Bayesian Variable Selection in Clustering High-Dimensional Data With Substructure

Michael D. SWARTZ, Qianxing MO, Mary E. MURPHY,
Joanne R. LUPTON, Nancy D. TURNER,
Mee Young HONG, and Marina VANNUCCI

In this article we focus on clustering techniques recently proposed for high-dimensional data that incorporate variable selection and extend them to the modeling of data with a known substructure, such as the structure imposed by an experimental design. Our method essentially approximates the within-group covariance by facilitating clustering without disrupting the groups defined by the experimenter. The method we adopt simultaneously determines which expression patterns are important, and which genes contribute to such patterns. We evaluate performance on simulated data and on microarray data from a colon carcinogenesis study. Selected genes are biologically consistent with current research and provide strong biological validation of the cluster configuration identified by the method.

Key Words: Bayesian inference; Designed experiments; Microarray analysis.

1. INTRODUCTION

The availability of high throughput data-collection techniques, especially in bioinformatics, offer statisticians a growing challenge: How to analyze data when the number p of variables far outnumbers the number of samples. These scenarios lead to the curse of dimensionality (Scott 1992), which essentially means that the data seem sparse across the p-dimensional space, and the usual asymptotics for frequentist theory commonly do not apply. Many researchers have therefore turned to Bayesian techniques for mining and analysis of these high-dimensional data to search for differentially expressed genes (Do et al. 2006; Sebastiani et al. 2003).

Michael D. Swartz is Instructor, Department of Epidemiology, M.D. Anderson Cancer Center, Houston, TX 77030. Qianxing Mo is Biostatistician, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10021. Mary E. Murphy is Research Associate, Texas A&M University, College Station, TX 77843. Joanne R. Lupton is Professor and Nancy D. Turner is Associate Professor, Nutrition and Food Science Department, Texas A&M University, College Station, TX 77843. Mee Young Hong is Assistant Professor, School of Exercise and Nutritional Sciences, San Diego State University, San Diego, CA 92182. Marina Vannucci is Professor, Department of Statistics, Rice University, Houston, TX 77251 (E-mail for correspondence: *marina@rice.edu*).

© 2008 American Statistical Association and the International Biometric Society

Journal of Agricultural, Biological, and Environmental Statistics, Volume 13, Number 4, Pages 407–423
DOI: 10.1198/108571108X378317

Many studies involving microarrays have substructure inherent to the data. This is the case, for example, with designed experiments that group the data within treatments. Recently, Bayesian methods have appeared in the literature that propose an approach to gene discovery in designed experiments [see Efron et al. (2001); Newton et al. (2001); Ibrahim et al. (2002); Kendziorski et al. (2003)]. Efron et al. (2001) developed a nonparametric approach for microarray analysis that uses permutation methods to estimate the null distribution of the summary statistics for gene expression. Nonparametric permutation methods, however, can be inconsistent with a small number of replicates per group. Ibrahim et al. (2002) introduced a parametric two-component mixture model that combines a point mass at a threshold value with a normal distribution component. This method applies to only two groups. Newton et al. (2001) also used a two-component mixture model for gene expression, assuming the components are parametric gamma distributions. Their original method applied to two groups and was recently extended to compare differentially expressed genes when considering multiple groups; see Kendziorski et al. (2003). This method works well in identifying patterns of differential expression, but it requires the enumeration of all possible patterns or some external justification to reduce the patterns.

Other approaches for microarray data without substructure use clustering analysis. In these studies, the focus is classifying individuals based on their gene expression values. To cluster individuals effectively, researchers must reduce the number of expression values because including large numbers of uninformative variables can greatly interfere with the recovering of the true cluster structure (see Brusco and Cradit 2001; Tadesse et al. 2005 and references therein). Therefore, clustering methods for microarrays must incorporate the information in the data to select the gene expression values that drive the clusters. Bayesian variable selection techniques applied to clustering offer a comprehensive method to both select the most informative genes (variables) and recover cluster structure. In addition, these methods allow the true number of clusters to be unknown.

Two recently introduced techniques combine Bayesian model selection techniques with model based clustering. The first technique, described by Tadesse et al. (2005), introduces a novel Bayesian approach to clustering high-dimensional data. This procedure jointly estimates the cluster patterns in the data and selects the variables that best define those patterns via the use of stochastic search and reversible jump Markov chain technologies. The second approach, described by Raftery and Dean (2006), uses the same mixture model approach with the model selection approach driven by Bayes factors and a greedy search algorithm. This algorithm is simplified by using the BIC to approximate Bayes factors. Both methods simultaneously recover the cluster structure in the data and select the individual variables that best define the cluster structure.

In this article we extend the work of Tadesse et al. (2005) to the modeling of data with a known substructure, such as the structure imposed by an experimental design. By jointly clustering the data and selecting the discriminatory variables, the method determines both which experimental treatments are important, and which genes have the most differentiating expression values affected by the treatments. By essentially approximating the within-group covariance, our approach facilitates clustering without disrupting the groups defined by the experimenter. This extension applies to any data with substructure, and in particular to microarrays used in preclinical animal designs, an important area of medical research, where often the differentiating genes are of more interest than the clusters they define.

The rest of the article is organized as follows: We close this Introduction with a brief description of the microarray study that has motivated this work. Section 2 describes the proposed extension of the model of Tadesse et al. (2005) to handle data with substructure. Performances of the method are exemplified on simulated data in Section 3 and on data from our case study in Section 4. Section 5 concludes the article.

1.1 MOTIVATING DATA

We briefly describe the experimental design of the specific microarray dataset from a colon carcinogenesis study that motivated the development of the method proposed in this article. Sprague-Dawley rats were prescribed a diet rich either in corn oil or in fish oil as the primary source of fatty acids, and were treated either with Dextran sodium sulfate (DSS) with 48-hour recovery before being sacrificed, or not treated at all (controls). Thus, by design we have four groups of rats: fish oil, control; corn oil, control; fish oil, DSS with recovery; corn oil, DSS with recovery. The original microarrays consisted of 54,184 genes [more details regarding study design can be found in Hong et al. (2005)]. The goal of the study is to discover a small subset of genes associated with the treatments that can be investigated further using biological assays.

2. METHODS

We focus on recently developed Bayesian methods for mixture models that simultaneously cluster the samples and select the variables and discuss how those methods can be adjusted to incorporate the correlation within subgroups.

