Bayesian Analysis (2006) **1**, Number 2, pp. 321–344

Model-based subspace clustering

Peter D. Hoff*

Abstract. We discuss 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 may depend on the cluster being considered. The method is based on a Pólya urn cluster model for multivariate means and variances, resulting in a multivariate Dirichlet process mixture model. This particular model-based approach accommodates outliers and allows for the incorporation of application-specific data features into the clustering scheme. For example, in an analysis of genetic CGH array data we are able to design a clustering method that accounts for spatial dependence of chromosomal abnormalities.

Keywords: COSA, Dirichlet process, mixture model, nonparametric Bayes, Pólya urn, unsupervised learning, variable selection

1 Introduction

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

In many applications it is quite possible that only a small number of the attributes differentiate groups of 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 estimates of the clusters memberships. More generally, procedures that identify clusters based on potentially non-overlapping subsets of the attributes are

*Center for Statistics and the Social Sciences, University of Washington, Seattle, WA., http://www.stat.washington.edu/hoff/

called subspace clustering algorithms (see Parsons, Haque and Liu 2004 for a review). These too are generally driven by heuristic criteria and search algorithms. In contrast, some model-based methodology has been developed for the case of binary data. Newton (2002) and Hoff (2005) provide model-based subspace clustering methods for binary data which allow the number of clusters, the cluster memberships, and the relevant attributes to be jointly estimated in a unified procedure. Newton's approach is somewhat specific to a particular applied problem in cancer genetics, and does not allow the same attributes to be relevant in different clusters. Hoff's approach is more general and does not have this restriction.

The clustering approach developed in this paper is based on finding groups which differ from each other in terms of their means and/or variances at one or more attributes. For example, in the case where differences between attribute distributions are described by difference in means, we are looking for a value K, a cluster membership function $c : \{1, \ldots, n\} \to \{1, \ldots, K\}$ and K m-dimensional means $\boldsymbol{\mu}_{(1)}, \ldots, \boldsymbol{\mu}_{(K)}$ such that the within-cluster residual sums of squares is small. However, we parameterize $\boldsymbol{\mu}_{(k)} = \boldsymbol{\mu} + \boldsymbol{r}_{(k)} \times \boldsymbol{\delta}_{(k)}$, with $\boldsymbol{r}_{(k)} \in \{0, 1\}^m$, $\boldsymbol{\delta}_{(k)} \in \mathbb{R}^m$ and "$\times$" indicating element-wise multiplication. The vector $\boldsymbol{r}_{(k)} \times \boldsymbol{\delta}_{(k)}$ is then a vector of "mean shifts" for group k, being potentially zero at many entries. A model for such a clustering approach can be obtained by writing $\boldsymbol{y}_i = \boldsymbol{\mu} + \boldsymbol{r}_i \times \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i$ and modeling the distribution of $\{(\boldsymbol{r}_i, \boldsymbol{\delta}_i) : i = 1, \ldots, n\}$ with a Pólya urn scheme. The resulting model for the $\boldsymbol{y}_i$'s is called a Dirichlet process mixture model (Antoniak 1974, MacEachern 1994). Such a model provides estimates of the cluster memberships and an identification of which attributes are likely to be defining each cluster. Additionally, this model-based approach allows for an assessment of uncertainty in the clustering and the incorporation of known data features into the clustering algorithm, as is shown in an analysis of spatially dependent genetic array data in Section 4.

For clarity, we begin with a detailed discussion of the approach in the simple case of subset clustering of mean shifts. Section 2 introduces a model for this simple case, discusses parameter estimation and model behavior, and provides a small simulation study. In this study the subset clustering approach outperforms a more standard Dirichlet mixture model which assumes all attributes are relevant for each cluster. The approach also outperforms COSA and the model-based clustering routine `Mclust` of Fraley and Raftery (2002). Section 3 extends the approach to allow for clusterings based on differences in means and variances, a goal similar to that of Friedman and Meulman's (2004) COSA procedure. The model-based procedure performs as well as COSA on a 10,000-attribute simulated dataset considered in Friedman and Meulman's article, and outperforms the COSA algorithm on a similar dataset. In Section 4 we apply a modified version of the model to spatially correlated genetic array data in which we try to identify groups of tumor cells having common patterns of chromosomal abnormalities. A discussion of the approach and some generalizations follow in Section 5.

2 Clustering mean shifts with a Pólya urn scheme

Given an $\alpha > 0$ and a distribution f_0 on $\{0,1\}^m \times \mathbb{R}^m$, the clustering procedure described in this section is as follows:

$$\begin{aligned} f &\sim \text{Dirichlet}(\alpha, f_0) \qquad (1)\\ \{\boldsymbol{r}_1, \boldsymbol{\delta}_1\}, \ldots, \{\boldsymbol{r}_n, \boldsymbol{\delta}_n\} &\sim \text{i.i.d. } f \\ \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n &\sim \text{i.i.d. multivariate normal}(\mathbf{0}, \text{diag}\{\sigma_1^2, \ldots, \sigma_m^2\}) \\ \boldsymbol{y}_i &= \boldsymbol{\mu} + \boldsymbol{r}_i \times \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i. \end{aligned}$$

The vectors $\boldsymbol{r}_1 \times \boldsymbol{\delta}_1, \ldots, \boldsymbol{r}_n \times \boldsymbol{\delta}_n$ are the "mean shifts" away from $\boldsymbol{\mu}$ described earlier, and f describes their distribution. Modeling f as a Dirichlet process results in what is called a Dirichlet process mixture model, as the density of $\boldsymbol{y}_i$ can be written as the mixture $p(\boldsymbol{y}_i|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, f) = \int p(\boldsymbol{y}_i|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\theta}) f(d\boldsymbol{\theta})$ where $\boldsymbol{\theta} = \{\boldsymbol{r}, \boldsymbol{\delta}\}$ and 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 (2003b) and others.

Model (1) provides the following:

(a) the possible values of f include all discrete distributions on $\{0,1\}^m \times \mathbb{R}^m$ (f is estimated nonparametrically);

(b) a sample of n mean shifts from f may have less than n unique values (the $\{\boldsymbol{r}, \boldsymbol{\delta}\}$'s "cluster");

(c) the attributes j for which $r_{(k),j} \times \delta_{(k),j} \neq 0$ may depend on k (relevant attributes may be cluster-specific).

Samples from a Dirichlet process are discrete, and so f has support on a countable number of $\{\boldsymbol{r}, \boldsymbol{\delta}\}$-values. This discreteness implies that a sample from f could have a number of ties, and thus forms 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 $f \sim \text{Dirichlet}(\alpha, f_0)$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ are i.i.d. samples from f, then unconditional on f the joint distribution of the $\boldsymbol{\theta}_i$'s is equal to that of an exchangeable sequence generated as follows:

1. sample $\boldsymbol{\theta}_1 \sim f_0$;
2. sample $\boldsymbol{\theta}_2 \sim \frac{\alpha}{\alpha+1} f_0 + \frac{1}{\alpha+1} \delta_{\boldsymbol{\theta}_1}(\cdot)$;

   $\vdots$

n. sample $\boldsymbol{\theta}_n \sim \frac{\alpha}{\alpha+n-1} f_0 + \frac{n-1}{\alpha+n-1} \hat{f}_{n-1}$,

where $\delta_{\boldsymbol{\theta}_1}(\cdot)$ is a point-mass measure on $\boldsymbol{\theta}_1$ and $\hat{f}_{n-1}$ is the empirical distribution of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{n-1}$. The above process is called a Pólya urn scheme with parameters α and

f_0. It is clear that, depending on α, the sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ 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 denote the number of draws from f_0 as K, and the values of the draws as $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(K)}$. The function mapping the unit labels $\{1, \ldots, n\}$ to the independent draws $\{1, \ldots, K\}$ is denoted c. As can be seen, α determines the distribution of K, whereas f_0 determines the distribution of the cluster-specific parameters $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(K)}$. Note that under the Pólya urn model there is positive prior probability that a given object is assigned to its own "singleton" cluster. This property allows for the identification of outliers and reduces the possibility that they will hinder the detection of coherent clusters.

Rewriting in terms of the Pólya urn representation, our full probability model can be alternatively described by replacing the first two lines of (1) with

$$\{\boldsymbol{r}_1, \boldsymbol{\delta}_1\}, \ldots, \{\boldsymbol{r}_n, \boldsymbol{\delta}_n\} \sim \text{Pólya urn}(\alpha, f_0).$$

A convenient choice for the baseline distribution f_0 is that of independent binary and normal random variables,

$$f_0(\boldsymbol{r}, \boldsymbol{\delta}) = \prod_{j=1}^{m} \text{binary}(r_j | \frac{e^{\lambda_j}}{1+e^{\lambda_j}}) \times \text{normal}(\delta_j | 0, \tau_j^2). \tag{2}$$

In this case λ_j represents the prior log-odds that the mean of a given attribute within a cluster differs from that of the other clusters. The parameter τ_j^2 represents the average squared magnitude of such a difference. For now we allow λ_j and τ_j^2 to vary among attributes, although in Section 2.3 we suggest a more parsimonious version of the model.

