# Bayesian Variable Selection in Clustering High-Dimensional Data

Mahlet G. TADESSE, Naijun SHA, and Marina VANNUCCI

Over the last decade, technological advances have generated an explosion of data with substantially smaller sample size relative to the number of covariates ($p \gg n$). A common goal in the analysis of such data involves uncovering the group structure of the observations and identifying the discriminating variables. In this article we propose a methodology for addressing these problems simultaneously. Given a set of variables, we formulate the clustering problem in terms of a multivariate normal mixture model with an unknown number of components and use the reversible-jump Markov chain Monte Carlo technique to define a sampler that moves between different dimensional spaces. We handle the problem of selecting a few predictors among the prohibitively vast number of variable subsets by introducing a binary exclusion/inclusion latent vector, which gets updated via stochastic search techniques. We specify conjugate priors and exploit the conjugacy by integrating out some of the parameters. We describe strategies for posterior inference and explore the performance of the methodology with simulated and real datasets.

KEY WORDS: Bayesian variable selection; Bayesian clustering; Label switching; Reversible-jump Markov chain Monte Carlo.

## 1. INTRODUCTION

Over the last decade, technological advances have generated an explosion of data for which the number of covariates is considerably larger than the sample size. These data pose a challenge to standard statistical methods and have revived a strong interest in clustering algorithms. The goals are to uncover the group structure of the observations and to determine the discriminating variables. The analysis of DNA microarray datasets is a typical high-dimensional example, where variable selection and outcome prediction have become a major focus of research. The technology provides an automated method for simultaneously quantifying thousands of genes, but its high cost constrains researchers to a few experimental units. The discovery of different types of tissues or subtypes of a disease and the identification of genes that best distinguish them is believed to provide a better understanding of the underlying biological processes. This in turn could lead to better treatment choice for patients in different risk groups, and the selected genes can serve as biomarkers to improve diagnosis and therapeutic intervention.

When dealing with high-dimensional datasets, the cluster structure of the observations is often confined to a small subset of variables. As pointed out by several authors (e.g., Fowlkes, Gnanadesikan, and Kettering 1988; Milligan 1989; Gnanadesikan, Kettering, and Tao 1995; Brusco and Cradit 2001), the inclusion of unnecessary covariates could complicate or even mask the recovery of the clusters. Common approaches to mitigating the effect of noisy variables or identifying those that define true cluster structure involve differentially weighting the covariates or selecting the discriminating ones. Gnanadesikan et al. (1995) showed that variable weighting schemes were often outperformed by variable selection procedures. In addition to improving the prediction of cluster membership, the variable selection task reduces the measurement and storage requirements for future samples, thereby providing more cost-effective predictors.

In this article we focus on the analysis of high-dimensional datasets characterized by a sample size, $n$, that is substantially smaller than the number of covariates, $p$. We propose a Bayesian approach for simultaneously selecting discriminating variables and uncovering the cluster structure of the observations. The article is organized as follows. In Section 2 we give a brief review of existing procedures and discuss their shortcomings. In Section 3 we describe our model specification, which makes use of model-based clustering and introduces latent indicators to identify discriminating variables. In Section 4 we present the prior assumptions and the resulting full conditionals. We also provide guidelines for choosing the hyperparameters. In Section 5 we discuss in detail the Markov chain Monte Carlo (MCMC) updates for each of the model parameters, and in Section 6 we address the issue of label switching and describe the inference mechanism. In Section 7 we assess the performance of our methodology using various datasets. We conclude the article with a brief discussion in Section 8.

## 2. REVIEW OF EXISTING METHODS

The most widely used approaches separate the variable selection/weighting and clustering tasks. One common variable selection approach involves fitting univariate models on each covariate and selecting a small subset that passes some threshold for significance (McLachlan, Bean, and Peel 2002). Another standard approach uses dimension-reduction techniques, such as principal component analysis, and focuses on the leading components (Ghosh and Chinnaiyan 2002). The reduced data are then clustered using the procedure of choice. These selection steps are suboptimal. The first approach does not assess the joint effect of multiple variables and could throw away potentially valuable covariates, which are not predictive individually but may provide significant improvement in conjunction with others. The second approach results in linear combinations and does not allow evaluation of the original

Mahlet G. Tadesse is Assistant Professor, Department of Biostatistics, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *mtadesse@cceb.upenn.edu*). Naijun Sha is Assistant Professor, Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968 (E-mail: *nsha@utep.edu*). Marina Vannucci is Associate Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *mvannucci@stat.tamu.edu*). Tadesse's research is supported by grant CA90301 from the National Cancer Institute, Sha's work is supported in part by BBRC/RCMI National Institute of Health grant 2G12RR08124 and a University Research Institute (UTEP) grant, and Vannucci's research is supported by National Science Foundation CAREER award number DMS-00-93208. The authors thank the associate editor and two referees for their thorough reading of the manuscript and their input, which have improved the article considerably.


Journal of the American Statistical Association
June 2005, Vol. 100, No. 470, Theory and Methods
DOI 10.1198/016214504000001565

variables. In addition, it has been shown that the leading components do not necessarily contain the most information about the cluster structure (Chang 1983).

The literature includes several procedures that combine the variable selection and clustering tasks. Fowlkes et al. (1988) used a forward selection approach in the context of complete linkage hierarchical clustering. The variables are added using information of the between-cluster and total sum of squares, and their significance is judged based on graphical information. The authors characterized this as an "informal" assessment. Brusco and Cradit (2001) proposed a variable selection heuristic for $k$-means clustering using a similar forward selection procedure. Their approach uses the adjusted Rand index to measure cluster recovery. Recently, Friedman, and Meulman (2003) proposed a hierarchical clustering procedure that uncovers cluster structure on separate subsets of variables. The algorithm does not explicitly select variables, but rather assigns them different weights, which can be used to extract relevant covariates. This approach is somewhat different from ours in that it detects subgroups of observations that preferentially cluster on different subsets of variables, rather than clustering on the same subset of variables. These methods all work in conjunction with $k$-means or hierarchical clustering algorithms, which have several limitations. One major shortcoming is the lack of a statistical criterion for assessing the number of clusters. These procedures also do not provide a measure of the uncertainty associated with the sample allocations. In addition, with respect to variable selection, the existing approaches use greedy deterministic search procedures, which can be stuck at local minima. They also have the limitation of presuming the existence of a single "best" subset of clustering variables. In practice, however, there may be several equally good subsets that define the true cluster structure.

The Bayesian paradigm offers a coherent framework for addressing the problems of variable selection and clustering simultaneously. Recent developments in MCMC techniques (Gilks, Richardson, and Spiegelhalter 1996) have led to substantial research in both areas. Bayesian variable selection methods so far have been developed mainly in the context of regression analysis. (See George and McCulloch 1997 for a review on prior specifications and MCMC implementations, and Brown, Vannucci, and Fearn 1998b for extensions to the case of multivariate responses.) The Bayesian clustering approach uses a mixture of probability distributions, each distribution representing a different cluster. Diebolt and Robert (1994) presented a comprehensive treatment of MCMC strategies when the number of mixture components is known. For the general case of an unknown number of components, Richardson and Green (1997) successfully applied the reversible-jump MCMC (RJMCMC) technique, but considered only univariate data. Stephens (2000a) proposed an approach based on continuous-time Markov birth–death processes and applied it to a bivariate setting.

We propose a method that simultaneously selects the discriminating variables and clusters the samples into $G$ groups, where $G$ is unknown. The computational burden of searching through all possible $2^p$ variable subsets is handled through the introduction of a latent $p$-vector with binary entries. We use a stochastic search method to explore the space of possible values for this latent vector. The clustering problem is formulated in terms of a multivariate mixture model with an unknown number of components. We work with a marginalized likelihood where some of the model parameters are integrated out and adapt the RJMCMC technique of Richardson and Green (1997) to the multivariate setting. Our inferential procedure provides selection of discriminating variables, estimates for the number of clusters and the sample allocations, and class prediction for future observations.

## 3. MODEL SPECIFICATION

### 3.1 Clustering via Mixture Models

The goal of cluster analysis is to partition a collection of observations into homogeneous groups. Model-based clustering has received great attention recently and has provided promising results in various applications. In this approach the data are viewed as coming from a mixture of distributions, each distribution representing a different cluster. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be independent $p$-dimensional observations from $G$ populations. The problem of clustering the $n$ samples can be formulated in terms of a mixture of $G$ underlying probability distributions

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

where $f(\mathbf{x}_i|\boldsymbol{\theta}_k)$ is the density of an observation $\mathbf{x}_i$ from the $k$th component and $\mathbf{w} = (w_1, \ldots, w_G)^T$ are the component weights ($w_k \geq 0$, $\sum_{k=1}^{G} w_k = 1$).

To identify the cluster from which each observation is drawn, latent variables $\mathbf{y} = (y_1, \ldots, y_n)^T$ are introduced, where $y_i = k$ if the $i$th observation comes from component $k$. The $y_i$'s are assumed to be independently and identically distributed with probability mass function $p(y_i = k) = w_k$. We consider the case where $f(\mathbf{x}_i|\boldsymbol{\theta}_k)$ is a multivariate normal density with parameters $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Thus for sample $i$, we have

$$\mathbf{x}_i|y_i = k, \mathbf{w}, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2}$$

### 3.2 Identifying Discriminating Variables

In high-dimensional data, it is often the case that a large number of variables provide very little, if any, information about the group structure of the observations. Their inclusion could be detrimental, because it might obscure the recovery of the true cluster structure. To identify the relevant variables, we introduce a latent $p$-vector, $\boldsymbol{\gamma}$, with binary entries. $\gamma_j$ takes value 1 if the $j$th variable defines a mixture distribution for the data and 0 otherwise. We use $(\gamma)$ and $(\gamma^c)$ to index the discriminating variables and those that favor a single multivariate normal density. Notice that this approach is different from what is commonly done in Bayesian variable selection for regression models, where the latent indicator is used to induce mixture priors on the regression coefficients. The likelihood function is then given by

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

$$\times \prod_{k=1}^{G} (2\pi)^{-p_\gamma n_k/2} \left|\boldsymbol{\Sigma}_{k(\gamma)}\right|^{-n_k/2} w_k^{n_k} \times \exp\left\{-\frac{1}{2}\sum_{x_i \in C_k} \left(\mathbf{x}_{i(\gamma)} - \boldsymbol{\mu}_{k(\gamma)}\right)^T \boldsymbol{\Sigma}_{k(\gamma)}^{-1} \left(\mathbf{x}_{i(\gamma)} - \boldsymbol{\mu}_{k(\gamma)}\right)\right\}, \tag{3}$$

where $(\boldsymbol{\eta}_{(\gamma^c)}, \boldsymbol{\Omega}_{(\gamma^c)})$ denote the mean and covariance parameters for the nondiscriminating variables, $(\boldsymbol{\mu}_{k(\gamma)}, \boldsymbol{\Sigma}_{k(\gamma)})$ are the mean and covariance parameters of cluster $k$, $C_k = \{x_i | y_i = k\}$ with cardinality $n_k$, and $p_\gamma = \sum_{j=1}^{p} \gamma_j$.

## 4. PRIOR SETTING AND FULL CONDITIONALS

### 4.1 Prior Formulation and Choice of Hyperparameters

For the variable selection indicator, $\boldsymbol{\gamma}$, we assume that its elements $\gamma_j$ are independent Bernoulli random variables,

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{p} \varphi^{\gamma_j} (1-\varphi)^{1-\gamma_j}. \tag{4}$$

The number of covariates included in the model, $p_\gamma$, thus follows a binomial distribution, and $\varphi$ can be elicited as the proportion of variables expected a priori to discriminate the different groups, $\varphi = p_{\text{prior}}/p$. This prior assumption can be relaxed by formulating a beta$(a, b)$ hyperprior on $\varphi$, which leads to a beta-binomial prior for $p_\gamma$ with expectation $p\frac{a}{a+b}$. A vague prior can be elicited by setting $a + b = 2$, which leads to setting $a = 2p_{\text{prior}}/p$ and $b = 2 - a$.

Natural priors for the number of components, $G$, are a truncated Poisson,

$$P(G = g) = \frac{e^{-\lambda}\lambda^g/g!}{1 - (e^{-\lambda}(1+\lambda) + \sum_{j=G_{\max}+1}^{\infty} (e^{-\lambda}\lambda^j)/j!)}, \quad g = 2, \ldots, G_{\max}, \tag{5}$$

or a discrete uniform on $[2, \ldots, G_{\max}]$, $P(G = g) = \frac{1}{G_{\max}-1}$, where $G_{\max}$ is chosen arbitrarily large. For the vector of component weights, $\mathbf{w}$, we specify a symmetric Dirichlet prior, $\mathbf{w}|G \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$.