2.1 MODEL-BASED CLUSTERING

Let $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n})$ denote n-independent p-dimensional observations from G underlying subpopulations. Clustering the n samples can be modeled as a mixture of the G subpopulation models:

$$f(\mathbf{x}_i|\mathbf{w}, \boldsymbol{\theta}) = \sum_{k=1}^{G} w_k f(\mathbf{x}_i|\boldsymbol{\theta}_k), \tag{2.1}$$

where $f(\mathbf{x}_i|\boldsymbol{\theta}_k)$ is the density for the observation from the kth subpopulation and $\mathbf{w}$ is the vector of nonnegative component weights w_k that sum up to 1, and $\boldsymbol{\theta}$ denotes the distribution parameters. The model is completed with a latent vector $\mathbf{y} = (y_1, \ldots, y_n)$ with elements indicating to which subpopulation component each observation belongs to. If we let the y_i's be iid, with $p(y_i = k) = w_k$, and define our subpopulation distributions to be multivariate normal with mean vector $\boldsymbol{\mu}_k$ and variance matrix $\boldsymbol{\Sigma}_k$, then we can model each sample i, conditional on y_i, as

$$(\mathbf{x}_i|y_i = k, \mathbf{w}, \boldsymbol{\theta}) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.2}$$

2.2 ADJUSTING FOR SUBSTRUCTURE

In order to account for substructure in the data, we consider the covariance structure of the data. If there are known subgroups within the data, as in a designed experiment, there will be within-group covariance and between-group covariance. The within-group and between-group covariances will obviously be different. The original method described by Tadesse et al. (2005) treats the covariance between the individuals as the same, regardless of whether the individuals are in the same group or in different groups. Therefore we can improve the method by accounting for this within-group and between-group difference in covariance. One way to do this is to construct a formal Bayesian model using blocked covariance matrices in the likelihood and/or priors that adequately reflect the within- and between-group variance structures. This approach, however, requires at least a $p \times p$ blocked covariance matrix and introduces a large number of parameters, especially in scenarios where $p \gg n$, bringing instability into the model. To avoid this, we approximate the within-group covariance structure and modify the cluster allocation proposal to reflect subgroups in the data. This approximation also makes the extension of the likelihood straightforward.

Here we impose structure on the data via the definition of the cluster allocation vector, $\mathbf{y}$. This vector now has elements indicating subgroups, that is, blocks of observations, rather than individual observations. Thus, all individuals in a given subgroup will be always assigned to the same cluster. When clustering the data, the original subgroups may collapse into bigger groups but they cannot be further divided into smaller groups.

2.3 LIKELIHOOD

In order to do variable selection we follow Tadesse et al. (2005) and employ a latent indicator to select the discriminatory gene expression values that best cluster the data. Let $\boldsymbol{\gamma}$ be such an indicator vector, where $\gamma_j = 1$ if the jth expression level (variable) contributes to differentiating the clusters and $\gamma_j = 0$ if the jth variable is nondiscriminatory. This generates a likelihood that is a product of the mixture model (2.1) and a single multivariate normal distribution that models the nondiscriminating variables. Following the notation used by Tadesse et al. (2005), we use $(\boldsymbol{\gamma})$ and $(\boldsymbol{\gamma}^{\boldsymbol{c}})$ to index the discriminating variables and those that do not discriminate, respectively.

Recall that $p(y = k) = w_k$. In the likelihood calculation we need to compute the exponent of the term corresponding to the weights w_k based on the number of subgroups belonging to cluster k (denoted m_k), rather than on the number of individuals in cluster k

(denoted n_k). The likelihood function is as follows:

$$
\begin{aligned}
&L(G, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta}, \boldsymbol{\Omega}, |\mathbf{X}, \mathbf{y}) \\
&= (2\pi)^{-(p-p_\gamma)n/2} |\boldsymbol{\Omega}_{(\gamma^c)}|^{-n/2} \\
&\quad \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_{i(\gamma^c)} - \boldsymbol{\eta}_{(\gamma^c)})^T \boldsymbol{\Omega}_{(\gamma^c)}^{-1} (\mathbf{x}_{i(\gamma^c)} - \boldsymbol{\eta}_{(\gamma^c)}) \right\} \\
&\quad \times \prod_{k=1}^{G} (2\pi)^{-p_\gamma n_k/2} |\boldsymbol{\Sigma}_{k(\gamma)}|^{-n_k/2} w_k^{m_k} \\
&\quad \times \exp\left\{ -\frac{1}{2} \sum_{x_i \in C_k} (\mathbf{x}_{i(\gamma)} - \boldsymbol{\mu}_{k(\gamma)})^T \boldsymbol{\Sigma}_{k(\gamma)}^{-1} (\mathbf{x}_{i(\gamma)} - \boldsymbol{\mu}_{k(\gamma)}) \right\}. \qquad (2.3)
\end{aligned}
$$

In the Equation (2.3), C_k denotes the kth mixture component, $\boldsymbol{\mu}_k$ denotes its mean, and $\boldsymbol{\eta}$ the mean of the nondiscriminatory distribution. Likewise, $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Omega}$ denotes the variance-covariance matrices. Notice that our likelihood depends on n, the total number of samples, n_k, the number of samples allocated to cluster k and also on m_k, the total number of subgroups allocated to component k, unlike the likelihood of Tadesse et al. (2005), which is only a function of n and n_k.

2.4 PRIOR MODEL

We adopt the same prior model as in Tadesse et al. (2005). The indicator variables $\boldsymbol{\gamma}$ are modeled as independent Bernoulli random variables, with common probability parameter φ. We elicit φ as the expected proportion of the variables that will be discriminating a priori. A natural prior for the number of clusters, G, is a truncated Poisson, with rate parameter λ:

$$
P(G = g) = \frac{e^{-\lambda}\lambda^g/g!}{1-(e^{-\lambda}(1+\lambda)+\sum_{j=G_{\max}+1}^{\infty}(e^{-\lambda}\lambda^j)/j!)}, \qquad g = 2, \ldots, G_{\max}. \qquad (2.4)
$$

For the vector of component weights, we use a symmetric Dirichlet prior, $\mathbf{w}|G \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$.

