial improvement over the standard approach when the number of relevant attributes per cluster is small ($\lambda = -1$ or -2).

5 Analysis of genomic abnormality data

In this section we develop a modified version of the model to analyze genomic abnormality data. Data of this type typically consist of a population of tissue or cell samples, each sample being examined for the presence or absence of certain types of genetic abnormalities. Notationally, to each sample $i \in \{1, \dots, n\}$ we associate an m -dimensional binary vector \mathbf{y}_i , where $y_{i,j}$ is the indicator that sample i exhibits abnormality of type j . An abnormality type can be more specific than just a genomic location: In what follows, amplification and deletion of genetic material at a common location are considered as two distinct abnormality types.

Cancer researchers studying a population of tumors are often interested in identifying groupings of tumors based on increased rates of certain abnormality types. The reasoning is that that it is the presence of an abnormality rather than its absence that may result in tumorigenesis. For a data analysis with this goal in mind we restrict the model parameters so that if abnormality type j is relevant for a given cluster then the rate $\tilde{\theta}_j$ in that cluster is higher than the background rate θ_j . Such a constraint may be parameterized as $\tilde{\theta}_j = \theta_j + \rho_j(1 - \theta_j)$ with $\rho_j \in [0, 1]$. Such a constraint modifies the estimation procedure outlined in Section 3 only slightly: A prior distribution on ρ_j

induces a prior on $\tilde{\theta}_j \in [\theta_j, 1]$, and in the case of a beta(\tilde{a}, \tilde{b}) prior for ρ_j , the log Bayes factor $\hat{\lambda}_j(k)$ is still calculated from (7), but the value is now

$$\begin{aligned} \hat{\lambda}_j(k) &= \log \frac{p(\{y_{i,j} : c(i) = k\} | r_j = 1)}{p(\{y_{i,j} : c(i) = k\} | r_j = 0)} \\ &= \log \left\{ \frac{\Gamma(\tilde{a} + \tilde{b})}{\Gamma(\tilde{a})\Gamma(\tilde{b})} \sum_{l=0}^{n_{k,j,1}} \left(\frac{1 - \theta}{\theta} \right)^l \binom{n_{k,j,1}}{l} \frac{\Gamma(\tilde{a} + l)\Gamma(\tilde{b} + n_{k,j,0})}{\Gamma(\tilde{a} + \tilde{b} + l + n_{k,j,0})} \right\}. \end{aligned} \quad (9)$$

The only other modification to the estimation procedure is that a beta prior is no longer conjugate for a baseline rate θ_j , so the baseline rates are resampled using Metropolis-Hastings steps in the MCMC scheme.

To illustrate the method we consider an analysis of $n = 116$ renal cell carcinomas collected by H. Moch and colleagues at the University of Basel, and previously analyzed by Jiang et al. (2000) and Newton (2002). The chromosomal material of each sample was evaluated for $m = 52$ different abnormality types. For these data, an abnormality is either an amplification or deletion of chromosomal material on a particular arm of a chromosome. For example, abnormality $+3p$ refers to an amplification of genetic material on the p -arm of chromosome 3, whereas $-3p$ refers to deletion on the same arm. For these data it is possible to have both amplification and deletion abnormalities on a single chromosomal arm.

A variety of permutation tests performed by this author and by Newton (2002) indicate that there is a large degree of positive dependence among the abnormalities. We investigate this further with the model described by equations (4)-(6) and (9). Parameters in this mixture model were estimated using the MCMC procedure detailed in Section 3. We also include estimation of λ in the procedure, as this parameter has some scientific interest: A biologically important abnormality may be involved in several pathways to tumorigenesis, and thus have a high marginal probability of relevance across clusters. Additionally, estimating this parameter allows a degree of information sharing across clusters.