2.1 Model behavior

Key to understanding model behavior and parameter estimation is the probability of the data within a cluster conditional on the clustering c but marginal over the values of the cluster-specific parameters $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(K)}$, where $\boldsymbol{\theta}_{(k)} = \{\boldsymbol{r}_{(k)}, \boldsymbol{\delta}_{(k)}\}$. The marginal distribution for the data from attribute j in cluster k can be obtained by summing and integrating:

$$\begin{aligned} p(\{y_{i,j} : c(i) = k\} | \mu_j, \sigma_j^2, c) &= \sum_{r_j=0}^{1} \frac{e^{\lambda_j r_j}}{1+e^{\lambda_j}} \times \\ &\quad \int \left\{ \prod_{i:c(i)=k} \text{normal}(y_{i,j} | \mu_j + r_j \times \delta_j, \sigma_j^2) \right\} f_0(\delta_j | \tau_j^2)\, d\delta_j \\ &= \frac{1+e^{\lambda_j + \hat{\lambda}_j(k)}}{1+e^{\lambda_j}} \times \prod_{i:c(i)=k} \text{normal}(y_{i,j} | \mu_j, \sigma_j^2) \end{aligned}$$

where $\hat{\lambda}_j(k)$ is given by

$$\begin{aligned}
\hat{\lambda}_j(k) &= \log \frac{p(\{y_{i,j} : c(i) = k\} | r_j = 1, \mu_j, \sigma_j^2, \tau_j^2)}{p(\{y_{i,j} : c(i) = k\} | r_j = 0, \mu_j, \sigma_j^2)} && (3) \\
&= \log \frac{\text{multivariate normal } (\boldsymbol{y}_{(k),j} | \mu_j \mathbf{1}, \sigma_j^2 \mathbf{I} + \tau_j^2 \mathbf{1}\mathbf{1}')}{\text{multivariate normal } (\boldsymbol{y}_{(k),j} | \mu_j \mathbf{1}, \sigma_j^2 \mathbf{I})} && (4) \\
&= \frac{1}{2} \left\{ \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2/n_k} \frac{n_k}{\sigma_j^2} \overline{\xi_j(k)}^2 + \log \frac{\sigma_j^2/n_k}{\sigma_j^2/n_k + \tau_j^2} \right\} && (5)
\end{aligned}$$

where $\overline{\xi_j(k)}$ is the average value of $y_{i,j} - \mu_j$, averaged over objects i in group k. The value of $\hat{\lambda}_j(k)$ can be thought of as the adjustment to the log odds of $r_{(k),j} = 1$, the existence of 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 τ_j^2. Alternatively, $\hat{\lambda}_j(k)$ is a log Bayes factor for evaluating $H : E(y_{i,j}) \neq \mu_j$ versus H^c for data in group k.

Taking the product over all clusters k and attributes j gives

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n | c, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\lambda}, \boldsymbol{\tau}^2) = \left\{ \prod_{k=1}^{K} \prod_{j=1}^{m} \frac{1 + e^{\lambda_j + \hat{\lambda}_j(k)}}{1 + e^{\lambda_j}} \right\} \times \left\{ \prod_{i=1}^{n} \prod_{j=1}^{m} \text{normal}(y_{i,j} | \mu_j, \sigma_j^2) \right\}. \quad (6)$$

Note that the second product on the right hand side does not depend on the clustering or the parameters describing the distribution of the cluster-specific parameters, and so the conditional distribution of the clustering c given the data and the other parameters is proportional to its prior times the first product. The k, jth term in this product is an increasing function of $\hat{\lambda}_j(k)$. Taking logs and using a Taylor series expansion of $\log(1 + e^{\lambda_j + \hat{\lambda}_j(k)})$ about $\hat{\lambda}_j(k) = 0$, the log-likelihood as a function of c is approximately $\sum_{k=1}^{K} \sum_{j=1}^{m} \frac{e^{\lambda_j}}{1+e^{\lambda_j}} \hat{\lambda}_j(k)$ plus a constant. This indicates that higher posterior probability is given to clusterings for which $\hat{\lambda}_j(k)$ is large across attributes, weighted by $e^{\lambda_j}/(1+e^{\lambda_j})$. Recall that $\hat{\lambda}_j(k)$ is the log-ratio of two multivariate normal densities, the numerator modeling the responses at attribute j within a cluster as having marginal correlation $\tau_j^2/(\tau_j^2 + \sigma_j^2)$, the denominator modeling the responses as independent. As can be seen from (5), $\hat{\lambda}_j(k)$ is large if $\overline{\xi_j(k)}^2$ is large compared to σ_j^2/n_k, its expected value under the hypothesis of no mean shift.

It is also informative to look at the likelihood (6) as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. Taking the derivative of the log-likelihood with respect to μ_j, the maximizer in μ_j is seen to satisfy

$$\mu_j = \frac{\sum_{k=1}^{K} [1 - v_j(k)] n_k \bar{y}_j(k)}{\sum_{k=1}^{K} [1 - v_j(k)] n_k}, \qquad \text{with } v_j(k) = \frac{e^{\lambda_j + \hat{\lambda}_j(k)}}{1 + e^{\lambda_j + \hat{\lambda}_j(k)}} \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2/n_k},$$

Without the weight $v_j(k)$ the conditional maximizer would be $\bar{y}_j$, the sample mean over all clusters. The weight has the effect of making the estimate of μ_j predominantly

based on data from clusters k for which there is unlikely to be a mean shift, i.e. k for which $e^{\lambda_j+\hat{\lambda}_j(k)}/(1+e^{\lambda_j+\hat{\lambda}_j(k)})$ is small. Similarly, the maximizer of the likelihood in σ_j^2 satisfies

$$\sigma_j^2 = \frac{\sum_{k=1}^K n_k\{s_j^2(k) + [\bar{y}_j(k)-\mu_j]^2[1-v_j(k)v_j'(k)]\}}{\sum_{k=1}^K [n_k - v_j(k)]}$$

where $s_j^2(k) = \sum_{i:c(i)=k}[y_{i,j}-\bar{y}_j(k)]^2/n_k$ and $v_j'(k) = 1+(\sigma_j^2/n_k)/(\sigma_j^2/n_k+\tau_j^2)$. Although less transparent than the likelihood equation for μ_j, the behavior is analogous: For a group k with a large probability of a mean shift ($v_j(k)$ large), $[\bar{y}_j(k)-\mu_j]^2$ is not a good estimate of σ_j^2 and so the contribution to the estimate of σ_j^2 from group k is dominated by $s_j^2(k)$. On the other hand, if it is unlikely that group k has a mean shift ($v_j(k)$ small), then the contribution to the numerator from group k is approximately $n_k(s_j^2(k)+[\bar{y}_j(k)-\mu_j]^2) = \sum_{i:c(i)=k}[y_{i,j}-\mu_j]^2$.

2.2 Parameter estimation

Inference can be made on $\{c, \alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$ by constructing a relatively straightforward Markov chain which converges to the posterior distribution $p(c, \alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. We suggest an algorithm based on Gibbs sampling of the cluster membership function and the other parameters. This approach is the "standard" estimation technique for Dirichlet process mixture models, and is discussed by MacEachern (1994) for mixtures of univariate normals, and in general by Neal (2000). Given a current state of $\{c, \alpha, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$, one such algorithm iteratively sample new states for each quantity as follows:

1. For $i \in \{1, \ldots, n\}$ in random order, sample $c(i)$ conditional on the data, $\boldsymbol{\mu}, \boldsymbol{\sigma}^2$ and α but marginal over the cluster-specific parameters;
2. For $k \in \{1, \ldots, K\}$, sample $\{\boldsymbol{r}_{(k)}, \boldsymbol{\delta}_{(k)}\}$ from its full conditional distribution;
3. For $j \in \{1, \ldots, m\}$, sample μ_j and σ_j^2 from their full conditional distributions;
4. Sample α from its full conditional distribution.

These steps are outlined in more detail below.

Sampling c: Unconditional on the observed data, the conditional distribution of $c(i)$ given the other values of c 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 $1, \ldots, K$ if unit i is currently in its own cluster. The conditional distribution of $c(i)$ is

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

where $n_{k,-i}$ is the number of objects 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 into a new cluster with probability proportional to α. Conditional on the data and the other parameters these probabilities are reweighted as

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

where the weights are given by

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

and $\hat{\lambda}_j^{+i}(k)$, $\hat{\lambda}_j^{-i}(k)$ are calculated as in (3) but including and excluding $\boldsymbol{y}_i$ 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. To improve mixing of the Markov chain, we suggest including split-merge Metropolis-Hastings steps in the algorithm as well. Details for the implementation of such steps can be found in Jain and Neal (2004) or Dahl (2003a).