For the component means and variances, as well as the mean and variance of the nondiscriminating variables, we use the usual conjugate priors.

$$
\begin{aligned}
\boldsymbol{\mu}_{k(\gamma)}|\boldsymbol{\Sigma}_{k(\gamma)}, G &\sim N(\boldsymbol{\mu}_{0(\gamma)}, h_1\boldsymbol{\Sigma}_{k(\gamma)}), \\
\boldsymbol{\eta}_{(\gamma^c)}|\boldsymbol{\Omega}_{(\gamma^c)} &\sim N(\boldsymbol{\mu}_{0(\gamma^c)}, h_0\boldsymbol{\Omega}_{(\gamma^c)}), \\
\boldsymbol{\Sigma}_{k(\gamma)}|G &\sim IW(\delta; \mathbf{Q}_{1(\gamma)}), \\
\boldsymbol{\Omega}_{(\gamma^c)} &\sim IW(\delta; \mathbf{Q}_{0(\gamma^c)}).
\end{aligned} \qquad (2.5)
$$

Keeping consistent notation with Tadesse et al. (2005), here, IW(δ; $\mathbf{Q}_{1(\gamma)}$) denotes the inverse-Wishart distribution, with shape parameter $\delta = n - p_\gamma + 1$, dimension p_γ, degrees of freedom n, and mean $\mathbf{Q}_{1(\gamma)/\delta-2}$. Also, as in Tadesse et al. (2005), we use $\delta = 3$ to denote an uninformative prior and define $\mathbf{Q}_1 = 1/\kappa_1\mathbf{I}_{p\times p}$ and $\mathbf{Q}_0 = 1/\kappa_0\mathbf{I}_{p\times p}$, where κ_1 and κ_0 are defined respectively as proportional to the upper and lower decile of the $n-1$ nonzero eigen values of cov($\mathbf{X}$). These choices follow the guidelines given by Tadesse et al. (2005)

who pointed out the sensitivity of their procedure to the hyperparameters of the covariance matrix. Some sensitivity to the parameter choices is typical of any model-based clustering method. For the mean parameters, each element of $\boldsymbol{\mu}_{\mathbf{0}}$ was set to the midpoint of the range of the variable, and h_0 and h_1 were chosen arbitrarily large, between 10 and 1,000, for flat priors. For more details on regarding the hyper-prior parameters $\boldsymbol{\mu}_{0(\boldsymbol{\gamma})}$, $\boldsymbol{\mu}_{0(\boldsymbol{\gamma}^c)}$, h_1, h_2, δ, $\mathbf{Q}_{1(\boldsymbol{\gamma})}$, and $\mathbf{Q}_{0(\boldsymbol{\gamma}^c)}$ (see Tadesse et al. 2005).

2.5 MCMC ALGORITHM

The mean and variance parameters were expertly integrated out in Tadesse et al. (2005), and our modification, described above, is constant with respect to these parameters, and therefore does not change the integration calculations. Thus, even after accounting for substructure, we only need to update the parameters $(\boldsymbol{\gamma}, \mathbf{w}, \mathbf{y}, G)$. We were able to use much of the MCMC machinery from Tadesse et al. (2005), and use the cluster indicator vector $\mathbf{y}$ based on subgroups, and the new likelihood (2.3) in calculating the necessary probabilities. Following Tadesse et al. (2005), we simulate from the posterior using a hybrid Gibbs sampler and Metropolis–Hastings algorithm that iterates sampling from the following distributions:

$$f(\mathbf{y}|G, \mathbf{w}, \boldsymbol{\gamma}, \mathbf{X}) \propto f(\mathbf{X}, \mathbf{y}|G, \mathbf{w}, \boldsymbol{\gamma}), \tag{2.6}$$

$$f(\boldsymbol{\gamma}|G, \mathbf{w}, \mathbf{y}, \mathbf{X}) \propto f(\mathbf{X}, \mathbf{y}|G, \mathbf{w}, \boldsymbol{\gamma})p(\boldsymbol{\gamma}|G), \tag{2.7}$$

and

$$\mathbf{w}|G, \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X} \sim \text{Dirichlet}(\alpha + n_1, \ldots, \alpha + n_G). \tag{2.8}$$

The vector $\boldsymbol{\gamma}$ is updated via (2.7) using the Metropolis search algorithm that has now become quite standard in variable selection; see Sha et al. (2004) and Tadesse et al. (2005). At a single iteration the vector $\boldsymbol{\gamma}$ is updated either by swapping two of its elements or by randomly selecting one element and changing its value from 0 to 1 or 1 to 0. The cluster allocation vector $\mathbf{y}$ is updated element by element using a Gibbs sampling strategy via Equation (2.6). According to our modified model, each element of $\mathbf{y}$ corresponds to an experimental group. The full conditional probability that the ith experimental group is in the kth cluster is therefore calculated as

$$f(y_i = k|\mathbf{X}, \mathbf{y}_{(-\mathbf{i})}, \boldsymbol{\gamma}, \mathbf{w}, \mathbf{G}) \propto \mathbf{f}(\mathbf{X}, y_i = k, \mathbf{y}_{(-\mathbf{i})}|\mathbf{G}, \mathbf{w}, \boldsymbol{\gamma}). \tag{2.9}$$

Here, $\mathbf{y}_{(-i)}$ is the standard notation denoting the vector of cluster assignments for all subgroups except the ith subgroup.

The weights are updated by Gibbs sampler via Equation (2.8). We simplify the calculations by sampling independent gamma random variables with common scale and shape parameters $\alpha + n_1, \ldots \alpha + n_G$, and scaling the random variates to sum to 1. As in the original model formulation of Tadesse et al. (2005) we allow the number of clusters, G, to be unknown and update this parameter using reversible jump Markov chain Monte Carlo (RJMCMC) technology (Green 1995; Richardson and Green 1997). Our RJMCMC construction updates G using a split/merge cluster move, and a birth/death move as in Tadesse et al. (2005). However, to calculate the acceptance ratio, we use the new likelihood (2.3), where the weights of the cluster depends on the number of experimental groups. This accounts for using the experimental subgroups as items to be clustered.

2.6 POSTERIOR INFERENCE

In order to make inference from the posterior samples, we first use the method proposed by Stephens (2000) to resolve cluster identifiability. Once the clusters are suitably relabeled to be consistent across all iterations, we calculate frequency approximations to the posterior probabilities since our quantities of interest are multinomial or binomial random variables. It has been our experience that these frequency estimates are more robust to correlation that may be present in the Markov chain than calculating the marginal posterior probabilities—especially when analyzing real data (Kim et al. 2006). For inference on cluster memberships, we condition on the most probable number of clusters and count how many iterations each experimental group appears in each cluster. For inference on the variables, we count the number of iterations that each variable is selected and divide that by the total number of iterations kept after burn in. For the simulations below, we found similar distributions using the posterior probability calculations detailed in Tadesse et al. (2005) and our frequency approximations.

3. SIMULATION STUDY

We evaluated performances of our method using two simulated datasets. For comparison purposes, we first used the same simulation setting as in Tadesse et al. (2005). Then, we simulated a different dataset with more substructure.