The priors on $\{e^{\lambda_1}/(1 + e^{\lambda_1}), \dots, e^{\lambda_m}/(1 + e^{\lambda_m})\}$ and $\{\theta_1, \dots, \theta_m\}$ were both taken to be the product of m independent uniform distributions. The conditional prior distribution on each relevant rate $\tilde{\theta}_j$ was uniform on $[\theta_j, 1]$. The prior on the clustering parameter α was such that $\alpha/(\alpha + 1)$ was uniform on $[0, 1]$, resulting in a prior predictive distribution for the number of clusters K having a mode of $K = 1$ (no clustering), but is heavy tailed. Two Markov chains of length 10,000 were run, one starting with all tumors in a single cluster ($K = 1$) and the other starting with each tumor in a separate ‘‘cluster’’ ($K = 116$). Both chains converged quickly to clusterings that were similar to the overall MAP estimate of c , which placed 75 tumors into one cluster, 25 into a second cluster, 11 into a third, and placed 5 tumors into separate ‘‘singleton’’ clusters of size 1, indicating that these five tumors had abnormality patterns that were somewhat inconsistent with those of the larger groups. For analysis of MCMC samples, we refer to the largest cluster in each scan as cluster 1, the second largest as cluster 2, and the third largest as cluster 3. We examine properties of these

three “clusters” averaged over MCMC samples even though the cluster memberships of the tumors are changing somewhat across samples. Such an approach would not be meaningful if the posterior were not roughly unimodal.

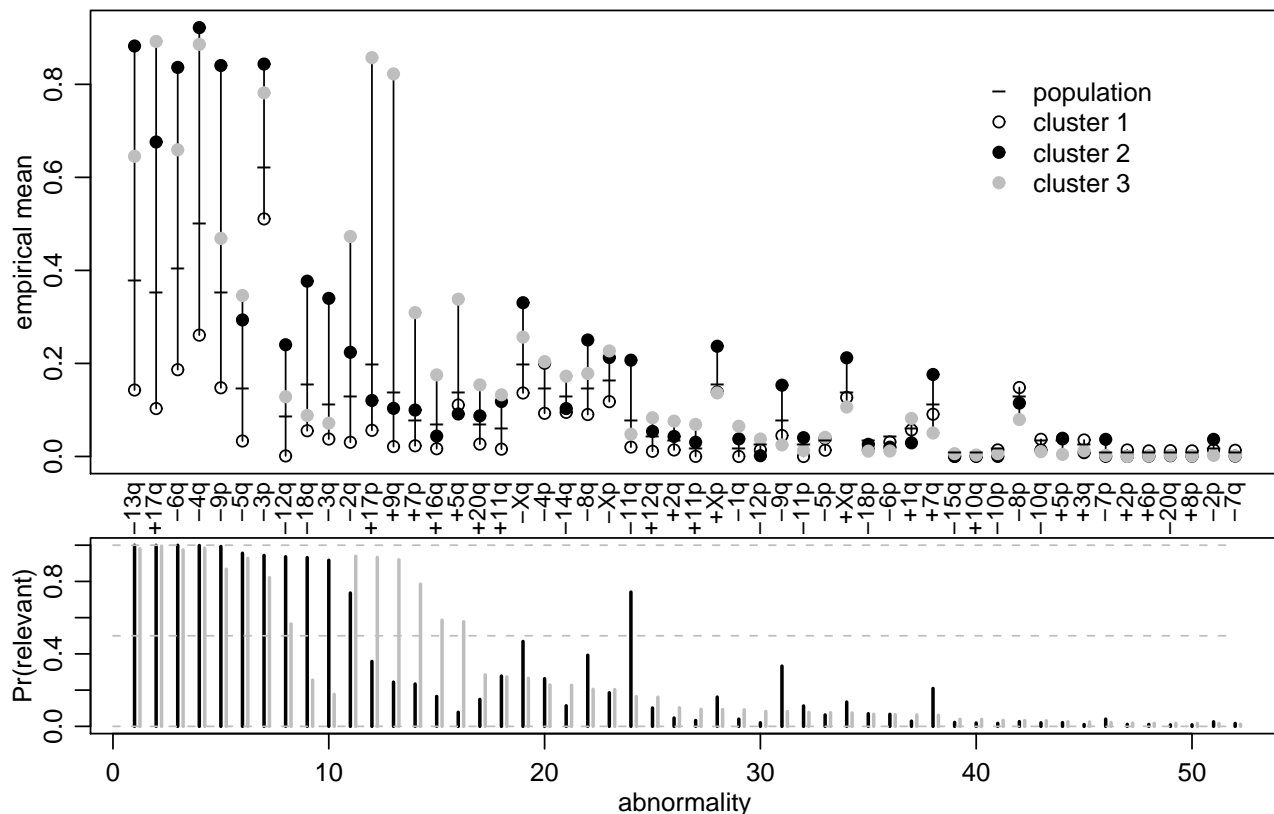


Figure 1: Differences in marginal abnormality rates among the three clusters, The bottom panel gives the probability that an abnormality is relevant for the two clusters