Updating the mean and covariance parameters is somewhat intricate. When deleting or creating new components, the corresponding appropriate changes must be made for the component parameters. However, it is not clear whether adequate proposals can be constructed for the reversible jump in the multivariate setting. Even if this difficulty can be overcome, as $\boldsymbol{\gamma}$ gets updated, the dimensions of $\boldsymbol{\mu}_{k(\gamma)}$, $\boldsymbol{\Sigma}_{k(\gamma)}$, $\boldsymbol{\eta}_{(\gamma^c)}$, and $\boldsymbol{\Omega}_{(\gamma^c)}$ change, requiring another sampler that moves between different dimensional spaces. The algorithm is much more efficient if we can integrate out these parameters. This also helps substantially accelerate model fitting, because the set of parameters that need to be updated would then consist only of $(\boldsymbol{\gamma}, \mathbf{w}, \mathbf{y}, G)$. The integration can be facilitated by taking conjugate priors

$$\begin{aligned} \boldsymbol{\mu}_{k(\gamma)} | \boldsymbol{\Sigma}_{k(\gamma)}, G &\sim \mathcal{N}\left(\boldsymbol{\mu}_{0(\gamma)}, h_1 \boldsymbol{\Sigma}_{k(\gamma)}\right), \\ \boldsymbol{\eta}_{(\gamma^c)} | \boldsymbol{\Omega}_{(\gamma^c)} &\sim \mathcal{N}\left(\boldsymbol{\mu}_{0(\gamma^c)}, h_0 \boldsymbol{\Omega}_{(\gamma^c)}\right), \\ \boldsymbol{\Sigma}_{k(\gamma)} | G &\sim \mathcal{IW}\left(\delta; \mathbf{Q}_{1(\gamma)}\right), \\ \boldsymbol{\Omega}_{(\gamma^c)} &\sim \mathcal{IW}\left(\delta; \mathbf{Q}_{0(\gamma^c)}\right), \end{aligned} \tag{6}$$

where $\mathcal{IW}(\delta; \mathbf{Q}_{1(\gamma)})$ indicates the inverse-Wishart distribution with dimension $p_\gamma$, shape parameter $\delta = n - p_\gamma + 1$, $n$ degrees of freedom, and mean $\mathbf{Q}_{1(\gamma)}/(\delta - 2)$ (Brown 1993). Weak prior information is specified by taking $\delta$ small, which we set to 3, the smallest integer such that the expectations of the covariance matrices are defined. We take $\mathbf{Q}_1 = 1/\kappa_1 \mathbf{I}_{p\times p}$ and $\mathbf{Q}_0 = 1/\kappa_0 \mathbf{I}_{p\times p}$. Some care in the choice of $\kappa_1$ and $\kappa_0$ is needed. These hyperparameters need to be in the range of variability of the data. At the same time, we do not want the prior information to overwhelm the data and drive the inference. We experimented with different values and found that values commensurate with the order of magnitude of the eigenvalues of $\text{cov}(\mathbf{X})$, where $\mathbf{X}$ is the full data matrix, lead to reasonable results. We suggest setting $\kappa_1$ between 1 and 10% of the upper decile of the $n - 1$ non-zero eigenvalues and $\kappa_0$ proportional to their lower decile.

For the mean parameters, we take the priors to be fairly flat over the region where the data are defined. Each element of $\boldsymbol{\mu}_0$ is set to the corresponding covariate interval midpoint, and $h_0$ and $h_1$ are chosen arbitrarily large. This prevents the specification of priors that do not overlap with the likelihood and allows for mixtures with widely different component means. We found that values of $h_0$ and $h_1$ between 10 and 1,000 performed reasonably well.

### 4.2 Full Conditionals

As we stated earlier, the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\eta}$, and $\boldsymbol{\Omega}$ are integrated out, and we need to sample only from the joint posterior of $(G, \mathbf{y}, \boldsymbol{\gamma}, \mathbf{w})$. We have

$$\begin{aligned} &f(\mathbf{X}, \mathbf{y} | G, \mathbf{w}, \boldsymbol{\gamma}) \\ &\quad = \int f(\mathbf{X}, \mathbf{y} | \boldsymbol{\gamma}, G, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta}, \boldsymbol{\Omega}) f(\boldsymbol{\mu} | G, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) f(\boldsymbol{\Sigma} | G, \boldsymbol{\gamma}) \\ &\qquad \times f(\boldsymbol{\eta} | \boldsymbol{\Omega}, \boldsymbol{\gamma}) f(\boldsymbol{\Omega} | \boldsymbol{\gamma})\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}\, d\boldsymbol{\eta}\, d\boldsymbol{\Omega}. \end{aligned} \tag{7}$$

For completeness, we provide the full conditionals under the assumption of equal and unequal covariances across clusters. However, in practice, unless there is prior knowledge that favors the former, we recommend working with heterogeneous covariances. After some algebra (see the Appendix), we get the following:

1. Under the assumption of homogeneous covariance,

$$\begin{aligned} &f(\mathbf{X}, \mathbf{y} | G, \mathbf{w}, \boldsymbol{\gamma}) \\ &\quad = \pi^{-np/2} K_{(\gamma)} \cdot \left|\mathbf{Q}_{1(\gamma)}\right|^{(\delta + p_\gamma - 1)/2} \\ &\qquad \times \left|\mathbf{Q}_{1(\gamma)} + \sum_{k=1}^{G} \mathbf{S}_{k(\gamma)}\right|^{-(n+\delta+p_\gamma-1)/2} \\ &\qquad \times H_{(\gamma^c)} \cdot \left|\mathbf{Q}_{0(\gamma^c)}\right|^{(\delta+p-p_\gamma-1)/2} \\ &\qquad \times \left|\mathbf{Q}_{0(\gamma^c)} + \mathbf{S}_{0(\gamma^c)}\right|^{-(n+\delta+p-p_\gamma-1)/2}, \end{aligned} \tag{8}$$

$$K_{(\gamma)} = \left[\prod_{k=1}^{G} w_k^{n_k} (h_1 n_k + 1)^{-p_\gamma/2}\right] \times \prod_{j=1}^{p_\gamma} \frac{\Gamma(\frac{1}{2}(n + \delta + p_\gamma - j))}{\Gamma(\frac{1}{2}(\delta + p_\gamma - j))},$$

$$H_{(\gamma^c)} = (h_0 n + 1)^{-(p-p_\gamma)/2} \times \prod_{j=1}^{p-p_\gamma} \frac{\Gamma(\frac{1}{2}(n+\delta+p-p_\gamma-j))}{\Gamma(\frac{1}{2}(\delta+p-p_\gamma-j))},$$

$$\mathbf{S}_{k(\gamma)} = \sum_{\mathbf{x}_{i(\gamma)} \in C_k} \big(\mathbf{x}_{i(\gamma)} - \bar{\mathbf{x}}_{k(\gamma)}\big)\big(\mathbf{x}_{i(\gamma)} - \bar{\mathbf{x}}_{k(\gamma)}\big)^T + \frac{n_k}{h_1 n_k + 1}\big(\boldsymbol{\mu}_{0(\gamma)} - \bar{\mathbf{x}}_{k(\gamma)}\big)\big(\boldsymbol{\mu}_{0(\gamma)} - \bar{\mathbf{x}}_{k(\gamma)}\big)^T,$$

$$\mathbf{S}_{0(\gamma^c)} = \sum_{i=1}^{n} \big(\mathbf{x}_{i(\gamma^c)} - \bar{\mathbf{x}}_{(\gamma^c)}\big)\big(\mathbf{x}_{i(\gamma^c)} - \bar{\mathbf{x}}_{(\gamma^c)}\big)^T + \frac{n}{h_0 n + 1}\big(\boldsymbol{\mu}_{0(\gamma^c)} - \bar{\mathbf{x}}_{(\gamma^c)}\big)\big(\boldsymbol{\mu}_{0(\gamma^c)} - \bar{\mathbf{x}}_{(\gamma^c)}\big)^T,$$

$\bar{\mathbf{x}}_{k(\gamma)}$ is the sample mean of cluster $k$, and $\bar{\mathbf{x}}_{(\gamma^c)}$ corresponds to the sample means of the nondiscriminating variables. As $h_1 \to \infty$, $\mathbf{S}_{k(\gamma)}$ reduces to the within-group sum of squares for the $k$th cluster, and $\sum_{k=1}^{G} \mathbf{S}_{k(\gamma)}$ becomes the total within-group sum of squares.

2. Assuming heterogeneous covariances across clusters,

$$\begin{aligned} f(\mathbf{X}, \mathbf{y} | G, \mathbf{w}, \boldsymbol{\gamma}) &= \pi^{-np/2} \prod_{k=1}^{G} \Big\{ K_{k(\gamma)} \cdot \big|\mathbf{Q}_{1(\gamma)}\big|^{(\delta+p_\gamma-1)/2} \\ &\quad \times \big|\mathbf{Q}_{1(\gamma)} + \mathbf{S}_{k(\gamma)}\big|^{-(n_k+\delta+p_\gamma-1)/2} \Big\} \\ &\quad \times H_{(\gamma^c)} \cdot \big|\mathbf{Q}_{0(\gamma^c)}\big|^{(\delta+p-p_\gamma-1)/2} \\ &\quad \times \big|\mathbf{Q}_{0(\gamma^c)} + \mathbf{S}_{0(\gamma^c)}\big|^{-(n+\delta+p-p_\gamma-1)/2}, \end{aligned} \tag{9}$$

where $K_{k(\gamma)} = w_k^{n_k} (h_1 n_k + 1)^{-p_\gamma/2} \prod_{j=1}^{p_\gamma} \frac{\Gamma(\frac{1}{2}(n_k+\delta+p_\gamma-j))}{\Gamma(\frac{1}{2}(\delta+p_\gamma-j))}$.

The full conditionals for the model parameters are then given by

$$f(\mathbf{y}|G, \mathbf{w}, \boldsymbol{\gamma}, \mathbf{X}) \propto f(\mathbf{X}, \mathbf{y}|G, \mathbf{w}, \boldsymbol{\gamma}), \tag{10}$$

$$f(\boldsymbol{\gamma}|G, \mathbf{w}, \mathbf{y}, \mathbf{X}) \propto f(\mathbf{X}, \mathbf{y}|G, \mathbf{w}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}|G), \tag{11}$$

and

$$\mathbf{w}|G, \boldsymbol{\gamma}, \mathbf{y}, \mathbf{X} \sim \text{Dirichlet}(\alpha + n_1, \ldots, \alpha + n_G). \tag{12}$$

## 5. MODEL FITTING

Posterior samples for the parameters of interest are obtained using Metropolis moves and RJMCMC embedded within a Gibbs sampler. Our MCMC procedure comprises of the following five steps:

1. Update $\boldsymbol{\gamma}$ from its full conditional in (11).
2. Update $\mathbf{w}$ from its full conditional in (12).
3. Update $\mathbf{y}$ from its full conditional in (10).
4. Split one mixture component into two, or merge two into one.
5. Birth or death of an empty component.

The birth/death moves specifically deal with creating/deleting empty components and do not involve reallocation of the observations. We could have extended these to include nonempty components and removed the split/merge moves. However, as described by Richardson and Green (1997), including both steps provides more efficient moves and improves convergence.

We would like to point out that as we iterate through these steps, the cluster structure evolves with the choice of variables. This makes the problem of variable selection in the context of clustering much more complicated than classification or regression analysis. In classification, the group structure of the observations is prespecified in the training data and guides the selection. Here, however, the analysis is fully data-based. In addition, in the context of clustering, including unnecessary variables can have severe implications, because it may obscure the true grouping of the observations.

### 5.1 Variable Selection Update

In step 1 we update $\boldsymbol{\gamma}$ using a Metropolis search, as suggested for model selection by Madigan and York (1995) and applied in regression by Brown et al. (1998a), among others, and recently by Sha et al. (2004) in multinomial probit models for classification. In this approach a new candidate, $\boldsymbol{\gamma}^{\text{new}}$, is generated by randomly choosing one of the following two transition moves:

1. Add/delete: Randomly choose one of the $p$ indices in $\boldsymbol{\gamma}^{\text{old}}$, and change its value from 0 to 1, or from 1 to 0, to become $\boldsymbol{\gamma}^{\text{new}}$.
2. Swap: Choose independently and at random a 0 and a 1 in $\boldsymbol{\gamma}^{\text{old}}$, and switch their values to get $\boldsymbol{\gamma}^{\text{new}}$.

The new candidate $\boldsymbol{\gamma}^{\text{new}}$ is accepted with probability $\min\{1, \frac{f(\boldsymbol{\gamma}^{\text{new}}|\mathbf{X},\mathbf{y},\mathbf{w},G)}{f(\boldsymbol{\gamma}^{\text{old}}|\mathbf{X},\mathbf{y},\mathbf{w},G)}\}$.

### 5.2 Update of Weights and Cluster Allocation

In step 2 we use a Gibbs move to sample $\mathbf{w}$ from its full conditional in (12). This can be done by drawing independent gamma random variables with common scale and shape parameters $\alpha + n_1, \ldots, \alpha + n_G$, then scaling them to sum to 1.

In step 3 we update the allocation vector $\mathbf{y}$ one element at a time via a sub-Gibbs sampling strategy. The full conditional probability that the $i$th observation is in the $k$th cluster is given by

$$f\big(y_i = k \big| \mathbf{X}, \mathbf{y}_{(-i)}, \boldsymbol{\gamma}, \mathbf{w}, G\big) \propto f\big(\mathbf{X}, y_i = k, \mathbf{y}_{(-i)} \big| G, \mathbf{w}, \boldsymbol{\gamma}\big), \tag{13}$$

where $\mathbf{y}_{(-i)}$ is the cluster assignment for all observations except the $i$th one.

### 5.3 Reversible-Jump MCMC

The last two steps, split/merge and birth/death moves, cause the creation or deletion of new components, and consequently require a sampler that jumps between different dimensional spaces. A popular method for defining such a sampler is RJMCMC (Green 1995; Richardson and Green 1997), which extends the Metropolis–Hastings algorithm to general state spaces. Suppose that a move type $m$ from a countable family of move types is proposed from $\psi$ to a higher-dimensional space $\psi'$. This will very often be implemented by drawing a vector of continuous random variables, $u$, independent of $\psi$, and setting $\psi'$ to be a deterministic and invertible function of $\psi$ and $u$. The reverse of this move (from $\psi'$ to $\psi$) can be accomplished by using the inverse transformation, so that in this

direction the proposal is deterministic. The move is then accepted with probability

$$\min\left\{1, \frac{f(\psi'|X)r_m(\psi')}{f(\psi|X)r_m(\psi)q(u)}\left|\frac{\partial\psi'}{\partial(\psi,u)}\right|\right\}, \tag{14}$$

where $r_m(\psi)$ is the probability of choosing move type $m$ when in state $\psi$ and $q(u)$ is the density function of $u$. The final term in the foregoing ratio is a Jacobian arising from the change of variable from $(\psi, u)$ to $\psi'$.