3.1 SCENARIO 1

The first dataset was simulated from the model

$$\begin{aligned} x_{ij} \sim\ & I_{[1\le i\le 4]}N(\mu_1,\sigma_1^2) + I_{[5\le i\le 7]}N(\mu_2,\sigma_2^2) \\ & + I_{[8\le i\le 13]}N(\mu_3,\sigma_3^2) + I_{[14\le i\le 15]}N(\mu_4,\sigma_4^2) \end{aligned} \tag{3.1}$$

for $i = 1,\ldots,15$, $j = 1,\ldots,20$. We used $\mu_1 = 5$, $\mu_2 = 2$, $\mu_3 = -3$, $\mu_4 = -6$, $\sigma_1^2 = 1.5$, $\sigma_2^2 = 0.1$, $\sigma_3^2 = 0.5$, and $\sigma_4^2 = 2$. We simulated an additional 480 variables as white noise, representing non-discriminating variables. The means and variances for this simulation were chosen such that the clusters were fairly well separated, and did not overlap for all practical purposes. Our purpose was to evaluate the performance of recovering original cluster structure while removing noise, or nondiscriminating variables. To impose substructure, we arbitrarily broke the third cluster consisting of six points into two subgroups of three points each. Then we assigned each of the other clusters to their own subgroup. We analyzed these data with both methods, the original method of Tadesse et al. (2005) and our method that allows for substructure. For each method, we ran four MCMC chains that started from different starting points, using the same prior setting. Since our method requires a starting cluster configuration, and starting variable configuration, we use two starting cluster configurations, each paired with two starting variable configurations. The four starting configurations can be denoted as ordered pairs (initial number of randomly selected variables, initial clusters): (1 randomly selected initial variable, 2 initial clusters); (1 randomly selected initial variable, 5 initial clusters); (100 randomly selected initial variables, 2 initial clusters); and (100 randomly selected initial variables, 5 initial clusters). When starting with 2 clusters, we allocated subgroups to each cluster so that half of the points are spread across the 2 clusters, while configurations with 5 initial clusters started with each subgroup in its own cluster. For both methods, we specified priors according to the guidelines described above and standardized the data by dividing each variable by its range. We also permuted the columns. Then κ_1 and κ_0 were chosen to be proportional to the last and first decile of the eigenvalues of the covariance matrix of the data. The prior mean for both the cluster mixtures and the nondiscriminating distribution was set as $\mu_{0j} = (1/2)\text{range}(x_j) + \min(x_j)$. For the weights, we assigned the symmetric Dirichlet distribution with prior parameter $\alpha = 1$, and used the truncated Poisson distribution with rate parameter $\lambda = 5$, and set the prior for the Bernoulli distribution on γ such that $p\varphi = 10$, denoting that a priori we expect on average 10 variables to be discriminating. We used 100,000 iterations with a burn-in of 40,000.

Both the original method and our method that accounts for substructure correctly recovered the cluster structure while results on the variable selection were somewhat different. For each chain, we considered the selected variables as those with marginal posterior probability greater than or equal to 0.5. Our method, using substructure, always selected the true variables, and only two of the chains selected one additional noisy variable. In contrast, the original method always missed one of the true variables. In Tadesse et al. (2005) the authors also compared their method to a detailed analysis of data simulated from this same model using the COSA algorithm of Friedman and Meulman (2004), which performs variable selection in the context of hierarchical clustering. Using the COSA approach, neither the single, average, nor complete linkage options for the hierarchical clustering were able to recover the true structure simulated in the data. As explained by the authors, this performance of COSA could be due to the fact that the method is designed to find clusters for which the discriminating variables have small variance. Since our method performs similarly to Tadesse et al. (2005) on this dataset, which has the same structure as that which was analyzed using COSA, we do not repeat the comparison to COSA.

3.2 SCENARIO 2

We then looked at performances on a second simulated dataset, that had more substructure. We simulated samples from the following model:

$$\begin{aligned} x_{ij} \sim\ & I_{[1\leq i\leq 10]} N(\mu_1, \sigma_1^2) + I_{[11\leq i\leq 20]} N(\mu_2, \sigma_2^2) \\ & + I_{[21\leq i\leq 30]} N(\mu_3, \sigma_3^2) \end{aligned} \tag{3.2}$$

for $i = 1, \ldots, 30$, $j = 1, \ldots, 20$. We simulated an additional 480 variables as white noise, representing nondiscriminating variables. We therefore had a 500-dimensional dataset with 30 samples that define 3 groups of 10 individuals on 20 discriminatory variables. Within

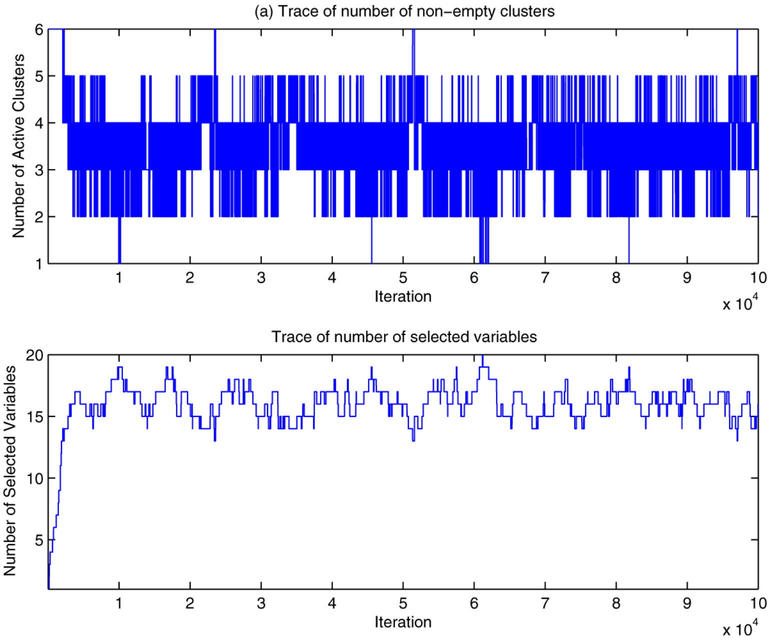

Figure 1. Simulated data: MCMC traces for a chain starting with 1 randomly selected variable and each subgroup as its own cluster. (a) Number of active components or clusters. (b) Number of selected variables.

these groups of 10, we arbitrarily assigned 5 individuals to a subgroup. Again, here we choose well-separated distributions with an equal number of samples in each subpopulation to better mirror an ANOVA type design.