Figure 1 graphically displays the differences in abnormality patterns for the three clusters. The top panel gives the posterior mean abnormality rates for cluster 1 (open circle), cluster 2 (black circle), cluster 3 (gray circle) and the entire population (horizontal line). Cluster 1 has a lower rate than the other two clusters for almost all of the abnormalities, and all abnormalities have a low probability of relevance for this cluster. Marginal posterior probabilities of relevance for the other two clusters ($\Pr(r_{(k),j} = 1 | \mathbf{y}), k = 2, 3$) are plotted in the second panel of Figure 1. Thirteen abnormalities have a high ($> 80\%$) probability of being relevant for either cluster 2 or 3. Seven of these ($-13q, +17q, -6q, -4q, -9p, -5q, -3p$) are highly relevant for both clusters, three ($-12q, -18q, -3q$) have a high relevance for only cluster 2, and three ($-2q, +17p, +9q$) have a high relevance for only cluster 3. The numerical values of these relevance probabilities are given in Table 3. We note that if clusters 2 and 3 are merged then the resulting cluster exhibits a large amount of positive dependence among some of the abnormalities, in particular $+17p$ and $+9q$. The

cluster	Pr(relevance)												
	$-13q$	$+17q$	$-6q$	$-4q$	$-9p$	$-5q$	$-3p$	$-12q$	$-18q$	$-3q$	$-2q$	$+17p$	$+9q$
2	1.00	1.00	1.00	1.00	0.99	0.96	0.94	0.94	0.93	0.92	0.74	0.36	0.24
3	0.98	0.99	0.97	0.98	0.87	0.93	0.82	0.57	0.26	0.18	0.94	0.93	0.92

Table 3: Probabilities of abnormality relevance for clusters two and three.

ability to identify such dependence and represent it in terms of overlapping subsets of attributes is scientifically desirable: Letting $a = \{-13q, +17q, -6q, -4q, -9p, -5q, -3p\}$, $b = \{-12q, -18q, -3q\}$ and $c = \{-2q, +17p, +9q\}$, recognizing that a large number of tumors exhibit abnormalities at “ a and b ” or “ a and c ” could be useful for classification of tumor subtypes and identification of different pathways to tumorigenesis.

The analysis in Newton (2002), which does not allow overlapping subsets of relevant attributes, detects two clusters with a large majority of tumors (≈ 105) belonging to one cluster and the small number of remaining tumors (≈ 11) belonging to another, although this smaller cluster was not detected in a secondary analysis. The relevant abnormalities for Newton’s large cluster were $-13q, +17q, -6q, -4q, -9p$ and $-3p$, which are six of the seven abnormalities relevant to both clusters 2 and 3 above. This group of six includes those abnormalities for which there is a very high *population* rate of occurrence as well as large difference between groups of tumors, but does not include abnormalities having moderate population rates such as $-5q$. We suspect that Newton’s analysis does not identify such abnormalities as relevant because he uses a single common marginal rate for all relevant abnormalities and a lower common marginal rate for all non-relevant abnormalities (although abnormality-specific rates could presumably be accommodated within his model). In some applications this assumption may be reasonable, but it could overlook important and interesting features of some datasets. For example, perhaps some of the measured attributes are not directly involved in biological mechanisms but only correlated with such mechanisms, or perhaps there is error in detecting certain abnormalities. Such attributes might be very relevant to defining subgroups of samples, even if they don’t have extremely high population rates of occurrence.

6 Discussion

In clustering multivariate genomic datasets with large numbers of measured attributes, it is possible that only a subset of the attributes are relevant for identifying groups of similar objects. If this is the case, it may be beneficial to search for clusterings based on differences at subsets of attributes. This article has discussed a model-based approach to clustering objects based on subsets of binary attributes, in which the attributes that define a cluster of objects may be cluster-specific. In a simulation study it was found that allowing for the possibility of subset clustering increased cluster accuracy over a more standard approach that presumes differences between clusters at every

attribute. In an analysis of genomic abnormality data on a population of renal cell carcinomas, the model represented the dependencies among the abnormalities via three clusters in which the largest cluster of tumors differed from the two smaller clusters at distinct but overlapping subsets of attributes.