*5.3.1 Split and Merge Moves.* In step 4 we make a random choice between attempting to split a nonempty cluster or combine two clusters, with probabilities $b_k = .5$ and $d_k = 1 - b_k$ ($k = 2, \ldots, G_{\max} - 1$, and $b_{G_{\max}} = 0$).

The split move begins by randomly selecting a nonempty cluster $l$ and dividing it into two components, $l_1$ and $l_2$, with weights $w_{l1} = w_l u$ and $w_{l2} = w_l(1-u)$, where $u$ is a beta(2, 2) random variable. The parameters $\boldsymbol{\psi} = (G, \mathbf{w}_{\text{old}}, \boldsymbol{\gamma}, \mathbf{y}_{\text{old}})$ are updated to $\boldsymbol{\psi}' = (G+1, \mathbf{w}_{\text{new}}, \boldsymbol{\gamma}, \mathbf{y}_{\text{new}})$. The acceptance probability for this move is given by $\min(1, A)$, where

$$\begin{aligned} A &= \frac{f(G+1, \mathbf{w}_{\text{new}}, \boldsymbol{\gamma}, \mathbf{y}_{\text{new}}|\mathbf{X})}{f(G, \mathbf{w}_{\text{old}}, \boldsymbol{\gamma}, \mathbf{y}_{\text{old}}|\mathbf{X})} \times \frac{q(\boldsymbol{\psi}|\boldsymbol{\psi}')}{q(\boldsymbol{\psi}'|\boldsymbol{\psi}) \times f(u)} \\ &\quad \times \left|\frac{\partial(w_{l1}, w_{l2})}{\partial(w_l, u)}\right| \\ &= \frac{f(\mathbf{X}, \mathbf{y}_{\text{new}}|\mathbf{w}_{\text{new}}, \boldsymbol{\gamma}, G+1) \times f(\mathbf{w}_{\text{new}}|G+1) \times f(G+1)}{f(\mathbf{X}, \mathbf{y}_{\text{old}}|\mathbf{w}_{\text{old}}, \boldsymbol{\gamma}, G) \times f(\mathbf{w}_{\text{old}}|G) \times f(G)} \\ &\quad \times \frac{d_{G+1} \times P_{\text{chos}}}{b_G \times G_1^{-1} \times P_{\text{alloc}} \times f(u)} \times w_l, \end{aligned} \tag{15}$$

$$\frac{f(\mathbf{w}_{\text{new}}|G+1)}{f(\mathbf{w}_{\text{old}}|G)} = \frac{w_{l1}^{\alpha-1} w_{l2}^{\alpha-1}}{w_l^{\alpha-1} B(\alpha, G\alpha)}$$

and

$$\frac{f(G+1)}{f(G)} = \begin{cases} 1 & \text{if } G \sim \text{uniform}\,[2, \ldots, G_{\max}] \\ \frac{\lambda}{G+1} & \text{if } G \sim \text{truncated Poisson}\,(\lambda). \end{cases}$$

Here $G_1$ is the number of nonempty components before the split, and $P_{\text{alloc}}$ is the probability that the particular allocation is made and is set to $u^{n_{l1}}(1-u)^{n_{l2}}$, where $n_{l1}$ and $n_{l2}$ are the number of observations assigned to $l_1$ and $l_2$. Note that Richardson and Green (1997) calculated $P_{\text{alloc}}$ using the full conditional of the allocation variables, $P(y_i = k|\text{rest})$. In our case, because the component parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are integrated out, the allocation variables are no longer independent, and calculating these full conditionals is computationally prohibitive. The random allocation approach that we took is substantially faster even with the longer chains that need to be run to compensate for the low acceptance rate of proposed moves. $P_{\text{chos}}$ is the probability of selecting two components for the reverse move. Obviously, we want to combine clusters that are closest to one another. However, choosing an optimal similarity metric in the multivariate setting is not straightforward. We define closeness in terms of the Euclidean distance between cluster means and consider three possible ways of splitting a cluster:

a. Split results into two mutually adjacent components, $\Rightarrow P_{\text{chos}} = \frac{G_1^*}{G^*} \times \frac{2}{G_1^*}$.

b. Split results into components that are not mutually adjacent; that is, $l_1$ has $l_2$ as its closest component, but there is another cluster that is closer to $l_2$ than $l_1 \Rightarrow P_{\text{chos}} = \frac{G_1^*}{G^*} \times \frac{1}{G_1^*}$.

c. One of the resulting components is empty, $\Rightarrow P_{\text{chos}} = \frac{G_0^*}{G^*} \times \frac{1}{G_0^*} \times \frac{1}{G_1^*}$.

Here $G^*$ is the total number of clusters after the split. $G_0^*$ and $G_1^*$ are the number of empty and nonempty components after the split. Thus splits that result in mutually close components are assigned a larger probability, and those that yield empty clusters have lower probability.

The merge move begins by randomly choosing two components, $l_1$ and $l_2$, to be combined into a single cluster, $l$. The updated parameters are $\boldsymbol{\psi} = (G, \mathbf{w}_{\text{old}}, \boldsymbol{\gamma}, \mathbf{y}_{\text{old}}) \to \boldsymbol{\psi}' = (G-1, \mathbf{w}_{\text{new}}, \boldsymbol{\gamma}, \mathbf{y}_{\text{new}})$, where $w_l = w_{l1} + w_{l2}$. The acceptance probability for this move is given by $\min(1, A)$, where

$$\begin{aligned} A &= \frac{f(\mathbf{X}, \mathbf{y}_{\text{new}}|\boldsymbol{\gamma}, \mathbf{w}_{\text{new}}, G-1) \times f(\mathbf{w}_{\text{new}}|G-1) \times f(G-1)}{f(\mathbf{X}, \mathbf{y}_{\text{old}}|\boldsymbol{\gamma}, \mathbf{w}_{\text{old}}, G) \times f(\mathbf{w}_{\text{old}}|G) \times f(G)} \\ &\quad \times \frac{b_{G-1} \times P_{\text{alloc}} \times (G_1 - 1)^{-1} \times f(u)}{d_G \times P_{\text{chos}}} \\ &\quad \times (w_{l1} + w_{l2})^{-1}, \end{aligned} \tag{16}$$

where $\frac{f(\mathbf{w}_{\text{new}}|G-1)}{f(\mathbf{w}_{\text{old}}|G)} = \frac{w_l^{\alpha-1} B(\alpha,(G-1)\alpha)}{w_{l1}^{\alpha-1} w_{l2}^{\alpha-1}}$ and $\frac{f(G-1)}{f(G)}$ equals 1 or $\frac{G}{\lambda}$ for the Uniform and Poisson priors on $G$. $G_1$ is the number of nonempty clusters before the combining move, $f(u)$ is the beta(2, 2) density for $u = \frac{w_{l1}}{w_l}$, and $P_{\text{alloc}}$ and $P_{\text{chos}}$ are defined similarly to the split move, but now $G_0^*$ and $G_1^*$ in $P_{\text{chos}}$ correspond to the number of empty and nonempty clusters after the merge.

*5.3.2 Birth and Death Moves.* We first make a random choice between a birth move and a death move using the same probabilities, $b_k$ and $d_k = 1 - b_k$, as earlier. For a birth move, the updated parameters are $\boldsymbol{\psi} = (G, \mathbf{w}_{\text{old}}, \boldsymbol{\gamma}, \mathbf{y}) \to \boldsymbol{\psi}' = (G+1, \mathbf{w}_{\text{new}}, \boldsymbol{\gamma}, \mathbf{y})$. The weight for the proposed new component is drawn using $w^* \sim \text{beta}(1, G)$, and the existing weights are rescaled to $w_k' = w_k(1-w^*)$, $k = 1, \ldots, G$ (where $k$ indexes the component labels), so that all of the weights sum to 1. Let $G_0$ be the number of empty components before the birth. The acceptance probability for a birth is $\min(1, A)$, where

$$\begin{aligned} A &= \frac{f(\mathbf{X}, \mathbf{y}|\mathbf{w}_{\text{new}}, \boldsymbol{\gamma}, G+1) \times f(\mathbf{w}_{\text{new}}|G+1) \times f(G+1)}{f(\mathbf{X}, \mathbf{y}|\mathbf{w}_{\text{old}}, \boldsymbol{\gamma}, G) \times f(\mathbf{w}_{\text{old}}|G) \times f(G)} \\ &\quad \times \frac{d_{G+1} \times (G_0 + 1)^{-1}}{b_G \times f(w^*)} \times (1-w^*)^{G-1}, \end{aligned} \tag{17}$$

$\frac{f(\mathbf{w}_{\text{new}}|G+1)}{f(\mathbf{w}_{\text{old}}|G)} = \frac{w^{*\alpha-1}(1-w^*)^{G(\alpha-1)}}{B(\alpha, G\alpha)}$, and $(1-w^*)^{G-1}$ is the Jacobian of the transformation $(\mathbf{w}_{\text{old}}, w^*) \to \mathbf{w}_{\text{new}}$.

For the death move, an empty component is chosen at random and deleted. Let $w_l$ be the weight of the deleted component. The remaining weights are rescaled to sum to 1 ($w_k' = w_k/(1-w_l)$) and $\boldsymbol{\psi} = (G, \mathbf{w}_{\text{old}}, \boldsymbol{\gamma}, \mathbf{y}) \to \boldsymbol{\psi}' = (G-1, \mathbf{w}_{\text{new}}, \boldsymbol{\gamma}, \mathbf{y})$. The acceptance probability for this move is $\min(1, A)$, where

$$\begin{aligned} A &= \frac{f(\mathbf{X}, \mathbf{y}|\mathbf{w}_{\text{new}}, \boldsymbol{\gamma}, G-1) \times f(\mathbf{w}_{\text{new}}|G-1) \times f(G-1)}{f(\mathbf{X}, \mathbf{y}|\mathbf{w}_{\text{old}}, \boldsymbol{\gamma}, G) \times f(\mathbf{w}_{\text{old}}|G) \times f(G)} \\ &\quad \times \frac{b_{G-1} \times f(w_l)}{d_G \times G_0^{-1}} \times (1-w_l)^{-(G-2)}, \end{aligned} \tag{18}$$

$\frac{f(\mathbf{w}_{\text{new}}|G-1)}{f(\mathbf{w}_{\text{old}}|G)} = \frac{B(\alpha,(G-1)\alpha)}{w_l^{\alpha-1}(1-w_l)^{(G-1)(\alpha-1)}}$, $G_0$ is the number of empty components before the death move, and $f(w_l)$ is the beta$(1, G)$ density.

## 6. POSTERIOR INFERENCE

We draw inference on the sample allocations conditional on $G$. Thus we first need to compute an estimate for this parameter, which we take to be the value most frequently visited by the MCMC sampler. In addition, we need to address the label-switching problem.

### 6.1 Label Switching

In finite mixture models, an identifiability problem arises from the invariance of the likelihood under permutation of the component labels. In the Bayesian paradigm, this leads to symmetric and multimodal posterior distributions with up to $G!$ copies of each "genuine" mode, complicating inference on the parameters. In particular, it is not possible to form ergodic averages over the MCMC samples. Traditional approaches to this problem impose identifiability constraints on the parameters, for instance, $w_1 < \cdots < w_G$. These constraints do not always solve the problem, however. Recently, Stephens (2000b) proposed a relabeling algorithm that takes a decision-theoretic approach. The procedure defines an appropriate loss function and postprocesses the MCMC output to minimize the posterior expected loss.

Let $\mathbf{P}(\boldsymbol{\xi}) = [p_{ik}(\boldsymbol{\xi})]$ be the matrix of allocation probabilities,

$$\begin{aligned} p_{ik}(\boldsymbol{\xi}) &= \Pr(y_i = k|\mathbf{X}, \boldsymbol{\gamma}, \mathbf{w}, \boldsymbol{\theta}) \\ &= \frac{w_k f(\mathbf{x}_{i(\gamma)}|\boldsymbol{\theta}_{k(\gamma)})}{\sum_j w_j f(\mathbf{x}_{i(\gamma)}|\boldsymbol{\theta}_{j(\gamma)})}. \end{aligned} \tag{19}$$

In the context of clustering, the loss function can be defined on the cluster labels $\mathbf{y}$,

$$\mathcal{L}_0(\mathbf{y}; \boldsymbol{\xi}) = -\sum_{i=1}^{n} \log\{p_{iy_i}(\boldsymbol{\xi})\}. \tag{20}$$

Let $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(M)}$ be the sampled parameters and let $\nu_1, \ldots, \nu_M$ be the permutations applied to them. The parameters $\boldsymbol{\xi}^{(t)}$ correspond to the component parameters $\mathbf{w}^{(t)}$, $\boldsymbol{\mu}^{(t)}$, and $\boldsymbol{\Sigma}^{(t)}$ sampled at iteration $t$. However, in our methodology, the component mean and covariance parameters are integrated out, and we do not draw their posterior samples. Therefore, to calculate the matrix of allocation probabilities in (19), we estimate these parameters and set $\tilde{\boldsymbol{\xi}}^{(t)} = \{w_1^{(t)}, \ldots, w_G^{(t)}, \bar{\mathbf{x}}_{1(\gamma)}^{(t)}, \ldots, \bar{\mathbf{x}}_{G(\gamma)}^{(t)}, \mathbf{S}_{1(\gamma)}^{(t)}, \ldots, \mathbf{S}_{G(\gamma)}^{(t)}\}$, where $\bar{\mathbf{x}}_{k(\gamma)}^{(t)}$ and $\mathbf{S}_{k(\gamma)}^{(t)}$ are the estimates of cluster $k$'s mean and covariance based on the model visited at iteration $t$.