As in the previous example, we specified priors as suggested by Tadesse et al. (2005) and standardize the variables by their range. We used 4 different MCMC chains with the same parameter settings and different starting pairs of (initial randomly selected variable, initial clusters) configurations. Chain 1 started at (100 variables, 6 clusters), chain 2 started at (1 variable, 6 clusters), chain 3 started at (1 variable, 2 clusters), and chain 4 started at (100 variables, 2 clusters). When starting at 6 clusters, each subgroup started as its own cluster, and when starting the chain at 2 clusters, we assigned 3 subgroups to each cluster. Each chain consisted of 100,000 iterations, with 40,000 iterations as burn-in.

Our method performed well, always recovering the true cluster structure, and recovering at least 95% of the true variables. We present summary plots associated with one of the chains. Plots for the other chains were similar. Figure 1(a) shows the trace plot for the number of visited components G and Figure 1(b) the trace plot for the number of variables in the visited models. We see that the MCMC chain mixes well, quickly stabilizing to stochastically hover around the correct number of clusters and variables in the burn in period. Table 1 lists the posterior probability for each number of components $P(G|\mathbf{X})$ for all 4 chains. In all chains there is strong evidence for $G = 3$. Figure 2 shows the probability of each of the 30 samples to belong to one of these 3 clusters, for the same chain represented in Figure 1. These probabilities are estimated as normalized frequencies of

Table 1. Simulated data: Posterior distribution of G, for all 4 chains.

	Chain 1	Chain 2	Chain 3	Chain 4
k	$p(G = k\|\mathbf{X})$	$p(G = k\|\mathbf{X})$	$p(G = k\|\mathbf{X})$	$p(G = k\|\mathbf{X})$
1	0	< 0.001	0	< 0.001
2	0.073	0.086	0.083	0.114
3	**0.653**	**0.680**	**0.654**	**0.689**
4	0.238	0.209	0.216	0.182
5	0.029	0.002	0.027	0.013
6	0.007	0.005	0.019	0.023

each point appearing in each cluster throughout the MCMC iterations. According to these results, each subgroup is most frequently allocated to the proper mixture component from which it was simulated. Notice that the estimated membership probability of all points within a subgroup is the same, reflecting our substructure allocation method.

As for variable selection, using a decision rule of 50% posterior probability of inclusion, chains 2, 3, and 4 recovered all variables without any false positives, while chain 1 missed variable 6. Also, when using the frequency of inclusion in the models visited by the chain, only chain 1 failed to recover all the true variables. This chain, however, recovered all true variables without introducing false positives if we reduce the selection criterion to occurring in 40% of the models with three mixture components visited by the stochastic

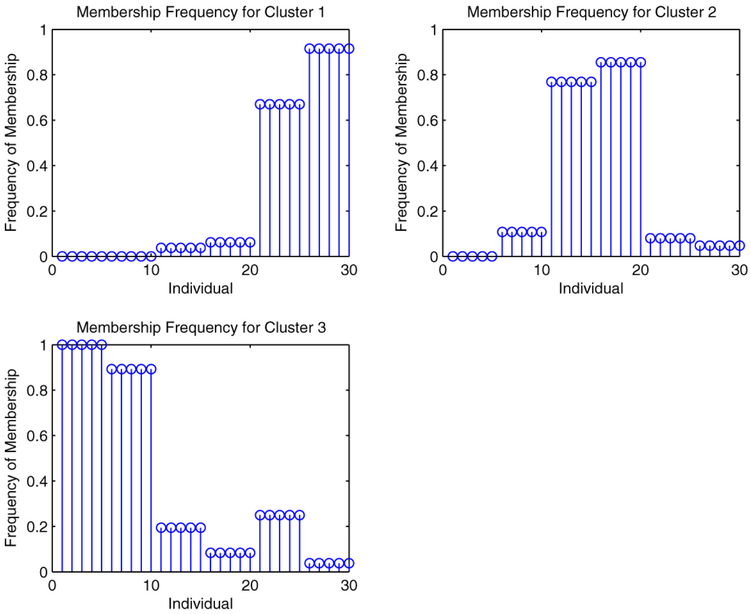

Figure 2. Simulated data: Cluster memberships.

search. Indeed, none of the chains introduced any false positives, even at the reduced 40% posterior marginal probability of inclusion criterion.

4. COLON CARCINOGENESIS STUDY

We finally present results on a microarray study on colon carcinogenesis.

4.1 DATA PREPROCESSING

Recall from the introduction that our data consist of four groups of rats by design: fish oil, control; corn oil, control; fish oil, DSS with recovery; corn oil, DSS with recovery, and the original microarrays consisted of 54,184 genes. For preprocessing, data were normalized by using the global median method. That is, for each array, the expression value of each gene was divided by the median expression value of the expressions on the array. An initial ANOVA analysis was performed, whose details and results will be described elsewhere. As a result of this analysis, 636 genes were found to be significant for a diet by treatment interaction at the 0.05 level. Here we apply our method to these 636 expression values, with the purpose of further refining the gene discovery.

4.2 ANALYSIS

To prepare for clustering, we divided each of the 636 median normalized gene expression values by the range of that gene. Next, we set our prior parameters using the same guidelines as described in the simulated examples. We chose $\kappa_0 = 0.023172$ and $\kappa_1 = 0.097160$, proportional to the first and last decile of the nonzero eigenvalues as our covariance parameters. The prior means for both cluster mixtures and the nondiscriminating distributions were set as $\mu_{0j} = (1/2)\text{range}(x_j) + \min(x_j)$ We set the symmetric Dirichlet distribution prior parameter as $\alpha = 1$, the truncated Poisson distribution with prior rate parameter $\lambda = 5$; and we set the prior probability for the Bernoulli distribution on γ such that $p\varphi = 10$. Here we defined the prior covariance matrices $\mathbf{Q}_{\mathbf{0}(\boldsymbol{\gamma})}$ and $\mathbf{Q}_{\mathbf{1}(\boldsymbol{\gamma}^c)}$ to be diagonal matrices with diagonal elements equal to the variances of each gene. Incorporating empirical variances has been shown to improve variable selection; see Swartz et al. (2006). We ran two MCMC chains. Both chains were run for 1,000,000 iterations, using the last 60,000 iterations for inference and the rest were considered burn-in. The first chain started with 100 randomly selected genes, and using each subgroup as an initial cluster. The second chain started with 50 randomly selected genes, and using two clusters: corn control with corn DSS recovery, and fish control with fish DSS recovery rats.

We performed the final inference using the pooled sets of samples from the two MCMC chains. After pooling the chains, our method grouped the data into two clusters and selected 17 genes using the marginal median model as cut off. Cluster membership probabilities for each rat are reported in Table 2. These clearly separate control rats from treated rats.