The subset clustering approach is based on MCMC estimation of a Dirichlet process mixture model, which allows for joint estimation of the number of clusters, the cluster memberships, and the cluster-specific relevant attributes. For genomic data a Bayesian model-based approach may be particularly desirable, so that the estimation procedure may incorporate information from previous experiments, or mimic certain biological assumptions, such as modeling the rates of relevant abnormalities as being higher than background rates. The MCMC procedure also makes posterior inference possible for a wide variety of quantities of interest. For example, the posterior probability that two given attributes are co-relevant in a cluster can be estimated by computing the empirical probability of this event over the MCMC samples. However, such a rich model output comes at a cost: It may be difficult to describe the posterior distribution for a cluster membership function c if it is not roughly unimodal. If multimodality is present, one possible approach to describing the posterior may be to select a metric on clusterings (Meilă 2002) and then “cluster the clusterings,” although this may seem a bit ad-hoc. Another approach used by some authors is to calculate $\Pr(c(i_1) = c(i_2)|\mathbf{y}_1, \dots, \mathbf{y}_2)$, the posterior probability of co-membership for each pair of objects, and then representing these “similarities” graphically via a dendrogram.

R-code for the method presented in this paper is available at the author’s website:
www.stat.washington.edu/hoff.

References

- Antoniak, C. E. (1974), “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems,” *Annals of Statistics*, 2, 1152–1174.
- Blackwell, D. and MacQueen, J. B. (1973), “Ferguson distributions via Pólya urn schemes,” *Annals of Statistics*, 1, 353–355.
- Brodeur, G. M., Tsiatis, A. A., Williams, D. L., Luthardt, F. W., and Green, A. A. (1982), “Statistical analysis of cytogenetic abnormalities in human cancer cells,” *Cancer Genetics and Cytogenetics*, 7, 137–152.
- Dahl, D. B. (2003), “An improved merge-split sampler for conjugate Dirichlet process mixture model,” Technical report no. 1086, Department of Statistics, University of Wisconsin-Madison.
- Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C. H., and Schaffer, A. A. (1999), “Inferring tree models for oncogenesis from comparative genome hybridization data,” *Journal of Computational Biology*, 6, 37–52.

- Escobar, M. D. and West, M. (1998), “Computing nonparametric hierarchical models,” in *Practical nonparametric and semiparametric Bayesian statistics*, vol. 133 of *Lecture Notes in Statistics*, pp. 1–22, Springer, New York.
- Friedman, J. H. and Meulman, J. J. (2004), “Clustering objects on subsets of attributes,” *Journal of the Royal Statistical Society*, to appear.
- Hoff, P. D. (2004), “Clustering based on Dirichlet mixtures of attribute ensembles,” Technical report no. 448, Department of Statistics, University of Washington, Seattle WA.
- 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,” *Journal of the American Statistical Association*, 90, 928–934.
- Kullback, S. (1959), *Information theory and statistics*, John Wiley and Sons, Inc., New York.
- MacEachern, S. N. (1994), “Estimating normal means with a conjugate style Dirichlet process prior,” *Communications in Statistics. Simulation and Computation*, 23, 727–741.
- Meilă, M. (2002), “Comparing clusterings,” Technical report no. 418, Department of Statistics, University of Washington, Seattle WA.
- Milligan, G. W., Soon, S. C., and Sokol, L. M. (1983), “The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5, 40–47.
- Neal, R. M. (2000), “Markov chain sampling methods for Dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, 9, 249–265.
- Newton, M. A. (2002), “Discovering combinations of genomic aberrations associated with cancer,” *Journal of the American Statistical Association*, 97, 931–942.
- Newton, M. A., Wu, S.-Q., and Reznikoff, C. A. (1994), “Assessing the significance of chromosome-loss data: Where are suppressor genes for bladder cancer?” *Statistics in Medicine*, 13, 839–858.
- 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.
- Patrikainen, A. and Mannila, H. (2004), “Subspace clustering of high dimensional binary data — a probabilistic approach,” in *SIAM Data Mining, Workshop on Clustering High Dimensional Data*.