Sampling $\{\boldsymbol{r}_{(k)},\boldsymbol{\delta}_{(k)}\}$: Under the prior distribution (2) and given the data, $\boldsymbol{\mu},\boldsymbol{\sigma}^2$ and c, $r_{(k),1},\ldots,r_{(k),m}$ are conditionally independent binary random variables having log-odds $\lambda_j+\hat{\lambda}_j(k)$. If $r_{(k),j}=1$, then $\delta_{(k),j}$ has a normal$(\hat{\delta}_j,\hat{\tau}_j^2)$ distribution, with $\hat{\tau}_j^2=(n_k/\sigma_j^2+1/\tau_j^2)^{-1}$ and $\hat{\delta}_j=\hat{\tau}_j^2(\sum_{i:c(i)=k}(y_{i,j}-\mu_j)/\sigma_j^2)$. Otherwise, $\delta_{(k),j}$ is normal $(0,\tau_j^2)$.

Sampling α: Fixed values of α and n provide a prior predictive distribution for the number of clusters K which is generally concentrated on a small set of integers. More diffuse priors 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 prior on $\alpha/(\alpha+1)$ results in a prior distribution for K that is monotonically decreasing from $K=1$ but has a reasonably heavy tail out to $K=n$.

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 α in 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. For arbitrary priors one can reparameterize in terms of $\pi=\frac{\alpha}{\alpha+1}\in(0,1)$, which represents the probability that a given pair of objects 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 $(0,1)$.

Sampling $\boldsymbol{\mu}, \boldsymbol{\sigma}^2$: The full conditional distribution of $\{\mu_j, \sigma_j^2\}$ depends on the data only through observations from attribute j. It is relatively straightforward to update these parameters for each $j = 1, \ldots, m$ using a Gibbs step in the case of conjugate prior distributions: Given c and $\{\boldsymbol{r}_{(k)}, \boldsymbol{\delta}_{(k)}\}, k = 1, \ldots, K$ we calculate $\varepsilon_{i,j} = y_{i,j} - r_{(c(i)),j} \times \delta_{(c(i)),j}$. These "residuals" at attribute j from all clusters $\{\varepsilon_{i,j}, i = 1, \ldots, n\}$ are conditionally i.i.d. normal(μ_j, σ_j^2). The full conditionals of μ_j and σ_j^2 are then normal and inverse-gamma respectively if conjugate priors are used:

$$\mu_j \sim \text{normal}(\hat{\mu}_j, \hat{\sigma}_j^2) \text{ , where } \hat{\sigma}_j^2 = (n/\sigma_j^2 + 1/v)^{-1},\ \hat{\mu}_j = \hat{\sigma}_j^2(\textstyle\sum_{i=1}^n \varepsilon_{i,j}/\sigma_j^2 + m/v)$$

$$1/\sigma_j^2 \sim \text{gamma}[\nu_1 + n/2, \nu_2 + \textstyle\sum_{i=1}^n (\varepsilon_{i,j} - \mu_j)^2)/2],$$

where m, v, ν_1 and ν_2 parameterize the priors of μ_j and σ_j^2.

2.3 Estimating hyperparameters

The types of clusters identified by the above modeling strategy are determined in part by the prior f_0 for the mean shifts, which in turn depends on the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}^2$. Ideally, one has a good idea of what types of clusters are of interest and is able to specify fixed values or informative prior distributions for some or all of these parameters. Alternatively, in many data analysis situations the attribute measurements may be of a common type. In this case it might be appropriate to model the relevance indicators r_j and mean shifts δ_j as depending on some parameters that are shared across attributes, and then estimating these parameters from the data. For example, one possibility is to model f_0 as follows:

$$f_0(\boldsymbol{r}, \boldsymbol{\delta}) = \prod_{j=1}^m \text{binary}(r_j|\frac{e^\lambda}{1+e^\lambda}) \times \text{normal}(\delta_j|0, \tau_j^2 = \eta \times \sigma_j^2), \tag{7}$$

and then include estimation of λ and η in the MCMC algorithm. In the above model, τ_j^2 has been parameterized as $\tau_j^2 = \eta\sigma_j^2$, and so η relates the magnitude of the mean shifts at an attribute to the variance of the attribute, or alternatively, $\eta/(\eta+1)$ is the marginal correlation at relevant attributes among observations in the same cluster. A prior distribution on λ and η amounts to modeling the elements of the relevance vectors $\boldsymbol{r}$ and $\boldsymbol{\delta}/\boldsymbol{\sigma}$ as marginally exchangeable but dependent. Such an exchangeable prior allows for differences across attributes of the cluster-specific parameters, but generally provides estimates with lower variability than if they were modeled as a priori independent. Even if the attributes represent measurements of very different types, it still might be desirable

to use such an exchangeable prior, as it establishes a common criterion for relevance across attributes.

Standard conjugate priors for these hyperparameters are beta (a_λ, b_λ) for $e^\lambda/(1+e^\lambda)$ and inverse-gamma (a_η, b_η) for η. Estimation of these parameters can be incorporated into the Markov chain described above by including Gibbs sampling steps for these parameters:

- sample $e^\lambda/(1+e^\lambda) \sim$beta$\{a_\lambda + \sum_{k=1}^K \sum_{j=1}^m r_{(k),j},\ b_\lambda + \sum_{k=1}^K \sum_{j=1}^m (1 - r_{(k),j})\}$;
- sample $\eta \sim$ inverse-gamma$\{a_\eta + \frac{1}{2}nK,\ b_\eta + \frac{1}{2}\sum_{k=1}^K \sum_{j=1}^m \delta^2_{(k),j}/\sigma_j^2\}$.

One additional modification to the MCMC algorithm required by model (7) is that, conditional on η, the mean shifts give information about $\boldsymbol{\sigma}^2$. As a result, the Gibbs sampling step for $\boldsymbol{\sigma}^2$ becomes

- sample $\sigma_j^2 \sim$ inverse-gamma$\{a+\frac{1}{2}(n+K), b+\frac{1}{2}(\sum_{i=1}^n (\varepsilon_{i,j}-\mu_{i,j})^2+\sum_{k=1}^K \delta^2_{(k),j}/\eta)\}$.

2.4 Simulation Study

We expect the accuracy of the clustering procedure to be high if the number of relevant attributes is large (large λ) and/or the magnitude of the mean shifts relative to the error variance is large (large η). We investigate these claims empirically with a simulation study. Eighteen datasets were generated as follows:

1. For $k = 1, \ldots, K = 10$, generate $\{\boldsymbol{r}_{(k)}, \boldsymbol{\delta}_{(k)}\}$ via
   - $r_{(k),1} \ldots, r_{(k),m} \sim$ i.i.d. binary$(e^\lambda/(1+e^\lambda))$, and
   - $\delta_{(k),1} \ldots, \delta_{(k),m} \sim$ i.i.d. normal$(0, \eta)$.
2. For $i = 1, \ldots, n$, sample $c(i)$ uniformly from $\{1, \ldots, K\}$ and set $\{\boldsymbol{r}_i, \boldsymbol{\delta}_i\} = \{\boldsymbol{r}_{(c(i))}, \boldsymbol{\delta}_{(c(i))}\}$.
3. For $i = 1, \ldots, n$ sample $\boldsymbol{y}_i \sim$ multivariate normal$(\boldsymbol{r}_i \times \boldsymbol{\delta}_i, \mathrm{I})$.

This data generating mechanism is the same as the model described by (1) and (7), except that the cluster memberships for the simulated data are not generated with a Pólya urn scheme. Eighteen datasets were generated in this manner, one for each combination of $\lambda \in \{-3, -2, -1\}$, $\eta \in \{1/2, 1, 2\}$ and $n \in \{50, 100\}$. These parameter values provide datasets exhibiting varying degrees of information about the clustering, from small mean shifts at roughly 5% of the attributes ($\eta = 1/2, \lambda = -3$) to large mean shifts at roughly 25% of the attributes ($\eta = 2, \lambda = -1$). To increase comparability across simulations, the same clustering was used for each of the nine datasets for a given sample size, and the clustering for the $n = 50$ datasets was a subclustering of the $n = 100$ clustering.

We used the model described by (1) and (7) to analyze the simulated data, with "unit information" priors (Kass and Wasserman 1995) for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ whereby a priori we have $\mu_j \sim \text{normal}(\bar{y}_{\cdot,j}, s_j^2)$ and $\sigma_j^2 \sim \text{inverse gamma}(1/2, s_j^2/2)$. The prior on $\alpha/(\alpha+1)$ was taken to be uniform(0,1), and the priors for $e^\lambda/(1+e^\lambda)$ and η were uniform(0,1) and inverse-gamma(1/2,1/2) respectively. A Markov chain of length 10,000 was run for each dataset, each starting with all n units in the same cluster. Convergence of the chains appeared to be rapid, typically occurring in the first few tens of scans and always occurring within the first half of the chain.

Figure 1 gives some indication of the mixing of the chains for three of the simulated datasets. The figure plots the Jaccard index of similarity (Jaccard 1912, Milligan, Soon and Sokal 1983) between sampled clusterings c and the clustering c_0 which generated the data for every tenth scan of the Markov chain. The Jaccard index is defined as $J(c, c_0) = N_{s \cdot s}/N_{s|s}$, with $N_{s \cdot s}$ being the number of pairs of objects in the same group under both clusterings, and $N_{s|s}$ the number of pairs in the same group under at least one clustering. With c_0 being the true clustering, this index measures how well a clustering c identifies pairs of objects that should be in the same group, but penalizes c if it puts too many objects in the same group. We also assess posterior mean performance of the method using the Jaccard index. The average values of $J(c, c_0)$, averaged over sampled clusterings from the second halves of the Markov chains, are shown in Table 1. As conjectured, the ability to accurately identify clusters increases with the number of relevant attributes (λ) and the magnitude of the mean shifts (η).