For comparison, we applied FDR multiple testing correction Benjamini and Hochberg (1995) to the ANOVA p-values for each gene. This is a standard method commonly used in microarray analysis. This method detected 243 genes as significant at the 0.05 level. Three

Table 2. Real data: Probability of cluster memberships

Treatment	P(member of cluster 1)	P(member of cluster2)
corn control	0.679	0.321
corn control	0.679	0.321
corn control	0.679	0.321
corn DSS recover	0.255	0.745
corn DSS recover	0.255	0.745
corn DSS recover	0.255	0.745
fish control	0.937	0.063
fish control	0.937	0.063
fish control	0.937	0.063
fish DSS recover	0.188	0.812
fish DSS recover	0.188	0.812
fish DSS recover	0.188	0.812

of the 17 genes we identified were also selected by the FDR method. A level of 0.1 detected a larger number of genes, and included 7 of the 17 genes we identified. The selection of a small set of genes is advantageous here. A small number of selected genes is appealing to biologists because they constitute a manageable set of candidates on which further studies can be performed via biological assays. Of course, if necessary, more genes can be selected by our Bayesian method by lowering the threshold of the 50% median model we used.

4.3 BIOLOGICAL FINDINGS

A heat map of the 17 genes selected as discriminatory can be seen in Figure 3. To date, only two of these genes have a listed function in the NCBI public database: cathepsin C (GE20388) and aquaporin 7 (GE20555). We recall that the rats were treated with DSS to induce inflammation, since the DSS treated rodent is a well-established experimental inflammation model to research inflammatory bowel disease and ulcerative colitis (Cooper et al. 1993; Dieleman et al. 1996; Okayasu et al. 1990; Shimizu et al. 2001). One of the genes selected by our method, cathepsin C, is a widely expressed lysosomal cystein protease, that plays an important role in inflammation, and induces the development of collagen-induced arthritis in mice. This mouse model for inflammation and arthritis shares many features with human rheumatoid arthritis (Hu and Pham 2005). In concordance with the biology, cathepsin C exhibits elevated expression in the DSS treated rats across both diets (cluster 2) versus the untreated rats (cluster 1). The second gene, aquaporin-7 is expressed in the epithelial cells of both the small and large intestines, and it resides in the biological pathway for intestinal function and/or fluid homeostasis (Laforenza et al. 2005). Previous research has shown aquaporin-7 to be down regulated in patients with ulcerative colitis, a known risk factor for colon caracinogenesis (Hardin et al. 2004) . In these data, aquaporin-7 had lower expression levels in the inflamed cells, or the DSS cluster (cluster 2) compared to the untreated cluster (cluster 1), consistent with prior research.

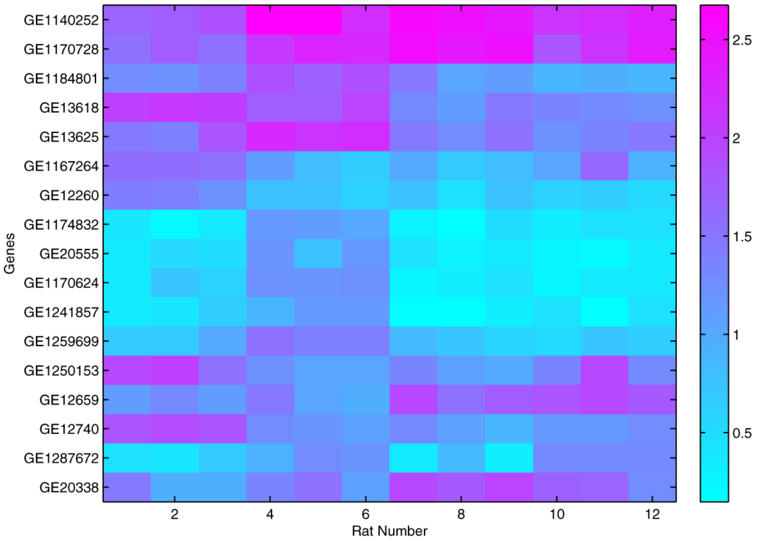

Figure 3. Real data: Heatmap of expression values for the selected genes. The rows represent the 17 genes, and the columns represents the rats. Rats 1-6 form cluster 1 and rats 7-12 cluster 2. Rat numbers 1,2 and 3 are corn diet, no treatment rats; rats 4, 5, and 6 are fish diet, no treatment rats; the others are the DSS treated rats.

Of the remaining 15 selected genes, 6 have been described as part of different pathways in the Gene Ontology database. One gene, GE1170624, codes molecules related to cell-cell adhesion. This gene was underexpressed in the DSS treated animals (cluster 2), which is also consistent with the literature: inflamed colon cells are known to have reduced adhesion molecules (Hanby et al. 1996; Karayiannakis et al. 1998; Syrigos and Karayiannakiks 2006). Another gene, GE12260, exhibits reduced expression in the DSS treated rats (cluster 2) versus the untreated rats (cluster 1). This gene codes for a ubiquitin ligase, and deficiency of ubiquitin ligases (underexpression) is expected under abnormal immune responses such as malignancy and inflammation (Liu 2004). Yet another gene, GE12659, codes for GTPase and is over-expressed in the inflamed DSS treated colon compared to the control cluster. GTPases are a large family of enzymes that play critical roles in signal transduction, are known to regulate adhesion and proliferation, and also play a role in inflammatory responses including intestinal inflammation (Segain et al. 2003). This gene is known to have altered signaling in inflammatory bowel disease patients (Heinzlmann et al. 2002), and to be up-regulated in ulcerative colitis patients (Ierardi et al. 2001).

The three remaining genes have only been shown to have an indirect link with colon cancer through function. GE1167264 is related to the SWI/SNF family and Helicase-Like Transcription factor (HLTF), which are involved in chromatin remodeling. Disruption of genes in chromatin remodeling, including the SWI/SNF family have been found in cancer cells (Moinova et al. 2002), and loss of HTLF expression was noted in nine of 34 colon cancer cell lines. GE1259699 is involved in insulin growth factor receptor (IGF-R) signaling. IGF-R is frequently over-expressed in colon cancer cells compared to normal colon cells (Khandwala et al. 2000; Bustin and Jenkins 2001; Hakam et al. 1999). The final gene

Table 3. Real data: List of selected genes and their posterior probabilities

Gene name	P(diff exp)	cluster 1 mean (no treatment)	cluster 2 mean (DSS and recovery)
GE1140252	0.529	0.886	0.980
GE1167264	0.520	1.678	1.375
GE1170624	0.565	0.344	0.128
GE1170728	0.584	0.865	1.040
GE1174832	0.526	5.316	2.730
GE1184801	0.536	3.348	2.270
GE12260	0.545	14.962	9.060
GE1241857	0.530	0.283	0.106
GE1250153	0.550	1.207	1.107
GE1259699	0.554	2.350	1.413
GE12659	0.571	3.446	5.335
GE12740	0.521	0.780	0.590
GE1287672	0.539	0.0824	0.093
GE13618	0.519	1.655	1.101
GE13625	0.510	11.851	8.769
GE20338	0.513	1.508	2.123
GE20555	0.512	0.749	0.308

of the 6, GE128762, is related to normal ionic transport and channel activity related to normal colonic function, and is therefore expected to be disregulated in DSS-induced colonic inflammation (Seidler et al. 2006). The remaining genes, yet uninvestigated, are listed in Table 3.