The relabeling algorithm proceeds by selecting initial values for the $\nu_t$'s, which we take to be the identity permutation, then iterating the following steps until a fixed point is reached:

a. Choose $\hat{\mathbf{y}}$ to minimize $\sum_{t=1}^{M} \mathcal{L}_0\{\hat{\mathbf{y}}; \nu_t(\tilde{\boldsymbol{\xi}}^{(t)})\}$.
b. For $t = 1, \ldots, M$, choose $\nu_t$ to minimize $\mathcal{L}_0\{\hat{\mathbf{y}}; \nu_t(\tilde{\boldsymbol{\xi}}^{(t)})\}$.

### 6.2 Posterior Densities and Posterior Estimates

Once the label switching is taken care of, the MCMC samples can be used to draw posterior inference. Of particular interest are the allocation vector, $\mathbf{y}$, and the variable selection vector, $\boldsymbol{\gamma}$.

We compute the marginal posterior probability that sample $i$ is allocated to cluster $k$, $y_i = k$, as

$$\begin{aligned} p(y_i = k|\mathbf{X}, G) &= \int p\big(y_i = k, \mathbf{y}_{(-i)}, \boldsymbol{\gamma}, \mathbf{w}\big|\mathbf{X}, G\big)\, d\mathbf{y}_{(-i)}\, d\boldsymbol{\gamma}\, d\mathbf{w} \\ &\propto \int p\big(\mathbf{X}, y_i = k, \mathbf{y}_{(-i)}, \boldsymbol{\gamma}, \mathbf{w}\big|G\big)\, d\mathbf{y}_{(-i)}\, d\boldsymbol{\gamma}\, d\mathbf{w} \\ &= \int p\big(\mathbf{X}, y_i = k, \mathbf{y}_{(-i)}\big|G, \boldsymbol{\gamma}, \mathbf{w}\big) \\ &\qquad \times p(\boldsymbol{\gamma}|G)p(\mathbf{w}|G)\, d\mathbf{y}_{(-i)}\, d\boldsymbol{\gamma}\, d\mathbf{w} \\ &\approx \sum_{t=1}^{M} p\big(\mathbf{X}, y_i^{(t)} = k, \mathbf{y}_{(-i)}^{(t)}\big|G, \boldsymbol{\gamma}^{(t)}, \mathbf{w}^{(t)}\big) \\ &\qquad \times p\big(\boldsymbol{\gamma}^{(t)}\big|G\big)p\big(\mathbf{w}^{(t)}\big|G\big), \end{aligned} \tag{21}$$

where $\mathbf{y}_{(-i)}^{(t)}$ is the vector $\mathbf{y}^{(t)}$ at the $t$th MCMC iteration without the $i$th element. The posterior allocation of sample $i$ can then be estimated by the mode of its marginal posterior density,

$$\widehat{y}_i = \arg\max_{1\le k\le G}\{p(y_i = k|\mathbf{X}, G)\}. \tag{22}$$

Similarly, the marginal posterior for $\gamma_j = 1$ can be computed as

$$\begin{aligned} p(\gamma_j = 1|\mathbf{X}, G) &= \int p\big(\gamma_j = 1, \boldsymbol{\gamma}_{(-j)}, \mathbf{w}, \mathbf{y}|\mathbf{X}, G\big)\, d\boldsymbol{\gamma}_{(-j)}\, d\mathbf{y}\, d\mathbf{w} \\ &\propto \int p\big(\mathbf{X}, \mathbf{y}, \gamma_j = 1, \boldsymbol{\gamma}_{(-j)}, \mathbf{w}\big|G\big)\, d\boldsymbol{\gamma}_{(-j)}\, d\mathbf{y}\, d\mathbf{w} \\ &= \int p\big(\mathbf{X}, \mathbf{y}\big|G, \gamma_j = 1, \boldsymbol{\gamma}_{(-j)}, \mathbf{w}\big) \\ &\qquad \times p(\boldsymbol{\gamma}|G)p(\mathbf{w}|G)\, d\boldsymbol{\gamma}_{(-j)}\, d\mathbf{y}\, d\mathbf{w} \\ &\approx \sum_{t=1}^{M} p\big(\mathbf{X}, \mathbf{y}^{(t)}\big|G, \gamma_j^{(t)} = 1, \boldsymbol{\gamma}_{(-j)}^{(t)}, \mathbf{w}^{(t)}\big) \\ &\qquad \times p\big(\gamma_j = 1, \boldsymbol{\gamma}_{(-j)}^{(t)}\big|G\big)p\big(\mathbf{w}^{(t)}\big|G\big), \end{aligned} \tag{23}$$

where $\boldsymbol{\gamma}_{(-j)}^{(t)}$ is the vector $\boldsymbol{\gamma}^{(t)}$ at the $t$th iteration without the $j$th element. The best discriminating variables can then be identified as those with largest marginal posterior, $p(\gamma_j = 1|X, G) > a$, where $a$ is chosen arbitrarily,

$$\widehat{\gamma}_j = I_{\{p(\gamma_j=1|\mathbf{X},G)>a\}}. \tag{24}$$

An alternative variable selection can be performed by considering the vector $\boldsymbol{\gamma}$ with largest posterior probability among all visited vectors,

$$\widehat{\boldsymbol{\gamma}}^* = \arg\max_{1\le t\le M}\big\{p\big(\boldsymbol{\gamma}^{(t)}\big|\mathbf{X}, G, \widehat{\mathbf{w}}, \widehat{\mathbf{y}}\big)\big\}, \tag{25}$$

where $\widehat{\mathbf{y}}$ is the vector of sample allocations estimated via (22) and $\widehat{\mathbf{w}}$ is given by $\widehat{\mathbf{w}} = \frac{1}{M}\sum_{t=1}^{M} \mathbf{w}^{(t)}$. The estimate given in (25) considers the joint density of $\boldsymbol{\gamma}$ rather than the marginal distributions of the individual elements as (24). In the same spirit, an

estimate of the joint allocation of the samples can be obtained as the configuration that yields the largest posterior probability,

$$\widehat{\mathbf{y}}^* = \arg\max_{1\le t\le M}\{p(\mathbf{y}^{(t)}|\mathbf{X}, G, \widehat{\mathbf{w}}, \widehat{\boldsymbol{\gamma}})\}, \tag{26}$$

where $\widehat{\boldsymbol{\gamma}}$ is the estimate in (24).

### 6.3 Class Prediction

The MCMC output can also be used to predict the class membership of future observations, $\mathbf{X}_f$. The predictive density is given by

$$\begin{aligned} p(y_f = k|\mathbf{X}_f, \mathbf{X}, G) &= \int p(y_f = k, \mathbf{y}, \mathbf{w}, \boldsymbol{\gamma}|\mathbf{X}_f, \mathbf{X}, G)\, d\mathbf{y}\, d\boldsymbol{\gamma}\, d\mathbf{w} \\ &\propto \int p(\mathbf{X}_f, \mathbf{X}, y_f = k, \mathbf{y}|G, \mathbf{w}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}|G) p(\mathbf{w}|G)\, d\mathbf{y}\, d\boldsymbol{\gamma}\, d\mathbf{w} \\ &\approx \sum_{t=1}^{M} p(\mathbf{X}_f, \mathbf{X}, y_f = k, \mathbf{y}^{(t)}|\boldsymbol{\gamma}^{(t)}, \mathbf{w}^{(t)}, G) \\ &\quad \times p(\boldsymbol{\gamma}^{(t)}|G) p(\mathbf{w}^{(t)}|G), \end{aligned} \tag{27}$$

and the observations are allocated according to $\widehat{y}_f$,

$$\widehat{y}_f = \arg\max_{1\le k\le G}\{p(y_f = k|\mathbf{X}_f, \mathbf{X}, G)\}. \tag{28}$$

## 7. PERFORMANCE OF OUR METHODOLOGY

In this section we investigate the performance of our methodology using three datasets. Because the Bayesian clustering problem with an unknown number of components via RJMCMC has been considered only in the univariate setting (Richardson and Green 1997), we first examine the efficiency of our algorithm in the multivariate setting without variable selection. For this, we use the benchmark iris data (Anderson 1935). We then explore the performance of our methodology for simultaneous clustering and variable selection using a series of simulated high-dimensional datasets. We also analyze these data using the COSA algorithm of Friedman and Meulman (2003). Finally, we illustrate an application with DNA microarray data from an endometrial cancer study.

### 7.1 Iris Data

This dataset consists of four covariates—petal width, petal height, sepal width, and sepal height—measured on 50 flowers from each of three species, *Iris setosa*, *Iris versicolor*, and *Iris virginica*. This dataset has been analyzed by several authors (see, e.g., Murtagh and Hernández-Pajares 1995) and can be accessed in S–PLUS as a three-dimensional array.

We rescaled each variable by its range and specified the prior distributions as described in Section 4.1. We took $\delta = 3$, $\alpha = 1$, and $h_1 = 100$. We set $G_{\max} = 10$ and $\kappa_1 = .03$, a value commensurate with the variability in the data. We considered different prior distributions for $G$: truncated Poisson with $\lambda = 5$ and $\lambda = 10$, and a discrete uniform on $[1, \ldots, G_{\max}]$. We conducted the analysis assuming both homogeneous and heterogeneous covariances across clusters. In all cases, we used a burn-in of 10,000 and a main run of 40,000 iterations for inference.

Let us first focus on the results using a truncated Poisson($\lambda = 5$) prior for $G$. Figure 1 displays the trace plots for the number of visited components under the assumption of heterogeneous and homogeneous covariances. Table 1, column 1, gives the corresponding posterior probabilities, $p(G = k|\mathbf{X})$. Table 2 shows the posterior estimates for sample allocations, $\widehat{\mathbf{y}}$, computed conditional on the most probable $G$, according to (22). Under the assumption of heterogeneous covariances across clusters, there is a strong support for $G = 3$. Two of the clusters successfully isolate all of the *setosa* and *virginica* plants, and the third cluster contains all of the *versicolor* except for five identified as *virginica*. These results are satisfactory, because the *versicolor* and *virginica* species are known to have some overlap, whereas the *setosa* is rather well separated. Under the assumption of constant covariance, there is a strong support for four clusters. Again, one cluster contains exclusively all of the *setosa*, and all of the *versicolor* are grouped

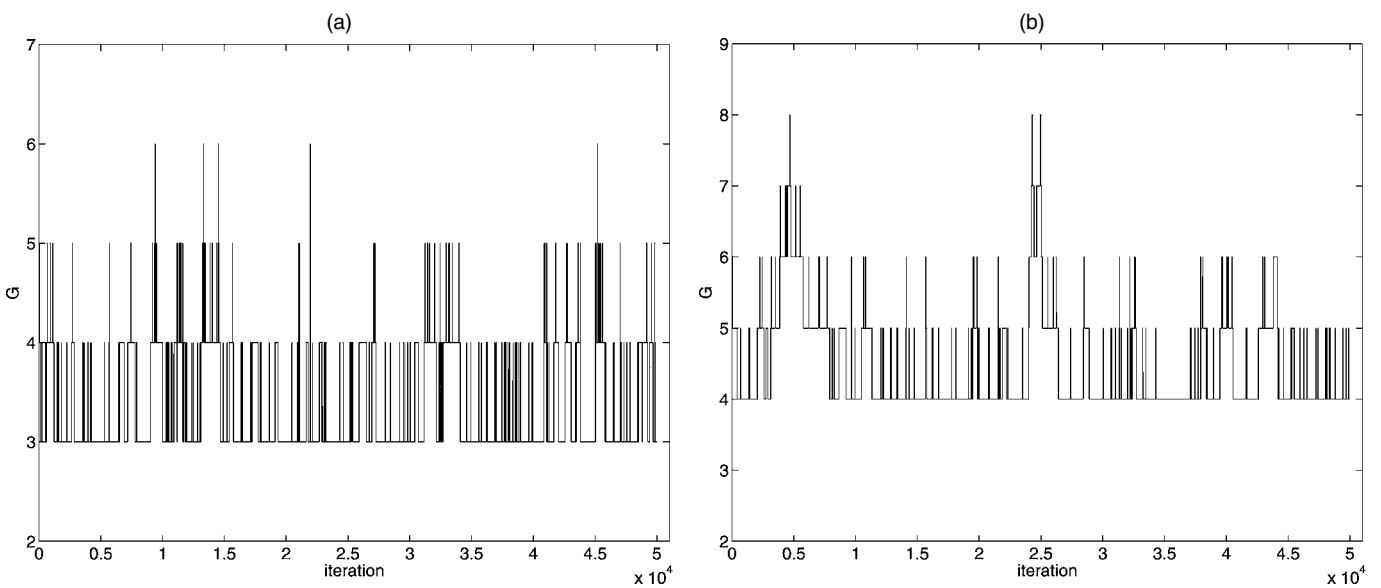


*Figure 1. Iris Data: Trace Plots of the Number of Clusters, G. (a) Heterogeneous covariance; (b) homogeneous covariance.*

*Table 1. Iris Data: Posterior Distribution of G*

| | Poisson($\lambda = 5$) | | Poisson($\lambda = 10$) | | Uniform[1, . . . , 10] | |
|---|---|---|---|---|---|---|
| k | Different $\Sigma_k$ | Equal $\Sigma$ | Different $\Sigma_k$ | Equal $\Sigma$ | Different $\Sigma_k$ | Equal $\Sigma$ |
| 3 | .6466 | 0 | .5019 | 0 | .8548 | 0 |
| 4 | .3124 | .7413 | .4291 | .4384 | .1417 | .6504 |
| 5 | .0403 | .3124 | .0545 | .2335 | .0035 | .3446 |
| 6 | .0007 | .0418 | .0138 | .2786 | 0 | .0050 |
| 7 | 0 | 0 | .0007 | .0484 | 0 | 0 |
| 8 | 0 | 0 | 0 | .0011 | 0 | 0 |

together except for one. Two *virginica* flowers are assigned to the *versicolor* group, and the remaining are divided into two new components. Figure 2 shows the marginal posterior probabilities, $p(y_i = k|\mathbf{X}, G = 4)$, computed using equation (21). We note that some of the flowers among the 12 *virginica* allocated to cluster IV had a close call; for instance, observation 104 had marginal posteriors $p(y_{104} = 3|\mathbf{X}, G) = .4169$ and $p(y_{104} = 4|\mathbf{X}, G) = .5823$.