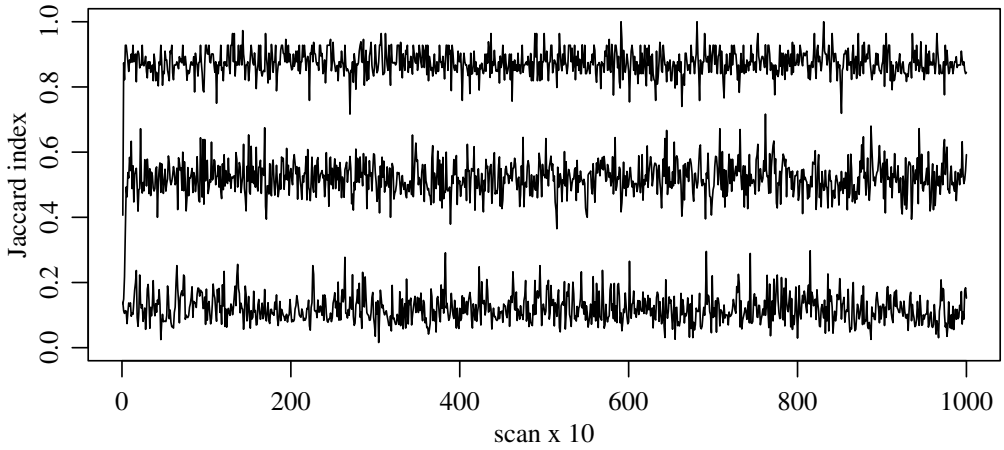

Figure 1: Plots of the Jaccard index of similarity between c_0 and every 10th sampled value of c for the three datasets $\eta \in \{1/2, 1, 2\}$ with $n = 50$ and $\lambda = -2$.

If the posterior distribution of c indicates a strong clustering then it may be of interest to identify the attributes which contribute substantially to the between-group differences. One way to do this is to obtain posterior estimates of $\boldsymbol{r}_{(1)}, \ldots, \boldsymbol{r}_{(K)}$ and

$n = 50$		η 1/2	1	2
λ	-3	0.07, 0.02	0.07, 0.02	0.15, 0.04
	-2	0.12, 0.08	0.52, 0.34	0.87, 0.68
	-1	0.28, 0.27	0.75, 0.72	0.91, 0.83

$n = 100$		η 1/2	1	2
λ	-3	0.08, 0.01	0.08, 0.02	0.36, 0.05
	-2	0.16, 0.08	0.67, 0.31	0.99, 0.88
	-1	0.76, 0.60	0.90, 0.90	0.97, 0.95

Table 1: Results of the simulation study: Numbers are the the Jaccard indices of similarity averaged over sampled clusterings from each of the Markov chains. The first number is the average index using the subspace clustering method, the second using a standard Dirichlet mixture model, clustering on all attributes.

$\boldsymbol{\delta}_{(1)}, \ldots, \boldsymbol{\delta}_{(K)}$ conditional on a single estimate of the clustering. We give an example of this for the $n = 50, \lambda = -2, \eta = 2$ dataset. The estimated modal clustering $\hat{c}$ (the value of c that was sampled the most number of times) had a Jaccard index of $J(c_0, \hat{c}) = 0.88$, placing 48 observations into 10 groups which were "pure" except for one misplaced object. The remaining two objects were not grouped within these ten clusters or with each other. The MCMC algorithm was run as above but with c fixed at $\hat{c}$. To summarize the conditional posterior distribution of $\boldsymbol{r}_{(1)}, \ldots, \boldsymbol{r}_{(K)}$ and $\boldsymbol{\delta}_{(1)}, \ldots, \boldsymbol{\delta}_{(K)}$, we categorize an attribute j as relevant for a given cluster k if $r_{(k),j} = 1$ for more than half of the saved scans. Figure 2 plots the MCMC sample mean of $r_{(k),j} \times \delta_{(k),j}$ at these attributes for the three largest clusters. As can be seen, such a criterion gives good estimates of the larger mean shifts but misses some of the smaller mean shifts.

An appropriate model to compare to the subset clustering approach is a Dirichlet process mixture model assuming all attributes are relevant to all clusters, i.e. a "standard" clustering method. This model can be viewed as a submodel of the one developed in this paper with $\lambda = \infty$. Results for this model applied to the 18 simulated datasets are also presented in Table 1. The subset clustering model outperformed this submodel for all simulated datasets, although as we would expect, the improvement decreases as the number of relevant attributes increases.

We also clustered the data using two other approaches: the model based clustering algorithm `Mclust` (Fraley and Raftery, 2002) as implemented in the R computing environment, and the COSA algorithm of Friedman and Meulman. The routine `Mclust` did not identify any clusters for any of the 18 simulated datasets (i.e. the preferred number of groups was $K = 1$ for each dataset). We speculate that this is due to the fact `Mclust` fits an m-dimensional mean vector for each group and that it uses BIC to penalize model complexity. For these simulated data the actual number of mean parameters required to differentiate between groups is much less than m, and so the

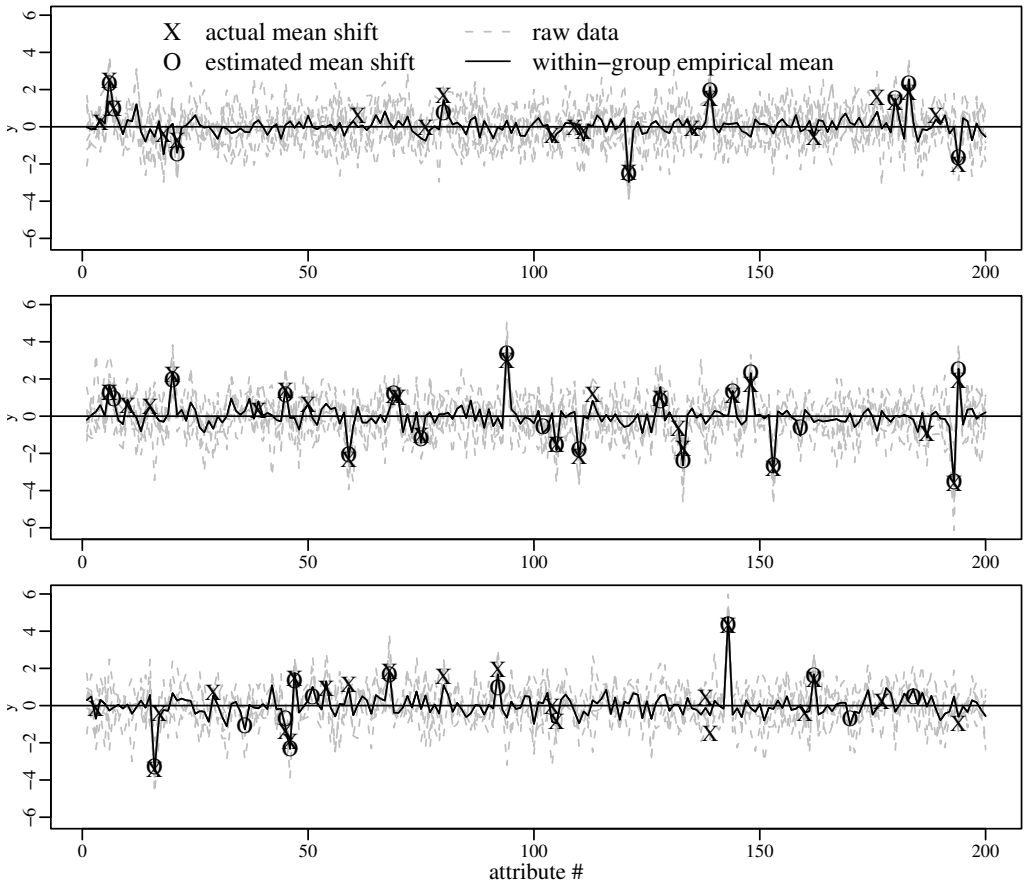

Figure 2: Data and estimated mean shifts from the three largest clusters from the $n = 50, \eta = 2, \lambda = -2$ dataset.

improvement in fit obtained by adding a cluster is small relative to the BIC complexity penalty of order m.

In contrast to `Mclust`, the goal of the COSA algorithm is to find clusters based on subsets of relevant attributes. The COSA algorithm generates a set of distances between objects which one can then input into a distance-based clustering method. We computed COSA distances for only the dataset with the strongest evidence of clustering ($n = 100, \lambda = -1, \eta = 2$). An average-linkage dendrogram of the COSA distances is shown in Figure 3, along with the group labels used to generate the data. As can be seen, the COSA distances are only weakly related to the true group labels. Additionally, the shape of the dendrogram does not reflect the strong clustering of the data. In contrast, the proposed model-based method produced a posterior mode clustering for these data

that identified the 10 clusters and correctly grouped 99 of the 100 objects. This lack of performance by COSA is partly due to the fact that COSA distances are largely based on attributes that have low variances within a group, as opposed to attributes where the within-group mean differs from that of the other groups. This is discussed further in Section 3.2 .