The description above highlights the fact that our method has successfully selected genes that are biologically consistent with current research and that provide strong biological validation of the cluster configuration suggested by the method.

5. DISCUSSION

We have proposed a method that takes advantage of known substructure in the data when simultaneously clustering high-dimensional data with an unknown number of clusters, and selecting the best discriminating variables for those clusters. This method approximates stronger within design group covariance by defining the cluster member indicator vector **y** to assign all members of a design group to the same cluster. The approach is similar to the idea of forcing the elements of the original vector **y**, indexed over individuals rather than subgroups, into subsets where all entries in the same subset have the same value. In this approach the likelihood is adjusted to compute the proper probability that corresponds with the reduced variation. Since this method was developed for designed experimental data with specific treatment groups, we assume that the experimental subjects of each design group are homogeneous, and therefore there is no need to split the groups. Also, given the structure of designed experiments, breaking this basic experimental structure would have no interpretation with regard to the experiment, other than providing evidence that the experiment was poorly designed. Additionally, by jointly finding structure in the data and selecting variables, here genes, we answer the researchers questions of, first, whether the treatments affect the subjects differently and, second, which genes define those differences.

The proposed method has some advantages over the original method of Tadesse et al. (2005). The computation time, in particular, is shorter, as the dimension for some of the calculations are reduced to the number of subgroups instead of the number of individual samples. As a result, using information on substructure of the data implies less memory requirements and the opportunity to handle bigger datasets. To be specific, this method performed 1,000 iterations on the biological data (638 genes for 12 rats) in 3 minutes on an Intel dual core, dual processor PC running at 3.0GHz with 3 Gigabytes of RAM running SUSIE linux 10.

In our simulation study we successfully recovered the clusters and true variables in various simulated datasets. However one limitation of this study is that we simulated the discriminatory variables independently, and this can be unrealistic. The true correlation of gene expression values is quite complex, and modeling this correlation structure is an interesting research question in its own right, let alone simulating from it. The underlying covariate selection mechanism we use for the selection of the discriminating variables has been shown to be effective in analyzing correlated covariates in studies with genetic markers, which is simpler to model than gene expression correlation (Swartz et al. 2006; Swartz and Shete 2007).

This study has shown that we can effectively use this method for studies with 5 or more subgroups in the data. It has also shown robustness to different sized subgroups. When analyzing real data, our method has found several genes that agree with current biological research and literature and that provide biological validation of the cluster configuration suggested by the method. Overall, our method can provide biologists with both useful and manageable information for further experimental research.

ACKNOWLEDGMENTS

The authors thank Mahlet Tadesse for sharing the code to perform variable selection in clustering, and the reviewers, editor and associate editor for their constructive comments that improved the clarity of this paper. Vannucci is supported by NIH/NHGRI grant R01HG003319 and by NSF award DMS-0605001. Hong was supported by CERH pilot project number 5P30-ES09106 and Swartz by NCI training grants 5R25CA09030-04, R25CA57730, and career development grant 1K07CA123109-01A.

[Received April 2007. Revised December 2007.]

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300.

Brusco, M.J., and Cradit, J.D. (2001), "A Variable Selection Heuristic for *k*-means Clustering," *Psychometrika*, 66, 249–270.

Bustin, S.A., and Jenkins, P.J. (2001), “The Growth Hormone-Insulin-Like Growth Factor-i Axis and Colorectal Cancer,” *Trends in Molecular Medicine*, 7, 447–454.

Cooper, H.S., Murthy, S.N.S., Shah, R.S., and Sedergram, D. J. (1993), “Clinicopathic Study of Dextran Sulfae Sodium Experimental Murine Colitis,” *Laboratory Investigation*, 69, 238–249.

Dieleman, L. A., Elson, C.O., Tennyson, G.S., and Beagley, K (1996), “Kinetics of Cytokine Expression During Healing of Acute Colitis in Mice,” *American Journal of Physiology*, 271, G130–G136.

Do, K.A., Mueller, P., and Vannucci, M. (eds.) (2006), *Bayesian Inference for Gene Expression and Proteomics*. New York: Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001), “Empirical Bayes Analysis of a Microarray Experiment,” *Journal of the American Statistical Association*, 96, 1151–1160.

Friedman, J.H., and Meulman, J.J. (2004), “Clustering Objects on Subsets of Attributes,” *Journal of the Royal Statistical Society*, Series B, 66, 815–849.

Green, P.J. (1995), “Reversible-Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination,” *Biometrika*, 82, 711–732.

Hakam, A., Yeatman, T.J., Lu, L., Mora, L., Marcet, G., Nicosia, S.V., Karl, R.C., and Coppola, D. (1999), “Expression of Insulin-Like Growth Factor-1 Receptor in Human Colorectal Cancer,” *Human Pathology*, 30, 1128–1133.

Hanby, A.M., Chinery, R., Poulsom, R., Playford, R.J., and Pignatelli, M. (1996), “Downregulation of e-cadherin in the Reparative Epithelium of the Human Gastrointestinal Tract,” *American Journal of Pathology*, 148, 723–729.

Hardin, J.A., Wallace, J.E., Wong, J.F., O’Loughlin, E.V., Urbanski, D.J., and Gall, W.K. (2004), “Aquaporin Expression is Downregulated in a Murine Model of Colitis and in Patients with Ulcerative Colitis Crohn’s Disease and Infectious Colitis,” *Cell and Tissue Research*, 318, 313–323.

Heinzlmann, M., Lang, S.M., Neynaber, S., Reinshagen, M., Emmrich, J., Stratakis, D.F., Heldwein, W., Wiebecke, B., and Loeschke, K. (2002), “Screening for p53 and K-ras Mutations in Whole-Gut Lavage in Chronic Inflammatory Bowel Disease,” *European Journal of Gastroenterology & Hepatology*, 14, 1061–1066.