We examined the sensitivity of the results to the choice of the prior distribution on $G$ and found them to be quite robust. The posterior probabilities $p(G = k|\mathbf{X})$, assuming $G \sim \text{Poisson}(\lambda = 10)$ and $G \sim \text{uniform}[1, \ldots, G_{\max}]$, are given in Table 1. Under both priors, we obtained sample allocations that are identical to those in Table 2. We also found little sensitivity to the choice of the hyperparameter $h_1$, with values between 10 and 1,000 giving satisfactory results. Smaller values of $h_1$ tended to favor slightly more components, but with some clusters containing very few observations. For example, for $h_1 = 10$, under the assumption of equal covariance across clusters, Table 3 shows that the MCMC sampler gives stronger support for $G = 6$. However, one of the clusters contains only one observation and another contains only four observations. Figure 3 displays the marginal posterior probabilities for the allocation of these five samples, $p(y_i = k|\mathbf{X}, G)$; there is little support for assigning them to separate clusters.

We use this example to also illustrate the label-switching phenomenon, described in Section 6.1. Figure 4 gives the trace plots of the component weights under the assumption of homogeneous covariance, before and after processing of the MCMC samples. We can easily identify two label-switching instances in the raw output of Figure 4(a). One of these occurred between the first and third components near iteration 1,000; the other occurred around iteration 2,400 between components 2, 3, and 4. Figure 4(b) shows the trace plots after postprocessing of the MCMC output. We note that the label switching is eliminated, and the multimodality in the estimates of the marginal posterior distributions of $w_k$ is removed. It is now straightforward to obtain sensible estimates using the posterior means.

*Table 2. Iris Data: Sample Allocation, $\hat{\mathbf{y}}$*

| | Heterogeneous covariance | | | Homogeneous covariance | | | |
|---|---|---|---|---|---|---|---|
| | I | II | III | I | II | III | IV |
| Iris setosa | 50 | 0 | 0 | 50 | 0 | 0 | 0 |
| Iris versicolor | 0 | 45 | 5 | 0 | 49 | 1 | 0 |
| Iris virginica | 0 | 0 | 50 | 0 | 2 | 36 | 12 |

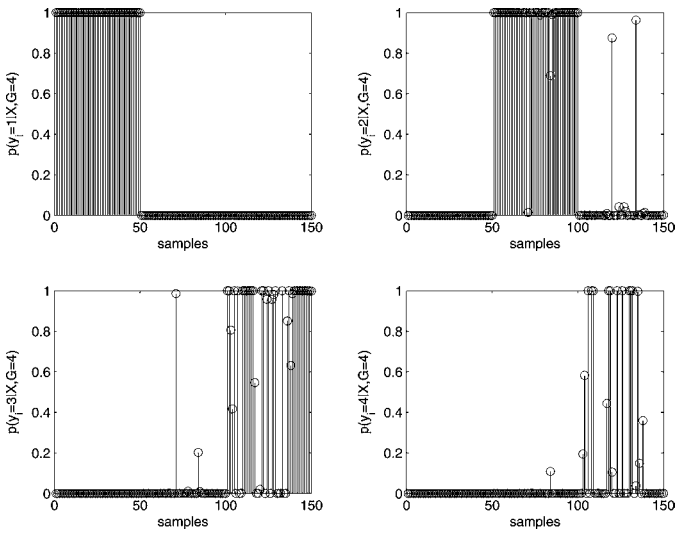


*Figure 2. Iris Data: Poisson($\lambda = 5$) With Equal $\Sigma$-Marginal Posterior Probabilities of Sample Allocations, $p(y_i = k|\mathbf{X}, G = 4)$, $i = 1, \ldots, 150$, $k = 1, \ldots, 4$.*

## 7.2 Simulated Data

We generated a dataset of 15 observations arising from four multivariate normal densities with 20 variables such that

$$x_{ij} \sim I_{\{1\le i\le 4\}}\mathcal{N}(\mu_1, \sigma_1^2) + I_{\{5\le i\le 7\}}\mathcal{N}(\mu_2, \sigma_2^2) + I_{\{8\le i\le 13\}}\mathcal{N}(\mu_3, \sigma_3^2) + I_{\{14\le i\le 15\}}\mathcal{N}(\mu_4, \sigma_4^2), \quad i = 1, \ldots, 15, j = 1, \ldots, 20,$$

where $I_{\{\cdot\}}$ is the indicator function equal to 1 if the condition is met. Thus the first four samples arise from the same distribution, the next three come from the second group, the following six are from the third group, and the last two are from the fourth group. The component means were set to $\mu_1 = 5$, $\mu_2 = 2$, $\mu_3 = -3$, and $\mu_4 = -6$, and the component variances were set to $\sigma_1^2 = 1.5$, $\sigma_2^2 = .1$, $\sigma_3^2 = .5$, and $\sigma_4^2 = 2$. We drew an additional set of $(p - 20)$ noisy variables that do not distinguish between the clusters from a standard normal density. We considered several values of $p$, $p = 50$, 100, 500, 1,000. These data thus contain a small subset of discriminating variables together with a large set of nonclustering variables. The purpose is to study the ability of our method to uncover the cluster structure and identify the relevant covariates in the presence of increasing number of noisy variables.

We permuted the columns of the data matrix, $\mathbf{X}$, to disperse the predictors. For each dataset, we took $\delta = 3$, $\alpha = 1$,

*Table 3. Iris Data. Sensitivity to Hyperparameter h: Results With h = 10*

| | Posterior distribution of G | | | | | | |
|---|---|---|---|---|---|---|---|
| k | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $p(G = k\|\mathbf{X})$ | .0747 | .1874 | .3326 | .2685 | .1120 | .0233 | .0015 |
| | Sample allocations, $\hat{\mathbf{y}}$ | | | | | | |
| | I | II | III | IV | V | VI | |
| Iris setosa | 50 | 0 | 0 | 0 | 0 | 0 | |
| Iris versicolor | 0 | 45 | 1 | 0 | 4 | 0 | |
| Iris virginica | 0 | 2 | 35 | 12 | 0 | 1 | |

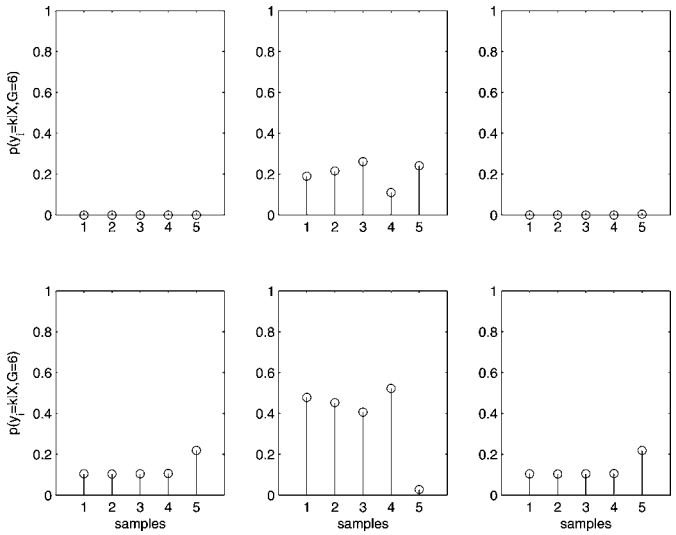


*Figure 3. Iris Data. Sensitivity to $h_1$: Results for $h_1 = 10$–Marginal Posterior Probabilities for Samples Allocated to Clusters V and VI, $p(y_i = k|\boldsymbol{X}, G = 6)$, $i = 1, \ldots, 5$, $k = 1, \ldots, 6$.*

$h_1 = h_0 = 100$, and $G_{\max} = n$ and assumed unequal covariances across clusters. Results from the previous example have shown little sensitivity to the choice of prior on $G$. Here we considered a truncated Poisson prior with $\lambda = 5$. As described in Section 4.1, some care is needed when choosing $\kappa_1$ and $\kappa_0$; their values need to be commensurate with the variability of the data. The amount of total variability in the different simulated datasets is, of course, widely different, and in each case these hyperparameters were chosen proportionally to the upper and lower decile of the non-zero eigenvalues. For the prior of $\boldsymbol{\gamma}$, we chose a Bernoulli distribution and set the expected number of included variables to 10. We used a starting model with one randomly selected variable and ran the MCMC chains for 100,000 iterations, with 40,000 sweeps as burn-in.

Here we do inference on both the sample allocations and the variable selection. We obtained satisfactory results for all datasets. In all cases, we successfully recovered the cluster structure used to simulate the data and identified the 20 discriminating variables. We present the summary plots associated with the largest dataset where $p = 1{,}000$. Figure 5(a) shows the trace plot for the number of visited components, $G$, and Table 4 reports the marginal posterior probabilities, $p(G|\mathbf{X})$. There is a strong support for $G = 4$ and 5, with a slightly larger probability for the former. Figure 6 shows the marginal posterior probabilities of the sample allocations, $p(y_i = k|\mathbf{X}, G = 4)$. We note that these match the group structure used to simulate the data. As for selection of the predictors, Figure 5(b) displays the number of variables selected at each MCMC iteration. In the first 30,000 iterations before burn-in, the chain visited models with around 30–35 variables. After about 40,000 iterations, the chain stabilized to models with 15–20 covariates. Figure 7 displays the marginal posterior probabilities $p(\gamma_j = 1|\mathbf{X}, G = 4)$. There were 17 variables with marginal probabilities greater than .9, all of which were in the set of 20 covariates simulated to effectively discriminate the four clusters. If we lower the threshold for inclusion to .4, then all 20 variables are selected. Conditional on the allocations obtained via (22), the $\boldsymbol{\gamma}$ vector with largest posterior probability (25) among all visited models contained 19 of the 20 discriminating variables.

We also analyzed these datasets using the COSA algorithm of Friedman and Meulman (2003). As we mentioned in Section 2, this procedure performs variable selection in conjunction with hierarchical clustering. We present the results for the analysis of the simulated data with $p = 1{,}000$. Figure 8 shows the dendrogram of the clustering results based on the non-targeted COSA distance. We considered single, average, and complete linkage, but none of these methods was able to recover the true cluster structure. The average linkage dendrogram [Fig. 8(b)] delineates one of the clusters that was partially recovered (the true cluster contains observations 8–13). Figure 8(d) displays the 20 highest relevance values of the variables that lead to this clustering. The variables that define the true cluster structure are indexed as 1–20, and we note that 16 of them appear in this plot.