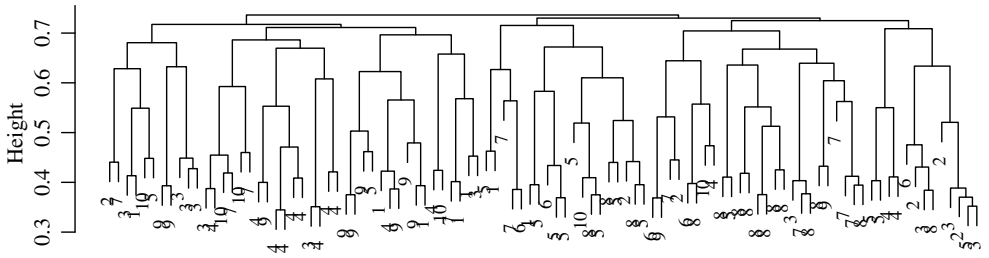

Figure 3: Average linkage dendrogram using COSA distances for the $n = 100, \eta = 2, \lambda = -1$ dataset. Plotting characters are the true group labels.

3 Clustering shifts in mean and variance

In many cases it is desirable to form groups of objects based on differences in means and variances. For example, one notion of subspace clustering is that the data for a relevant attribute within a cluster are all very similar (having a small variance), whereas the data for an irrelevant attributes vary widely (having a high variance). This is the notion of clustering that Friedman and Meulman (2004) consider with their COSA algorithm. In other applications we might want to allow for a positive mean-variance relationship, such as with skewed data or in biostatistical applications for which biological activity may be marked by increases in mean and variance.

Recall that the modeling approach discussed in the previous section identifies clusters by modeling the variance and covariance at a relevant attribute j within a cluster, unconditional on the mean shift, as

$$\mathrm{Var}(y_{i,j}) = \sigma_j^2 + \tau_j^2, \quad \mathrm{Cov}(y_{i_1,j}, y_{i_2,j}) = \tau_j^2, \quad \mathrm{Cor}(y_{i_1,j}, y_{i_2,j}) = \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2} = \frac{\eta}{\eta + 1},$$

where the last equality holds under the parameterization in (7). A model extension allowing for cluster-specific variance at relevant attributes can be parameterized as

$$\mathrm{Var}(y_{i,j}) = \omega^2_{(k),j}(\sigma_j^2 + \tau_j^2), \ \mathrm{Cov}(y_{i_1,j}, y_{i_2,j}) = \omega^2_{(k),j}\tau_j^2, \ \mathrm{Cor}(y_{i_1,j}, y_{i_2,j}) = \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2} = \frac{\eta}{\eta + 1},$$

where $\omega^2_{(k),j}$ represents the shift in variance at attribute j within cluster k. This can be

written as the following modification to model (1):

$$\begin{aligned} f &\sim \text{Dirichlet}(\alpha, f_0) \\ \{\boldsymbol{r}_1, \boldsymbol{\delta}_1, \boldsymbol{\omega}_1^2\}, \ldots, \{\boldsymbol{r}_n, \boldsymbol{\delta}_n, \boldsymbol{\omega}_n^2\} &\sim \text{i.i.d. } f \\ \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n &\sim \text{i.i.d. multivariate normal}(\mathbf{0}, \text{diag}\{\sigma_1^2, \ldots, \sigma_m^2\}) \\ y_{i,j} &= \mu_j + r_{(k),j} \times \delta_{(k),j} \times \omega_{(k),j} + \omega_{(k),j}^{r_{(k),j}} \times \epsilon_{i,j} \end{aligned} \tag{8}$$

where now the cluster-specific parameters are $\boldsymbol{\theta}_{(k)} = \{\boldsymbol{r}_{(k)}, \boldsymbol{\delta}_{(k)}, \boldsymbol{\omega}_{(k)}^2\}$. A simple conjugate base measure is given by

$$f_0(\boldsymbol{r}, \boldsymbol{\delta}, \boldsymbol{\omega}^2) = \prod_{j=1}^{m} \text{binary}(r_j | \frac{e^\lambda}{1+e^\lambda}) \times \text{normal}(\delta_j | 0, \tau_j^2 = \eta\sigma_j^2) \times \text{inverse gamma}(\omega_j^2 | a_\omega, b_\omega). \tag{9}$$

The values of a_ω and b_ω determine what types of variance shifts the model will detect. For example, to detect COSA-like clusterings where the variance at relevant attributes is lower than the background variance we might fix $a_\omega \geq b_\omega + 1$ so that $\omega_{(k),j}^2$ is less that one on average. Alternatively we could put a prior on (a_ω, b_ω) along with the other hyperparameters as discussed in Section 2.3.

3.1 Parameter Estimation

Posterior calculations for the model given by (8) and (9) can be made using the algorithm outlined in Section 2.2 and 2.3 with the modification that the numerator of $\hat{\lambda}_j(k)$ in (3) now must be obtained by integrating out the cluster specific values of $\omega_{(k),j}^2$ in addition to those of $r_{(k),j}$ and $\delta_{(k),j}$. The within-cluster distribution of the data at a relevant attribute, marginal over $\omega_{(k),j}^2$ and $\delta_{(k),j}$, is a multivariate t-distribution with density

$$\begin{aligned} p(\{y_{i,j} : c(i) = k\} | r_j = 1) &= (2\pi)^{-n_k/2} |\sigma_j^2 I + \tau_j^2 11'|^{-1/2} \frac{\Gamma(a_\omega + n_k/2)}{\Gamma(a_\omega)} b_\omega^{a_\omega} \times \\ &\left[b_\omega + \frac{n_k}{\sigma_j^2} \left\{ \overline{\xi_j(k)^2} - \frac{\tau_j^2}{\tau_j^2 + \sigma_j^2/n_k} \overline{\xi_j(k)}^2 \right\} \right]^{-(a_\omega + n_k/2)} \end{aligned} \tag{10}$$

The quantity $\hat{\lambda}_j(k)$ is given by $\hat{\lambda}_j(k) = \log[p(\{y_{i,j} : c(i) = k\} | r_j = 1) / p(\{y_{i,j} : c(i) = k\} | r_j = 0)]$ as before, and is used to perform Gibbs sampling of the cluster function c and the relevance vectors $\boldsymbol{r}_{(1)}, \ldots, \boldsymbol{r}_{(K)}$ as described in Section 2.2. Gibbs sampling of the other model parameters also proceeds as in Section 2.2, with the modification that quantities at relevant attributes are rescaled by the shifts in standard deviation $\boldsymbol{\omega}_{(1)}, \ldots, \boldsymbol{\omega}_{(K)}$. Details of the rescaling are straightforward, and the procedure is implemented in R-code available at the author's website. In order to perform the rescaling, conditional samples of $\boldsymbol{\omega}_{(1)}^2, \ldots, \boldsymbol{\omega}_{(K)}^2$ are required, and can be sampled as $\omega_{(k),j}^2 \sim$ inverse-gamma $(a_\omega + n_k/2,\ b_\omega + \frac{1}{2}\boldsymbol{\xi}'_{(k),j}\Sigma_j\boldsymbol{\xi}_{(k),j})$ if $r_{(k),j} = 1$, where $\boldsymbol{\xi}_{(k),j} = \{y_{i,j} - \mu_j : c(i) = k\}$ and $\Sigma_j = \sigma_j^2\mathbf{I} + \tau_j^2\mathbf{1}\mathbf{1}'$, and $\omega_{(k),j}^2 \sim$ inverse-gamma (a_ω, b_ω) if $r_{(k),j} = 0$.

3.2 Comparison to COSA

We evaluate the model given by (8) and (9) with two simulated datasets, one considered by Friedman and Meulman (2004) and one similar to it. The dataset considered by Friedman and Meulman consists of 10,000 attributes measured on 100 objects, in which the responses from all but 150 of the attributes were generated from a standard normal distribution. For one cluster of 85 objects, these 150 attributes are also standard normal, but for a second cluster of 15 objects these attributes were sampled from a normal distribution with mean 1.5 and standard deviation of 0.2. Thus the two groups differ in both mean and variance at 150 out of 10,000 attributes. As shown in the article, the COSA algorithm does a good job of distinguishing the two groups. To examine the ability of COSA to detect mean shifts in the absence of variance reductions, Hoff (2004) applied the COSA algorithm to a dataset similar to the one described above, having the mean shift of 1.5 but lacking any change in variance. As shown in Hoff (2004), COSA was unable to detect the clusters, indicating that the algorithm may be insensitive to group differences based on means alone.