Hong, M-Y, Bancroft, L.K., Turner, N.D., Davidson, L.A., Murphy, M.E., Carroll, R.J., Chapkin, R.S., and Lupton, J.R. (2005), “Fish Oil Decreases Oxidative DNA Damage by Enhancing Apoptosis in Rat Colon,” *Nutrition and Cancer*, 52, 166–175.

Hu, Y., and Pham, C.T.N. (2005), “Diphptidyl Peptidase I Regulates the Development of Collagen-Induced Arthritis,” *Arthritis and Rhemuatism*, 52, 2553–2558.

Ibrahim, J.G., Chen, M.-H., and Gray, R.J. (2002), “Bayesian Models for Gene Expression with DNA Microarray Data,” *Journal of the American Statistical Association*, 97, 88–99.

Ierardi, E., Principi, M., Francavilla, R., Passaro, S., Noviello, R., Burattini, O., and Farancavilla, A. (2001), “Epithelial Proliferation and ras p21 Oncoprotein Expression in Rectal Mucosa of Patients with Ulcerrative Colitis,” *Digestive Diseases and Sciences*, 46, 1083–1087.

Karayiannakis, A.J., Syrigos, K.N., Efstathiou, J., Valizadeh, A., Noda, M., Playford, R.J., Kmiot, W., and Pignatelli, M. (1998), “Expression of Catenins and e-cadherin During Epithelial Restitution in Inflammatory Bowel Disease,” *Journal of Pathology*, 185, 413–418.

Kendziorski, C.M., Newton, M.A., Lan, H., and Gould, M.N. (2003), “On Parametric Empirical Bayes Methods for Comparing Multiple Groups using Replicated Gene Expression Profiles,” *Statistics in Medicine*, 22, 3899–3914.

Khandwala, H.M., McCutcheon, I.E., Flyvbjerg, A., and Friend, K.E. (2000), “The Effects of Insulin Like Growth Factors on Tumorigenesis and Neoplastic Growth,” *Endocrine Reviews*, 21, 215–244.

Kim, S., Tadesse, M.G., and Vannucci, M. (2006), “Variable Selection in Clustering via Dirichlet Process Mixture Models,” *Biometrika*, 93(4), 877–893.

Laforenza, U., Gastaldi, G., Grazioli, M., Cova, E., Tritto, S., Faelli, A., Calamita, G., and Ventura, U. (2005), “Expression and Immunolocalization of aquaporin-7 in Rat Gastrointestinal Tract,” *Biology of the Cell*, 97, 605–613.

Liu, Y.C. (2004), "Ubiquitin Ligases and the Immune Response," *Annual Reviews of Immunology*, 22, 81–127.

Moinova, H.R., Chen, W.D., Shen, L., Smiraglia, D., Olechnowicz, J., Ravi, L., Kasturi, L., Myeroff, L., Plass, C., Parsons, R., Minna, J., Wilson, J.K., Green, S.B., Issa, J.P., and Markowitz, S.D. (2002), "HLTF Gene Silencing in Human Colon Cancer," *Proceedings of the National Academy of Science*, 99, 4562–4567.

Newton, M.A., Kendziorski, C.M., Richmod, C.S., Blattner, F.R., and Tsui, K.W. (2001), "On Differential Variability of Expression Ratio: Improving Statistical Inference About Gene Expression Changes from Microarray Data," *Journal of Computational Biology*, 8, 37–52.

Okayasu, I, Hatekeyama, S., M., Yamada, Ohkusa, T., Inagaki, Y. and Nakaya, R. (1990), "A Novel Method in the Induction of Reliable Experimental Acute and Chronic Ulcerative Colitis in Mice," *Gastroenterology*, 98, 694–702.

Raftery, A.E., and Dean, N. (2006), "Variable Selection for Model-Based Clustering," *Journal of the American Statistical Assosciation*, 101, 168–178.

Richardson, S., and Green, P.J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components" (with discussion), *Journal of the Royal Statistical Society*, Series B, 59, 731–792.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: Wiley.

Sebastiani, P., Ramoni, M., and Kohane, I.S. (2003), "Bayesian Clustering of Gene Expression Dynamics," in *The Analysis of Gene Expression Data: Methods and Software*, eds G. Parmigiani, E.S. Garrett, R.A. Irizarry, and S.L. Zeger, New York: Springer, pp. 409–427.

Segain, J.P., de la Bletiere, D.R., Sauzeau, V., Bourreille, A., Hilaret, G., Cario-Toumaniatz, C., Pacaud, P., Galmiche, J.P. and Loirand, G. (2003), "Rho Kinase Blockade Prevents Inflammation via Nuclear Factor $\kappa\beta$ Inhibition: Evidence in Crohn's Disease and Experimental Colitis," *Gastroenterology*, 124, 1180–1187.

Seidler, U., Lenzen, H., Cinar, A., Tessema, T., Bleich, A., and Riederer, B. (2006), "Molecular Mechanisms of Disturbed Elecrolyte Transport in Intestinal Inflammation," *Annals of the New York Academy of Sciences*, 1072, 262–275.

Sha, N., Vannucci, M., Tadesse, M.G., Brown, P.J., Davies, N., Roberts, T., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004), "Bayesian Variable Selection in Multinomial Probit Models to Identify Molecular Signatures of Disease Stage," *Biometrics*, 60, 812–819.

Shimizu, T, Igarashi, J., Ohtuka, Y., Oguchi, S., Kaneko, K., and Yamashiro, Y (2001), "Effects of n-3 Polyunsaturated Fatty Acids and Vitamin E on Colonic Mucosal Leukotriene Generation, Lipid Peroxidation, and Microcirculation in Rats with Experimental Colitis," *Digestion*, 63, 49–54.

Stephens, M. (2000), "Dealing with Label Switching in Mixture Models," *Journal of the Royal Statistical Society*, Series B, 62, 795–809.

Swartz, M. D., and Shete, S. (2007), "The Null Distribution of Stochastic Search Gene Suggestion: A Bayesian Approach to Gene Mapping," *BMC Proceedings*, Suppl 1, S113–S118.

Swartz, M.D., Kimmel, M., Mueller, P., and Amos, C.I. (2006), "Stochastic Search Gene Suggestion: A Bayesian Hierarchical Model for Gene Mapping," *Biometrics*, 62, 495–503.

Syrigos, K.N., and Karayiannakiks, A.J. (2006), "Adhesion Molecules as Targets for the Treatment of Neoplastic Diseases," *Current Pharmaceutical Designs*, 12, 2849–2861.

Tadesse, M.G., Sha, N., and Vannucci, M. (2005), "Bayesian Variable Selection in Clustering High Dimensional Data," *Journal of the American Statistical Association*, 100, 602–617.