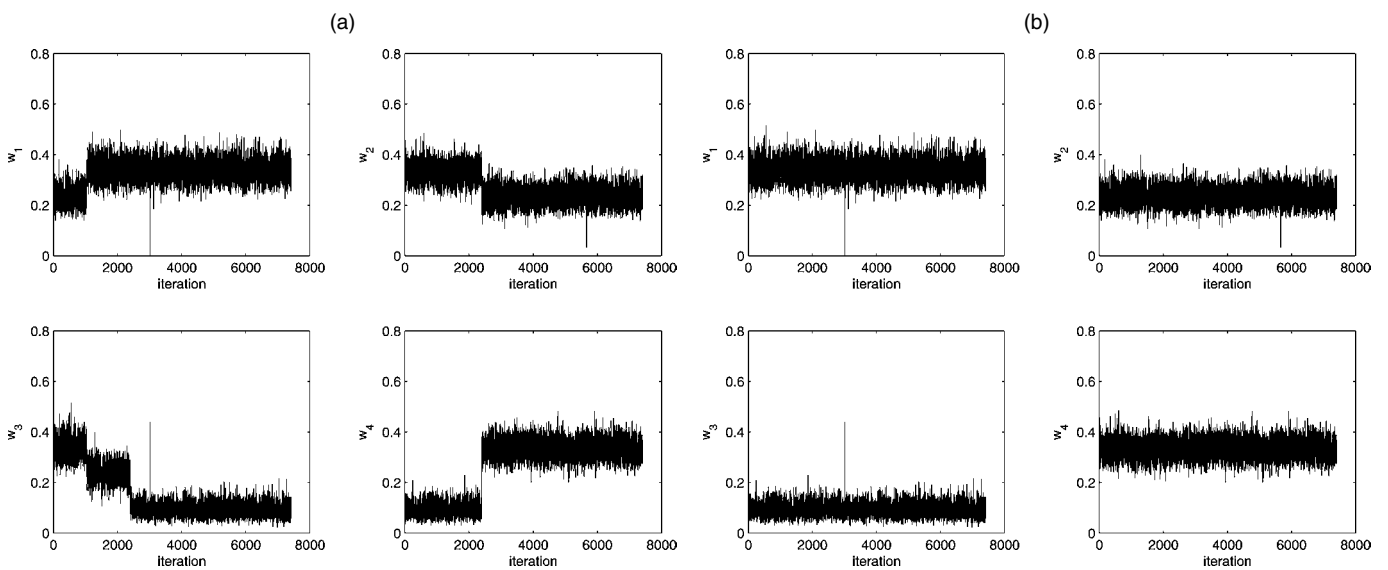


*Figure 4. Iris Data: Trace Plots of Component Weights Before and After Removal of the Label Switching. (a) Raw output; (b) processed output.*

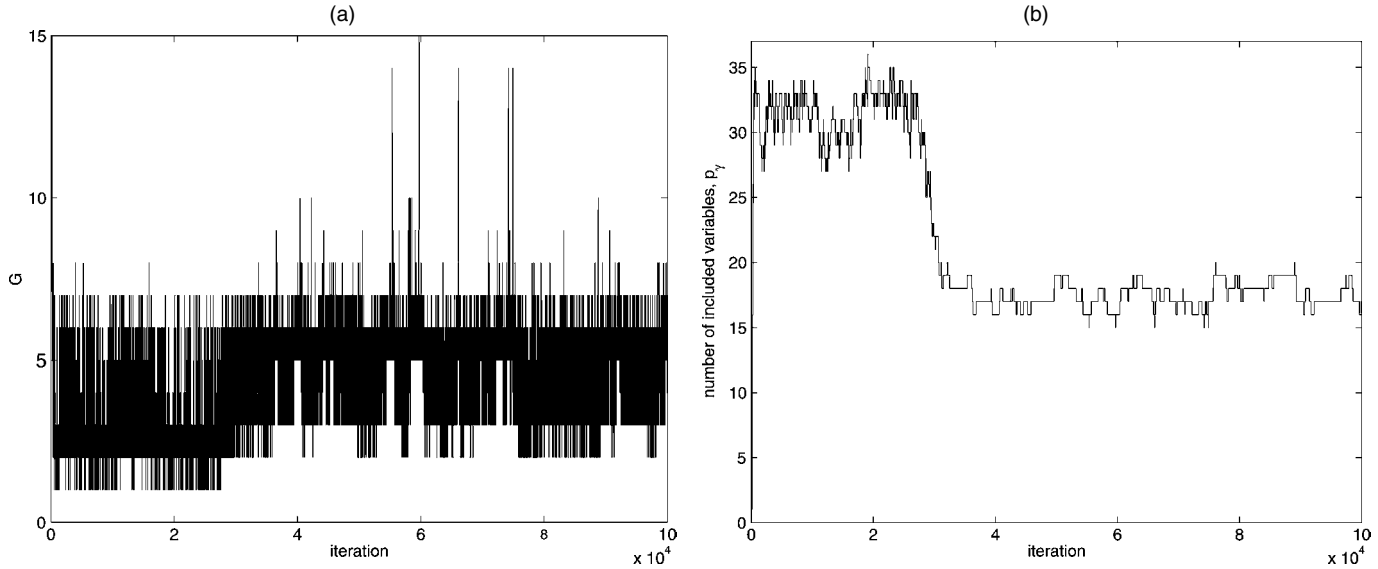


*Figure 5. Trace Plots for Simulated Data With p = 1,000. (a) Number of clusters, G; (b) number of included variables, $p_\gamma$.*

*Table 4. Simulated Data With p = 1,000: Posterior Distribution of G*

| *k* | *p(G = k\|**X**)* |
|---|---|
| 2 | .0111 |
| 3 | .1447 |
| 4 | .3437 |
| 5 | .3375 |
| 6 | .1326 |
| 7 | .0216 |
| 8 | .0040 |
| 9 | .0015 |
| 10 | .0008 |
| 11 | .0005 |
| 12 | .0007 |
| 13 | .0012 |
| 14 | .0002 |

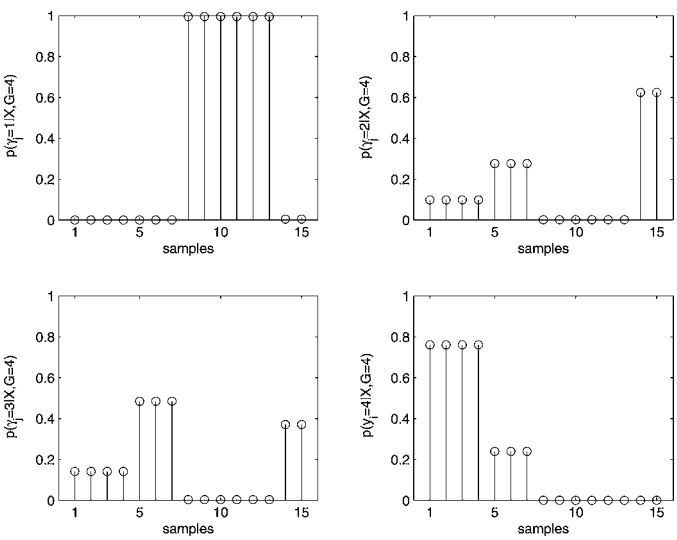


*Figure 6. Simulated Data With p = 1,000: Marginal Posterior Probabilities of Sample Allocations, $p(y_i = k|\mathbf{X}, G = 4)$, $i = 1, \ldots, 15$, $k = 1, \ldots, 4$.*

## 7.3 Application to Microarray Data

We now illustrate the methodology on microarray data from an endometrial cancer study. Endometrioid endometrial adenocarcinoma is a common gynecologic malignancy arising within the uterine lining. The disease usually occurs in postmenopausal women and is associated with the hormonal risk factor of protracted estrogen exposure unopposed by progestins. Despite its prevalence, the molecular mechanisms of its genesis are not completely known. DNA microarrays, with their ability to examine thousands of genes simultaneously, could be an efficient tool to isolate relevant genes and cluster tissues into different subtypes. This is especially important in cancer treatment, where different clinicopathologic groups are known to vary in their response to therapy, and genes identified to discriminate among the different subtypes may represent targets for therapeutic intervention and biomarkers for improved diagnosis.

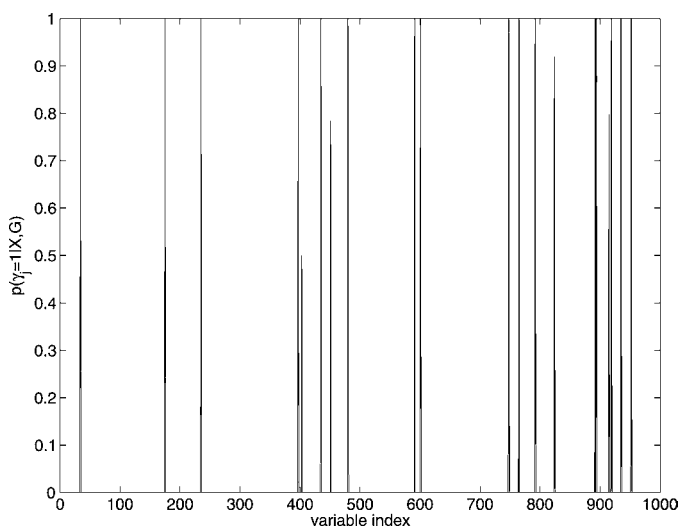


*Figure 7. Simulated Data With p = 1,000: Marginal Posterior Probabilities for Inclusion of Variables, $p(\gamma_j = 1|\mathbf{X}, G = 4)$.*

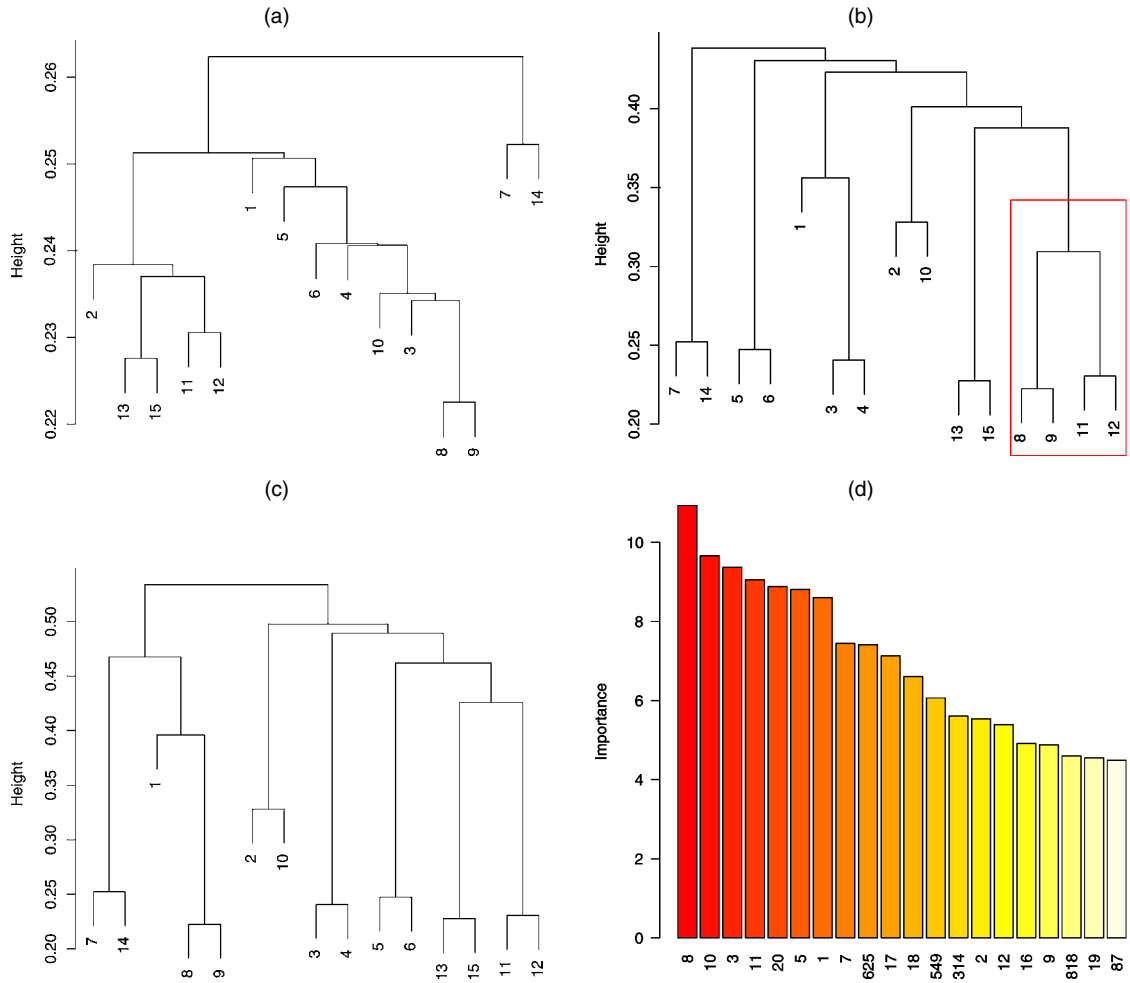


*Figure 8. Analysis of Simulated Data With p = 1,000 Using COSA. (a) Single linkage; (b) average linkage; (c) complete linkage; (d) relevance for delineated cluster.*

Four normal endometrium samples and 10 endometrioid endometrial adenocarcinomas were collected from hysterectomy specimens. RNA samples were isolated from the 14 tissues and reverse-transcribed. The resulting cDNA targets were prepared following the manufacturer's instructions and hybridized to Affymetrix Hu6800 GeneChip arrays, which contain 7,070 probe sets. The scanned images were processed with the Affymetrix software, version 3.1. The average difference derived by the software was used to indicate a transcript abundance, and a global normalization procedure was applied. The technology does not reliably quantify low expression levels, and it is subject to spot saturation at the high end. For this particular array, the limits of reliable detection were previously set to 20 and 16,000 (Tadesse, Ibrahim, and Mutter 2003). We used these same thresholds and removed probe sets with at least one unreliable reading from the analysis. This left us with $p = 762$ variables. We then log-transformed the expression readings to satisfy the assumption of normality, as suggested by the Box–Cox transformation. We also rescaled each variable by its range.

As described in Section 4.1, we specified the priors with the hyperparameters $\delta = 3$, $\alpha = 1$, and $h_1 = h_0 = 100$. We set $\kappa_1 = .001$ and $\kappa_0 = .01$ and assumed unequal covariances across clusters. We used a truncated Poisson$(\lambda = 5)$ prior for $G$, with $G_{\max} = n$. We took a Bernoulli prior for $\boldsymbol{\gamma}$, with an expectation of 10 variables to be included in the model.

To avoid possible dependence of the results on the initial model, we ran four MCMC chains with widely different starting points: (1) $\gamma_j$ set to 0 for all $j$'s except one randomly selected; (2) $\gamma_j$ set to 1 for 10 randomly selected $j$'s; (3) $\gamma_j$ set to 1 for 25 randomly selected $j$'s; and (4) $\gamma_j$ set to 1 for 50 randomly selected $j$'s. For each of the MCMC chains, we ran 100,000 iterations, with 40,000 sweeps taken as a burn-in.