We evaluated the Dirichlet mixture model given by (8) and (9) on these two 10,000 attribute datasets. Prior distributions for all parameters were as in the simulation study in Section 2.4, and the hyperparameters (a_ω, b_ω) were fixed at $(3, 2)$. This gives a relatively diffuse prior for the ω^2's, having a prior mean of 1 and a prior mode of $1/2$. This prior favors clusterings having a decrease in variance at relevant attributes, but allows for no change in variance and even the possibility of an increase in variance. Two Markov chains of length 1000 were run, one for each of the two simulated datasets. In the case of the data having both a mean and variance shift, all sampled clusterings after scan 222 were equal to the true clustering. For the data with just a mean shift, all scans after scan 90 were equal to the true clustering. This indicates that the true clustering is a very strong mode of the posterior distributions for both datasets. This should not be too surprising, as the class of models described by (8) includes the distributions which generated both datasets. However, it seems important to note that the COSA procedure fails to identify the clusters in the second dataset, a dataset having a relatively simple cluster structure that is apparently easy to identify with a correct model class.

4 A model extension for chromosomal abnormalities

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 has two copies of each chromosome, whereas cells having undergone errors in duplication may have lost or replicated certain regions of chromosomes, potentially resulting in one copy of chromosomal material at a given location (deletion) or more than two copies (amplification). If a cell lineage undergoes tumorigenesis, then such copy number changes are passed on to the descendant cells that eventually make up a tumor. If tumors from different, unrelated individuals all have the same types of deletion or amplification events at a combination of locations, then this is some evidence

that these locations play a role in tumorigenesis. 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 clustering model for binary chromosomal abnormality data used to identify mechanisms of tumorigenesis.

In this section we consider a first step in such an analysis: determining the extent to which a population of tumors can be divided into groups having similar patterns of chromosomal abnormalities. 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 samples from tumor i at genome 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 amplification abnormalities for tumor i at location j. The researchers provided data on chromosomes 1, 6, 16, and 17 for statistical analysis. Hybridization levels were measured at 345, 183, 150, and 133 locations on these four chromosomes respectively, for a total of $m = 811$ observations for each of the $n = 44$ tumors. Figure 4 shows hybridization data from chromosome

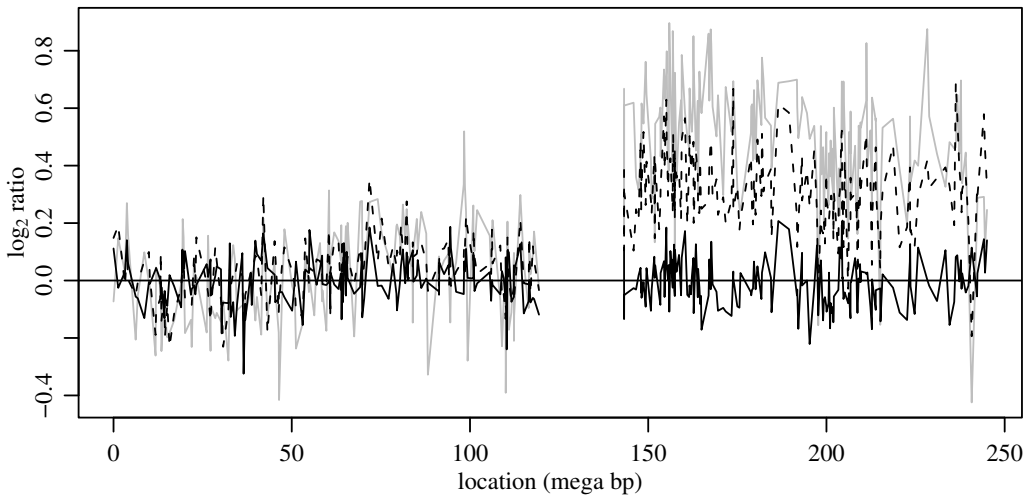

Figure 4: Hybridization ratio data from chromosome 1 for three tumors. The gap is the centromere.

1 for three tumors. The three tumors show a similar profile along the first part of chromosome 1, whereas only two of the three show an amplified response along the second part of the chromosome, suggesting a potential clustering based on subsets of genome locations. Additionally, there is some indication that the amplified responses are accompanied by increases in variance. Finally, the figure suggests that abnormalities occur at contiguous regions of the chromosome, so that the presence of abnormalities along the chromosome is a spatially dependent process.

4.1 Modeling spatial genetic events

To analyze these data we use the model for shifts in means and variances as outlined by (8) and (9), but modify f_0 to account for the spatial nature of the abnormalities. This is done by modeling the vector $\boldsymbol{r}$ as a binary Markov sequence, parameterized as

$$\text{log odds}(r_j = 1 | r_{j-1}) = \begin{cases} \gamma_1 + \gamma_2 r_{j-1} & \text{if locations } j \text{ and } j\text{-1 are next to each other;} \\ \gamma_1 & \text{otherwise.} \end{cases}$$

Two consecutively numbered locations are not next to each other if they are on different chromosomes or on different arms of the same chromosome. This model can also be written in exponential family form,

$$f_{0r}(\boldsymbol{r}) = \kappa(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)^{-1} \exp\{\sum_{j=1}^{m} \lambda_{1,j} r_j + \sum_{j=2}^{m} \lambda_{2,j} r_j r_{j-1}\} \tag{11}$$

where $\lambda_{1,j} = \gamma_1 + \log(1 + e^{\gamma_1}) - \log(1 + e^{\gamma_1 + \lambda_{2,j}})$ and $\lambda_{2,j} = \gamma_2$ if locations j and $j-1$ are next to each other and $\lambda_{2,j} = 0$ otherwise. Alternatively we could model the dependence between measurements at j and $j-1$ as a function of the genetic distance between them. We do not pursue this model complication, as the measurement locations are approximately evenly spaced along the four chromosomes. One additional model modification we do make is to constrain $\boldsymbol{\mu} = \mathbf{0}$, as we are interested in detecting common regions of chromosomal gain or loss rather than deviations from an average amount of gain or loss.

The full probability model for all data and unknown quantities is:

- Sampling model: $\boldsymbol{y}_i \sim$ multivariate normal $(\boldsymbol{r}_i \times \boldsymbol{\delta}_i \times \boldsymbol{\omega}_i, \text{diag}\{\boldsymbol{\sigma}^2 \times (\boldsymbol{\omega}_i^2)^{\boldsymbol{r}_i}\})$.
- Prior for variances :
  - $\sigma_1^2, \ldots, \sigma_m^2 \sim$ i.i.d. inverse gamma (ν_1, ν_2) .
  - $(\log \nu_1, \log \nu_2) \sim$ multivariate normal $(\mathbf{0}, 100 \times \mathbf{I})$.
- Clustering model: $\{\boldsymbol{r}_1, \boldsymbol{\delta}_1, \boldsymbol{\omega}_1^2\}, \ldots, \{\boldsymbol{r}_n, \boldsymbol{\delta}_n, \boldsymbol{\omega}_n^2\} \sim$ Pólya urn(α, f_0).
- Prior for α: $\frac{\alpha}{\alpha+1} \sim$ uniform(0,1) .
- Prior for f_0: $f_0(\boldsymbol{r}, \boldsymbol{\delta}, \boldsymbol{\omega}^2) = \prod_{j=1}^{m} \text{binary}(r_j | r_{j-1}, \gamma_1, \gamma_2) \times \text{normal}(\delta_j | 0, \eta\sigma_j^2) \times$ inverse-gamma$(\omega_j^2 | a, b)$;
  - $\frac{e^{\gamma_1}}{1+e^{\gamma_1}}, \frac{e^{\gamma_1+\gamma_2}}{1+e^{\gamma_1+\gamma_2}} \sim$ independent uniform (0,1);
  - $\eta \sim$ inverse gamma $(1/2, 1/2)$;
  - $(\log a, \log b) \sim$ multivariate normal $(\mathbf{0}, 100 \times \mathbf{I})$.

The prior on the variances $\sigma_1^2, \ldots, \sigma_m^2$ is an exchangeable prior, in that unconditional on (ν_1, ν_2) the variances are dependent but exchangeable. This has the effect of reducing the variability of the σ_j^2's, and is justifiable in the sense that each attribute measurement is of a similar type (CGH array data), so that knowledge of the variance at one attribute gives some information about the variance at another. The prior distribution on the parameters in the base measure f_0 are proper but diffuse. These can also be viewed as generating a prior for the cluster-specific parameters $\{\boldsymbol{r}, \boldsymbol{\delta}, \boldsymbol{\omega}^2\}$ which is exchangeable across attributes.

4.2 Parameter estimation

Parameter estimation for this more complicated model is only slightly more difficult than that for the model described in Section 3. As before, a useful quantity for estimation is the marginal probability of the data in a cluster given the parameters and clustering:

$$p(\{\boldsymbol{y}_i : c(i) = k\}) = \left\{ \prod_{k=1}^{K} \frac{\kappa(\boldsymbol{\lambda}_1 + \hat{\boldsymbol{\lambda}}(k), \boldsymbol{\lambda}_2)}{\kappa(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)} \right\} \times \left\{ \prod_{j=1}^{m} \prod_{i:c(i)=k} \text{normal}(y_{i,j} : \mu_j, \sigma_j^2) \right\},$$

where $\hat{\boldsymbol{\lambda}}(k) = \{\hat{\lambda}_1(k), \ldots, \hat{\lambda}_m(k)\}$ is as defined by (3) and (10).