We looked at both the allocation of tissues and the selection of genes whose expression best discriminate among the different groups. The MCMC sampler mixed steadily over the iterations. As shown in the histogram of Figure 9(a), the sampler visited mostly between two and five components, with a stronger support for $G = 3$ clusters. Figure 9(b) shows the trace plot for the number of included variables for one of the chains, which visited mostly models with 25–35 variables. The other chains behaved similarly. Figure 10 gives the marginal posterior probabilities, $p(\gamma_j = 1|\mathbf{X}, G = 3)$, for each of the chains. Despite the very different starting points, the four chains visited similar regions and exhibit broadly similar marginal plots. To assess the concordance of the results across the four chains, we looked at the correlation coefficients between the frequencies for inclusion of genes. These are reported in

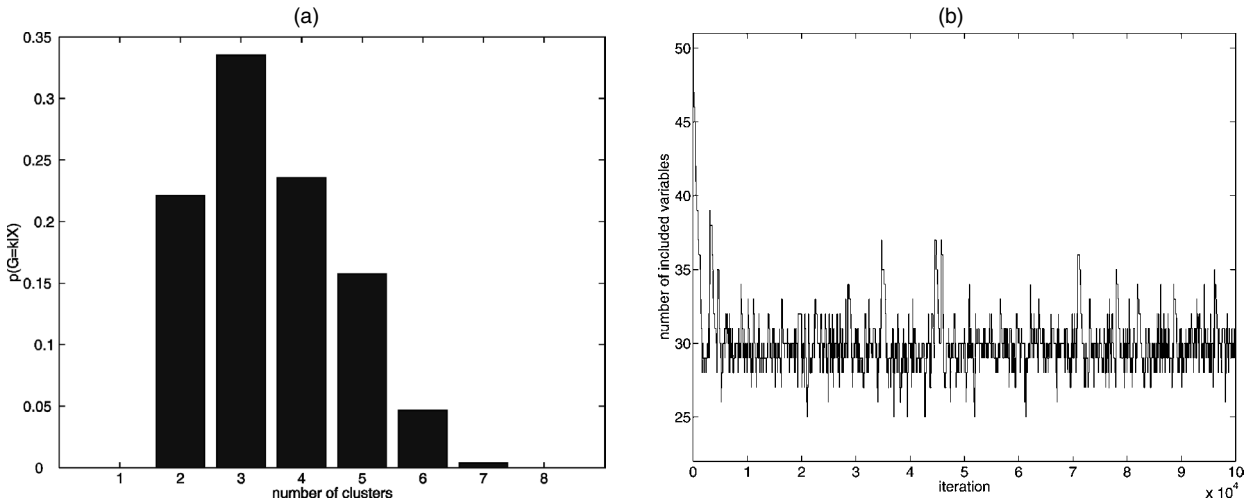


*Figure 9. MCMC Output for Endometrial Data. (a) Number of visited clusters, G; (b) number of included variables, $p_\gamma$.*

Figure 11, along with the corresponding pairwise scatterplots. We see that there is very good agreement across the different MCMC runs. The marginal probabilities, on the other hand, are not as concordant across the chains, because their derivation makes use of model posterior probabilities [see (23)], which are affected by the inclusion of correlated variables. This is a problem inherent to DNA microarray data, where multiple genes with similar expression profiles are often observed.

We pooled and relabeled the output of the four chains, normalized the relative posterior probabilities, and recomputed $p(\gamma_j = 1|\mathbf{X}, G = 3)$. These marginal probabilities are displayed in Figure 12. There are 31 genes with posterior probability greater than .5. We also estimated the marginal posterior probabilities for the sample allocations, $p(y_i = k|\mathbf{X}, G)$, based on the 31 selected genes. These are shown in Figure 13. We successfully identified the four normal tissues. The results also seem to suggest that there are possibly two subtypes within the malignant tissues.

We also present the results from analyzing this dataset with the COSA algorithm. Figure 14 shows the resulting single linkage and average linkage dendrograms, along with the top 30 variables that distinguish the normal and tumor tissues. There is some overlap between the genes selected by COSA and those identified by our procedure. Among the sets of 30 "best" discriminating genes selected by the two methods, 10 are common to both.

## 8. DISCUSSION

We have proposed a method for simultaneously clustering high-dimensional data with an unknown number of components and selecting the variables that best discriminate the different groups. We successfully applied the methodology to various datasets. Our method is fully Bayesian, and we provided standard default recommendations for the choice of priors.

Here we mention some possible extensions and directions for future research. We have drawn posterior inference conditional on a fixed number of components. A more attractive

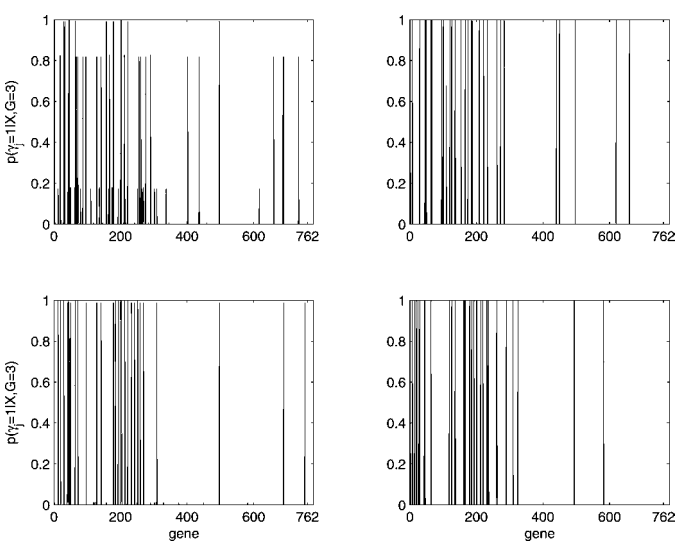


*Figure 10. Endometrial Data. Marginal posterior probabilities for inclusion of variables, $p(\gamma_j = 1|\boldsymbol{X}, G = 3)$, for each of the four chains.*

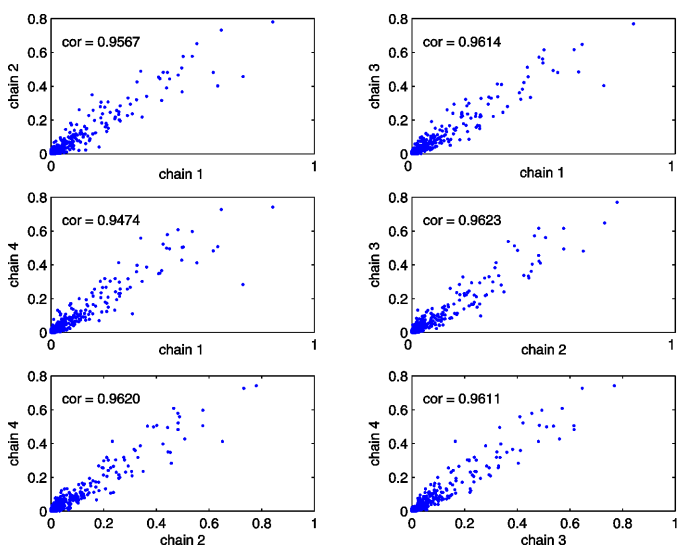


*Figure 11. Endometrial Data. Concordance of results among the four chains: Pairwise scatterplots of frequencies for inclusion of genes.*

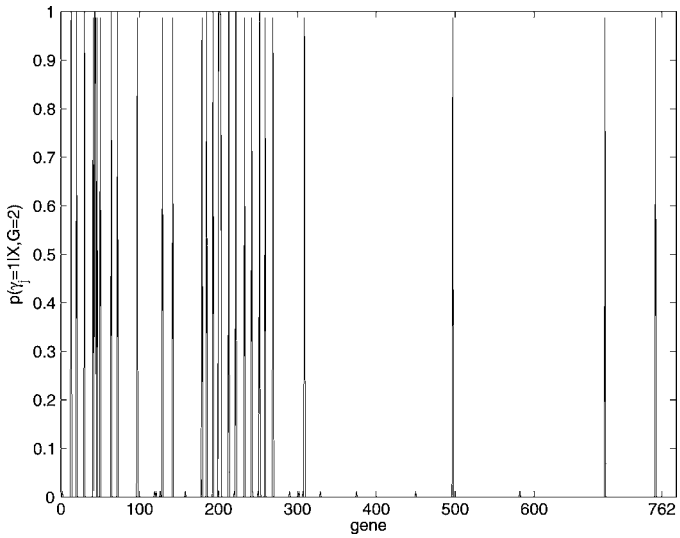


*Figure 12. Endometrial Data. Union of four chains: Marginal posterior probabilities $p(\gamma_j = 1|\boldsymbol{X}, G = 3)$.*

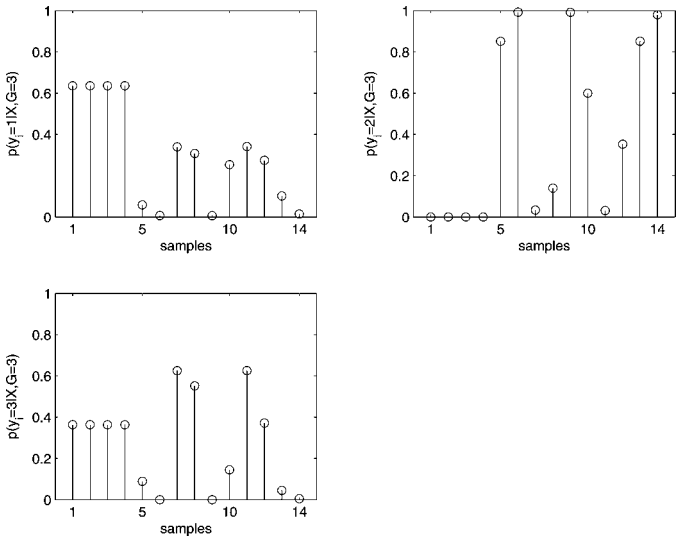


*Figure 13. Endometrial Data. Union of four chains: Marginal posterior probabilities of sample allocations, $p(y_i = k|\boldsymbol{X}, G = 3)$, $i = 1, \ldots, 14$, $k = 1, \ldots, 3$.*

approach would incorporate the uncertainty on this parameter and combine the results for all different values of $G$ visited by the MCMC sampler. To our knowledge, this is an open problem that requires further research. One promising approach involves considering the posterior probability of pairs of ob-

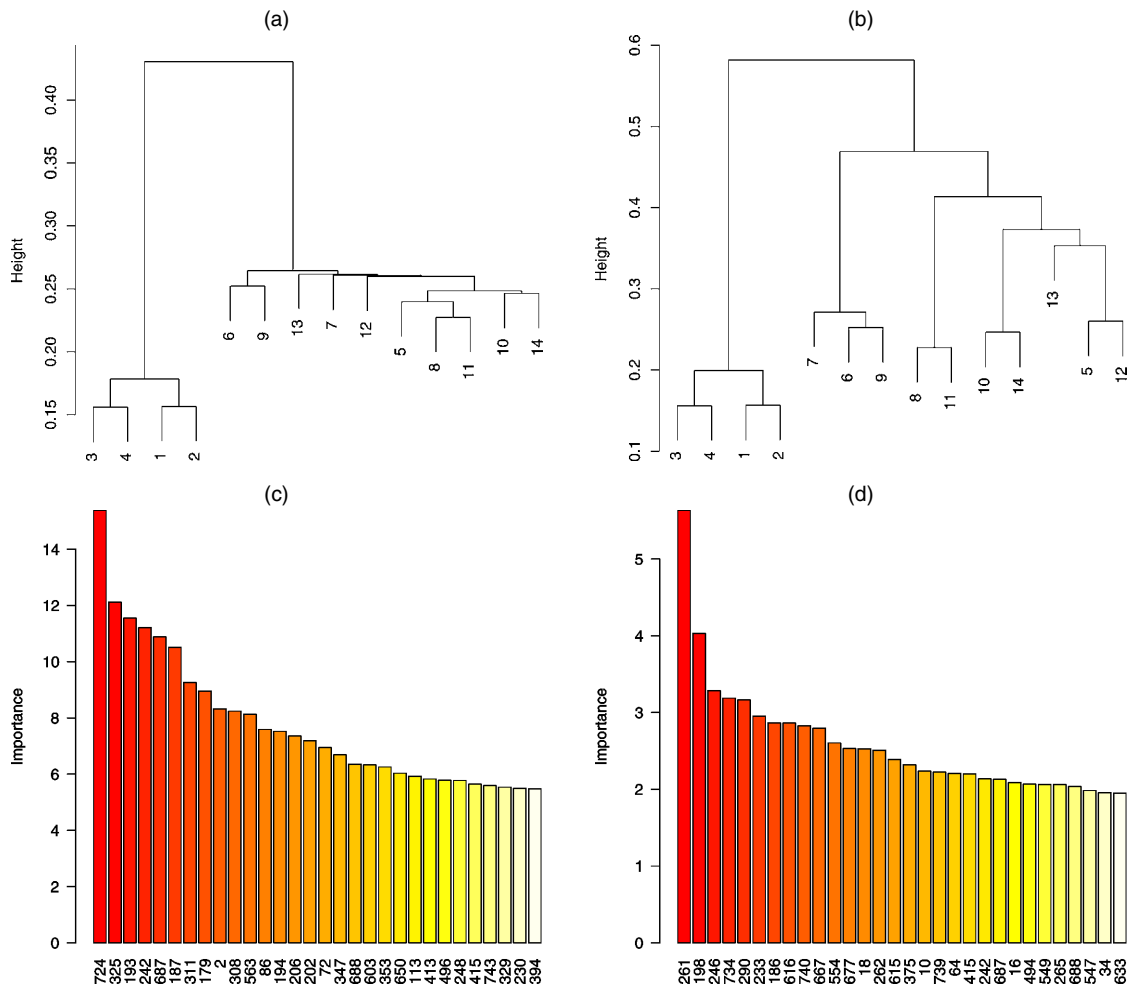


*Figure 14. Analysis of Endometrial Data Using COSA. (a) Single linkage; (b) average linkage; (c) variables leading to cluster 1; (d) variables leading to cluster 2.*

servations being allocated to the same cluster regardless of the number of visited components. As long as there is no interest in estimating the component parameters, the problem of label switching would not be a concern. But this leads to $\binom{n}{2}$ probabilities, and it is not clear how to best summarize them. For example, Medvedovic and Sivaganesan (2002) used hierarchical clustering based on these pairwise probabilities as a similarity measure.