Gibbs sampling of the cluster memberships proceeds as in Section 2.2 but with the following modification to the weights $w_1, \ldots, w_k$:

$$\begin{aligned} w_k &= \frac{\kappa(\boldsymbol{\lambda}_1 + \hat{\boldsymbol{\lambda}}^{+i}(k), \boldsymbol{\lambda}_2)}{\kappa(\boldsymbol{\lambda}_1 + \hat{\boldsymbol{\lambda}}^{-i}(k), \boldsymbol{\lambda}_2)} \quad \text{if } k < K+1, \\ w_K &= \frac{\kappa(\boldsymbol{\lambda}_1 + \hat{\boldsymbol{\lambda}}^{+i}(k), \boldsymbol{\lambda}_2)}{\kappa(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)} \end{aligned}$$

Conditional on the clustering and the other parameters, $\boldsymbol{r}_{(k)}$ can be sampled from the (exponentially parameterized) Markov model (11) with parameters $\{\boldsymbol{\lambda}_1 + \hat{\boldsymbol{\lambda}}(k), \boldsymbol{\lambda}_2\}$. The remaining parameters in the model can be updated by sampling either from full conditionals as described in Sections 2 and 3 or Metropolis-Hastings steps for ν_1, ν_2, a, and b.

4.3 Posterior Inference

Four Markov chains of length 25,000 scans each were generated as described above. One chain was begun with $K = 1$, one with $K = 44$, and the two remaining chains were given randomly sampled starting values for the clustering function. The Markov chains arrived at similar regions of the parameter space after a few hundred scans. After dropping the first 1000 scans from each Markov chain, the modal clustering $\hat{c}$ (the clustering that was sampled the most number of times) placed the 44 tumors into three groups of sizes 14, 11 and 9, and put 10 tumors into their own "outlier" groups.

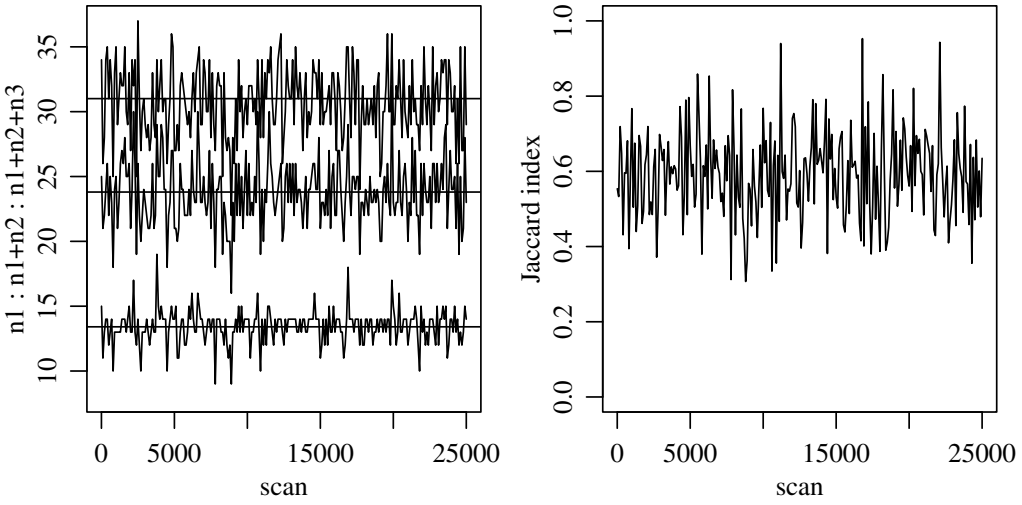

Figure 5: Summaries of the posterior distributions of the clustering. Panel 1 plots the cumulative sample sizes of the three largest groups for every 100th scan of the Markov chain. Panel 2 plots the Jaccard index between $\hat{c}$ and every 100th sampled value of c.

Examination of the Markov chains indicated good mixing and that cluster composition was fairly similar to that of $\hat{c}$ across MCMC scans and chains. Some aspects of this are presented in Figure 5. The first panel gives the cumulative sample sizes of the largest cluster, second largest cluster and third largest cluster for every 100th scan of one of the Markov chains The second panel gives the Jaccard index of similarity between $\hat{c}$ and each 100th sampled value of c. Results for the other Markov chains are similar.

On average over all scans and chains, the sample sizes of the three largest groups were 13.5, 10.4, and 7.2 respectively, making up 31.1 (71%) of the 44 tumors. The ten tumors that were placed into their own outlier groups under $\hat{c}$ were also typically outliers across scans of the Markov chains. The fraction of scans in which these ten tumors were outliers were .97, .96, .93, .92, .91, .89, .86, .84, .74 and .38. Of the remaining 34 tumors, one had a 14% posterior probability of being an outlier and the rest had probabilities of less than 3%. On average across scans and Markov chains, 8.7 tumors (20%) were placed in their own outlier groups.

Data from the three largest groups under $\hat{c}$ are shown in Figures 6, 7 and 8. Hybridization ratios within these groups are generally characterized by having high variances relative to $\boldsymbol{\sigma}^2$ and having some common regions of over or under expression. The main features of the largest group include large, consistent amplification of genetic material on the second arm of chromosome 1, moderate amplification on the first arm of chromosome 16, and deletion on the second arm of chromosome 16. The second largest group also exhibits some amplification on chromosome 1, but has higher variances than the first group and lacks a pattern of gain and loss on chromosome 16. The third largest

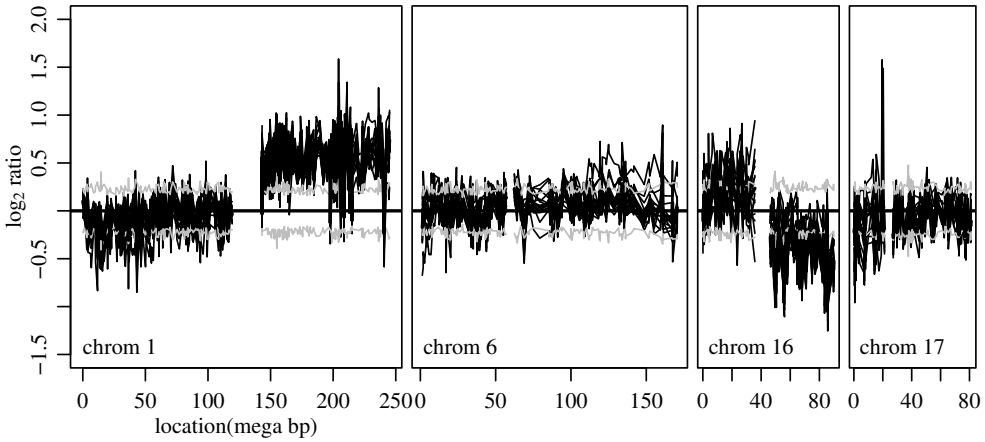

Figure 6: Data from the largest cluster. Gray lines are $\pm 2\hat{\boldsymbol{\sigma}}$.

group exhibited a similar pattern to that of the largest group, but with a smaller shift on the second arm of chromosome 1 and an indication of deletion on the first arm of chromosome 17.

For comparison, data from one of the outlier tumors is shown in Figure 9. This tumor was identified as an outlier in 97% of the saved scans of the Markov chains, which seems like a desirable classification: This tumor distinguishes itself from the others by exhibiting no sign of amplification or deletion anywhere on chromosomes 1, 6 or 16, but has a very clear abnormality pattern on 17, one that is unlike that of the other tumors.

As can be seen from these plots, the data are quite noisy even within a cluster. However, for each of these clusters there exist many regions of the chromosomes in which the hybridization ratios are all either above zero or all below zero. As a quick ad-hoc check, for each of the three groups under $\hat{c}$ we counted the number of locations at which the measurements were either all above zero or all below zero, and compared these counts to the expected number of such locations from randomly sampled subsets of tumors, the subsets having sizes of $n_k \in \{14, 11, 9\}$. The ratios of observed to expected counts were 2.4, 1.2, and 1.5 for the clusters of size 14,11 and 9 respectively (or "p-values" of $< .0001$, .142 and .02). In contrast, when taken together, the set of 10 outliers had an observed to expected ratio of 0.7 (or a "p-value" of .96). This simply checks that the clustering procedure grouped together tumors having similar patterns of over and under expression.

5 Discussion

This paper has developed a model-based method of finding clusters based on subsets of attributes. The types of clusterings this method is designed to identify include, but

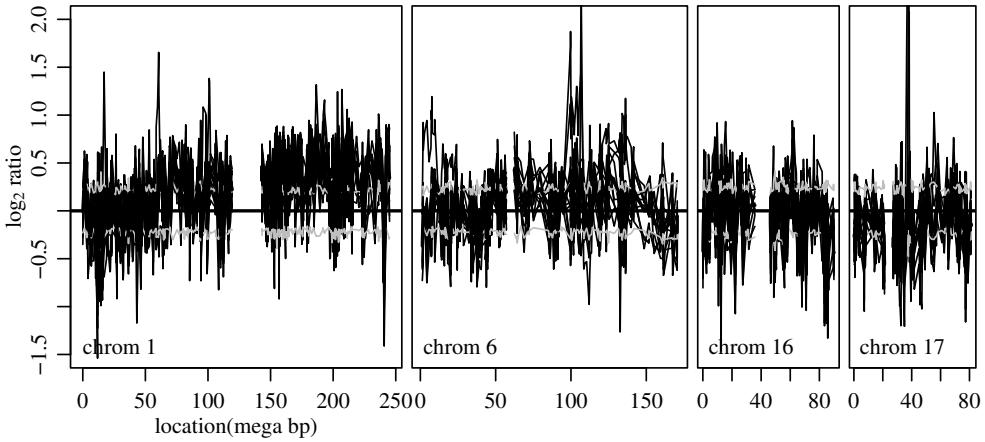

Figure 7: Data from the second largest cluster.

are not limited to, clusterings where all attributes differ among groups and where a fixed subset of attributes all differ among groups, and so this approach is more general than a variable selection procedure. The method also has the feature (shared by all Dirichlet process mixture models) that an object can potentially be put into a group 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 influence the features of the groups. Additionally, the basic method presented here for identifying clusterings based on attribute subsets is extendable to more general data analysis situations. For example, non-normal data can be modeled with exponential family distributions, using $\boldsymbol{\mu}_{(k)} = \boldsymbol{\mu} + \boldsymbol{r}_{(k)} \times \boldsymbol{\delta}_{(k)}$ to represent the canonical parameters.

The method described in this article is based on a nonparametric mixture of sequences of independent normal random variables. As a result, the estimated clustering has the task of representing any bivariate correlation of the attributes as well as higher-order dependence patterns. If the attributes are highly correlated then the number of components required by the mixture model to fit the data might be large. A quick, ad-hoc solution to this is to perform the subspace clustering on the principal components of the data. Alternatively, one can estimate parameters in the model $\boldsymbol{y}_i = \boldsymbol{\mu} + \boldsymbol{r}_{(k)} \times \boldsymbol{\delta}_{(k)} + \boldsymbol{\epsilon}_i$ with $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \Sigma$ being an arbitrary covariance matrix. Posterior calculations for this model are made difficult by the fact that the marginal distribution of the data within a cluster, unconditional on the cluster parameters $\boldsymbol{r}_{(k)}$ and $\boldsymbol{\delta}_{(k)}$ is given by the sum $\sum_{\boldsymbol{r}_{(k)} \in \{0,1\}^m} p(\boldsymbol{r}_{(k)}) p(\{\boldsymbol{y}_i : c_i = k\} | \boldsymbol{r}_{(k)})$ where, as a function of $\boldsymbol{r}_{(k)}$, $p(\{\boldsymbol{y}_i : c_i = k\} | \boldsymbol{r}_{(k)})$ is proportional to $|V|^{1/2} \exp\{-\boldsymbol{\theta} V^{-1} \boldsymbol{\theta}/2\}$ with $V = [R_{(k)} \Sigma^{-1} R_{(k)} + \tau^{-2} I]$, $R_{(k)} = \mathrm{diag}(\boldsymbol{r}_{(k)})$ and $\boldsymbol{\theta}$ is a function of V, $\Sigma, \boldsymbol{\mu}$ and $\{\boldsymbol{y}_i : c(i) = k\}$. Computing this sum over $\boldsymbol{r}_{(k)}$ is impractical for any realistic number of attributes m, and so sampling values of $c(i)$ from its full conditional is not possible. However, one can construct an approximate full conditional for $c(i)$ by replacing Σ with

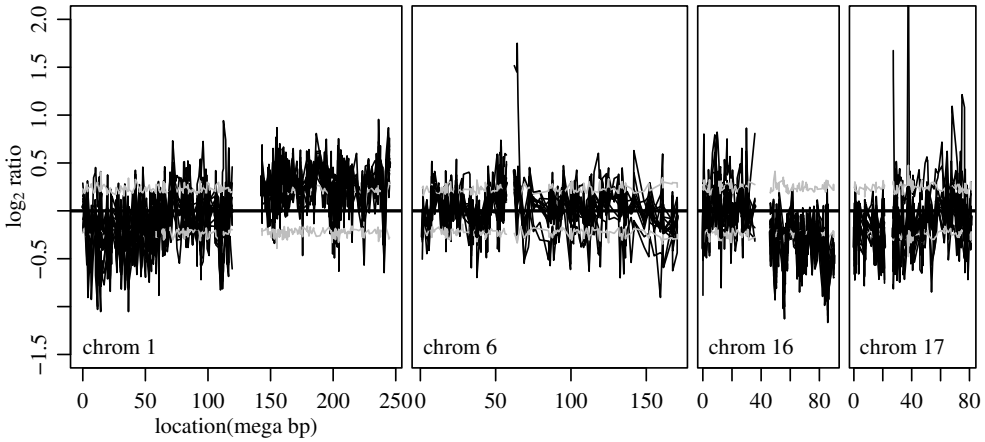

Figure 8: Data from the third largest cluster.

its diagonal and using the proposal distribution of Section 2.2. Having sampled a proposed value of $\boldsymbol{r}$ to go along with the proposed $c(i)$, the new values can be accepted or rejected with the appropriate probability. Model-based subspace clustering of such correlated data, with the possibility of cluster-specific correlation matrices, is a current research area of the author.

Computer code in the R-language and example analyses are available at the author's website http://www.stat.washington.edu/hoff/research.html.

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.

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.

— (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. (1995). "Bayesian density estimation and inference using mixtures." *J. Amer. Statist. Assoc.*, 90(430): 577–588.

— (1998). "Computing nonparametric hierarchical models." In *Practical nonparametric*

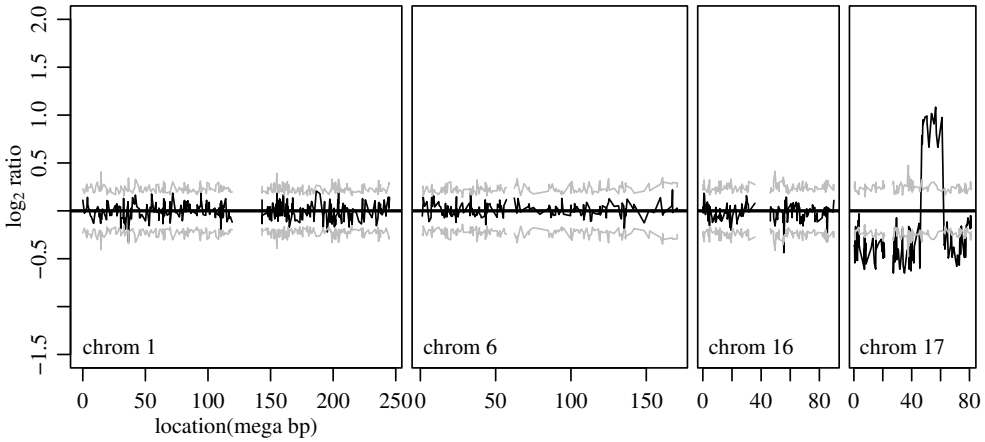

Figure 9: Data from an "outlier".

and semiparametric Bayesian statistics, volume 133 of *Lecture Notes in Statist.*, 1–22. New York: Springer.

Fraley, C. and Raftery, A. E. (2002). "Model-based clustering, discriminant analysis, and density estimation." *J. Amer. Statist. Assoc.*, 97(458): 611–631.

Friedman, J. H. and Meulman, J. J. (2004). "Clustering objects on subsets of attributes." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(4): 815–849. With discussion and a reply by the authors.

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(1): 87–90.

Hoff, P. D. (2004). "Discussion of 'Clustering objects on subsets of attributes' by Friedman and Meulman." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(4): 845–846.

— (To appear). "Subset Clustering of Binary Sequences, with an Application to Genomic Abnormality Data." *Biometrics*.

Jaccard, P. (1912). "The distribution of flora in the alpine zone." *The New Phytologist*, 11: 37–50.

Jain, S. and Neal, R. M. (2004). "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *J. Comput. Graph. Statist.*, 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(431): 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(3): 727–741.

MacEachern, S. N. and Müller, P. (1998). "Estimating mixture of Dirichlet process models." *J. Comput. Graph. Statist.*, 7: 223–238.

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.

Mirkin, B. (1996). *Mathematical classification and clustering*, volume 11 of *Nonconvex Optimization and its Applications*. Dordrecht: Kluwer Academic Publishers.

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *J. Comput. Graph. Statist.*, 9(2): 249–265.

Newton, M. A. (2002). "Discovering combinations of genomic aberrations associated with cancer." *J. Amer. Statist. Assoc.*, 97(460): 931–942.

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)*, 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.

Acknowledgments

This research was supported by National Cancer Institute grant CA077607-04 and Office of Naval Research grant N00014-02-1-1011. 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.