An interesting future avenue would be to investigate the performance of our method when covariance constraints are imposed, as was proposed by Banfield and Raftery (1993). Their parameterization allows one to incorporate different assumptions on the volume, orientation, and shape of the clusters.

Another possible extension is to use empirical Bayes estimates to elicit the hyperparameters. For example, conditional on a fixed number of clusters, a complete log-likelihood can be computed using all covariates and a Monte Carlo EM approach developed for estimation. Alternatively, if prior information were available or subjective priors were preferable, then the prior setting could be modified accordingly. If interactions among predictors were of interest, then additional interaction terms could be included in the model, and the prior on $\boldsymbol{\gamma}$ could be modified to accommodate the additional terms.

## APPENDIX: FULL CONDITIONALS

Here we give details on the derivation of the marginalized full conditionals under the assumption of equal and unequal covariances across clusters.

### Homogeneous Covariance: $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_G = \boldsymbol{\Sigma}$

$$
\begin{aligned}
&f(\mathbf{X}, \mathbf{y}|G, \mathbf{w}, \boldsymbol{\gamma}) \\
&\quad = \int f(\mathbf{X}, \mathbf{y}|G, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta}, \boldsymbol{\Omega}, \boldsymbol{\gamma}) f(\boldsymbol{\mu}|G, \boldsymbol{\Sigma}, \boldsymbol{\gamma}) f(\boldsymbol{\Sigma}|G, \boldsymbol{\gamma}) \\
&\qquad \times f(\boldsymbol{\eta}|\boldsymbol{\Omega}, \boldsymbol{\gamma}) f(\boldsymbol{\Omega}|\boldsymbol{\gamma})\, d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}\, d\boldsymbol{\eta}\, d\boldsymbol{\Omega} \\
&\quad = \int \prod_{k=1}^{G} \{ w_k^{n_k} |\boldsymbol{\Sigma}_{(\gamma)}|^{-(n_k+1)/2} (2\pi)^{-(n_k+1)p_\gamma/2} h_1^{-p_\gamma/2} \} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \sum_{k=1}^{G} \sum_{\mathbf{x}_{i(\gamma)} \in C_k} (\mathbf{x}_{i(\gamma)} - \boldsymbol{\mu}_{k(\gamma)})^T \boldsymbol{\Sigma}_{(\gamma)}^{-1} (\mathbf{x}_{i(\gamma)} - \boldsymbol{\mu}_{k(\gamma)}) \right\} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \sum_{k=1}^{G} (\boldsymbol{\mu}_{k(\gamma)} - \boldsymbol{\mu}_{0(\gamma)})^T (h_1 \boldsymbol{\Sigma}_{(\gamma)})^{-1} (\boldsymbol{\mu}_{k(\gamma)} - \boldsymbol{\mu}_{0(\gamma)}) \right\} \\
&\qquad \times 2^{-p_\gamma(\delta+p_\gamma-1)/2} \pi^{-p_\gamma(p_\gamma-1)/4} \left( \prod_{j=1}^{p_\gamma} \Gamma\left( \frac{\delta + p_\gamma - j}{2} \right) \right)^{-1} \\
&\qquad \times |\mathbf{Q}_{1(\gamma)}|^{(\delta+p_\gamma-1)/2} |\boldsymbol{\Sigma}_{(\gamma)}|^{-(\delta+2p_\gamma)/2} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}_{(\gamma)}^{-1} \mathbf{Q}_{1(\gamma)}) \right\} \\
&\qquad \times |\boldsymbol{\Omega}_{(\gamma^c)}|^{-(n+1)/2} (2\pi)^{-(n+1)(p-p_\gamma)/2} h_0^{-(p-p_\gamma)/2} \\
&\qquad \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\} \\
&\qquad \times \exp\left\{ -\frac{1}{2} (\boldsymbol{\eta}_{(\gamma^c)} - \boldsymbol{\mu}_{0(\gamma^c)})^T (h_0 \boldsymbol{\Omega}_{(\gamma^c)})^{-1} (\boldsymbol{\eta}_{(\gamma^c)} - \boldsymbol{\mu}_{0(\gamma^c)}) \right\} \\
&\qquad \times 2^{-(p-p_\gamma)(\delta+p-p_\gamma-1)/2} \pi^{-(p-p_\gamma)(p-p_\gamma-1)/4} \\
&\qquad \times \left( \prod_{j=1}^{p-p_\gamma} \Gamma\left( \frac{\delta + p - p_\gamma - j}{2} \right) \right)^{-1} \\
&\qquad \times |\mathbf{Q}_{0(\gamma^c)}|^{(\delta+p-p_\gamma-1)/2} |\boldsymbol{\Omega}_{(\gamma^c)}|^{-(\delta+2(p-p_\gamma))/2} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \operatorname{tr}(\boldsymbol{\Omega}_{(\gamma^c)}^{-1} \mathbf{Q}_{0(\gamma^c)}) \right\} d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}\, d\boldsymbol{\eta}\, d\boldsymbol{\Omega} \\
&\quad = \int \prod_{k=1}^{G} \left\{ (2\pi)^{-(n_k+1)p_\gamma/2} h_1^{-p_\gamma/2} w_k^{n_k} |\boldsymbol{\Sigma}_{(\gamma)}|^{-(n_k+1)/2} \right. \\
&\qquad \left. \times \exp\left\{ -\frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}_{(\gamma)}^{-1} \mathbf{S}_{k(\gamma)}) \right\} \right\} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \sum_{k=1}^{G} \left( \boldsymbol{\mu}_{k(\gamma)} - \frac{\sum_{\mathbf{x}_{i(\gamma)} \in C_k} \mathbf{x}_{i(\gamma)} + h_1^{-1} \boldsymbol{\mu}_{0(\gamma)}}{n_k + h_1^{-1}} \right)^T \right. \\
&\qquad \left. \times (n_k + h_1^{-1}) \boldsymbol{\Sigma}_{(\gamma)}^{-1} \left( \boldsymbol{\mu}_{k(\gamma)} - \frac{\sum_{\mathbf{x}_{i(\gamma)} \in C_k} \mathbf{x}_{i(\gamma)} + h_1^{-1} \boldsymbol{\mu}_{0(\gamma)}}{n_k + h_1^{-1}} \right) \right\} \\
&\qquad \times 2^{-p_\gamma(\delta+p_\gamma-1)/2} \pi^{-p_\gamma(p_\gamma-1)/4} \left( \prod_{j=1}^{p_\gamma} \Gamma\left( \frac{\delta + p_\gamma - j}{2} \right) \right)^{-1} \\
&\qquad \times |\mathbf{Q}_{1(\gamma)}|^{(\delta+p_\gamma-1)/2} |\boldsymbol{\Sigma}_{(\gamma)}|^{-(\delta+2p_\gamma)/2} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}_{(\gamma)}^{-1} \mathbf{Q}_{1(\gamma)}) \right\} \\
&\qquad \times (2\pi)^{-(n+1)(p-p_\gamma)/2} h_0^{-(p-p_\gamma)/2} |\boldsymbol{\Omega}_{(\gamma^c)}|^{-(n+1)/2} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \operatorname{tr}(\boldsymbol{\Omega}_{(\gamma^c)}^{-1} \mathbf{S}_{0(\gamma^c)}) \right\} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \left( \boldsymbol{\eta}_{(\gamma^c)} - \frac{\sum_{i=1}^{n} \mathbf{x}_{i(\gamma^c)} + h_0^{-1} \boldsymbol{\mu}_{0(\gamma^c)}}{n + h_0^{-1}} \right)^T \right. \\
&\qquad \left. \times (n + h_0^{-1}) \boldsymbol{\Omega}_{(\gamma^c)}^{-1} \left( \boldsymbol{\eta}_{(\gamma^c)} - \frac{\sum_{i=1}^{n} \mathbf{x}_{i(\gamma^c)} + h_0^{-1} \boldsymbol{\mu}_{0(\gamma^c)}}{n + h_0^{-1}} \right) \right\} \\
&\qquad \times 2^{-(p-p_\gamma)(\delta+p-p_\gamma-1)/2} \pi^{-(p-p_\gamma)(p-p_\gamma-1)/4} \\
&\qquad \times \left( \prod_{j=1}^{p-p_\gamma} \Gamma\left( \frac{\delta + p - p_\gamma - j}{2} \right) \right)^{-1} \\
&\qquad \times |\mathbf{Q}_{0(\gamma^c)}|^{(\delta+p-p_\gamma-1)/2} |\boldsymbol{\Sigma}_{(\gamma^c)}|^{-(\delta+2(p-p_\gamma))/2} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}_{0(\gamma^c)}^{-1} \mathbf{Q}_{0(\gamma^c)}) \right\} d\boldsymbol{\mu}\, d\boldsymbol{\Sigma}\, d\boldsymbol{\eta}\, d\boldsymbol{\Omega} \\
&\quad = \left[ \prod_{k=1}^{G} (h_1 n_k + 1)^{-p_\gamma/2} w_k^{n_k} \right] \pi^{-np_\gamma/2} \\
&\qquad \times |\mathbf{Q}_{1(\gamma)}|^{(\delta+p_\gamma-1)/2} 2^{-p_\gamma(\delta+n+p_\gamma-1)/2} \pi^{-p_\gamma(p_\gamma-1)/4} \\
&\qquad \times \left( \prod_{j=1}^{p_\gamma} \Gamma\left( \frac{\delta + p_\gamma - j}{2} \right) \right)^{-1} \\
&\qquad \times \int_{\boldsymbol{\Sigma}} |\boldsymbol{\Sigma}_{(\gamma)}|^{-(\delta+n+2p_\gamma)/2} \\
&\qquad \times \exp\left\{ -\frac{1}{2} \operatorname{tr}\left( \boldsymbol{\Sigma}_{(\gamma)}^{-1} \left[ \mathbf{Q}_{1(\gamma)} + \sum_{k=1}^{G} \mathbf{S}_{k(\gamma)} \right] \right) \right\} d\boldsymbol{\Sigma}
\end{aligned}
$$

$$
\begin{aligned}
&\times (h_0 n+1)^{-(p-p_\gamma)/2}\pi^{-n(p-p_\gamma)/2}\\
&\times \left|\mathbf{Q}_{0(\gamma^c)}\right|^{(\delta+p-p_\gamma-1)/2}\\
&\times 2^{-(p-p_\gamma)(\delta+n+p-p_\gamma-1)/2}\pi^{-(p-p_\gamma)(p-p_\gamma-1)/4}\\
&\times \left(\prod_{j=1}^{p-p_\gamma}\Gamma\left(\frac{\delta+p-p_\gamma-j}{2}\right)\right)^{-1}\\
&\times \int_{\mathbf{\Omega}}\left|\mathbf{\Omega}_{(\gamma^c)}\right|^{-(\delta+n+2(p-p_\gamma))/2}\\
&\times \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Omega}_{(\gamma^c)}^{-1}\left[\mathbf{Q}_{0(\gamma)}+\mathbf{S}_{0(\gamma^c)}\right]\right)\right\}d\Omega\\
&=\pi^{-np/2}\left[\prod_{k=1}^{G}(h_1 n_k+1)^{-p_\gamma/2}w_k^{n_k}\right](h_0 n+1)^{-(p-p_\gamma)/2}\\
&\times\prod_{j=1}^{p_\gamma}\left(\Gamma\left(\frac{\delta+n+p_\gamma-j}{2}\right)\Big/\Gamma\left(\frac{\delta+p_\gamma-j}{2}\right)\right)\\
&\times\prod_{j=1}^{p-p_\gamma}\left(\Gamma\left(\frac{\delta+n+p-p_\gamma-j}{2}\right)\Big/\Gamma\left(\frac{\delta+p-p_\gamma-j}{2}\right)\right)\\
&\times\left|\mathbf{Q}_{1(\gamma)}\right|^{(\delta+p_\gamma-1)/2}\left|\mathbf{Q}_{0(\gamma^c)}\right|^{(\delta+p-p_\gamma-1)/2}\\
&\times\left|\mathbf{Q}_{1(\gamma)}+\sum_{k=1}^{G}\mathbf{S}_{k(\gamma)}\right|^{-(\delta+n+p_\gamma-1)/2}\\
&\times\left|\mathbf{Q}_{0(\gamma^c)}+\mathbf{S}_{0(\gamma^c)}\right|^{-(\delta+n+p-p_\gamma-1)/2},
\end{aligned}
$$

where $\mathbf{S}_{k(\gamma)}$ and $\mathbf{S}_{0(\gamma)}$ are given by

$$
\begin{aligned}
\mathbf{S}_{k(\gamma)}&=\sum_{\mathbf{x}_{i(\gamma)}\in C_k}\mathbf{x}_{i(\gamma)}\mathbf{x}_{i(\gamma)}^T+h_1^{-1}\boldsymbol{\mu}_{0(\gamma)}\boldsymbol{\mu}_{0(\gamma)}^T\\
&\quad-(n_k+h_1^{-1})^{-1}\Bigg(\sum_{\mathbf{x}_{i(\gamma)}\in C_k}\mathbf{x}_{i(\gamma)}+h_1^{-1}\boldsymbol{\mu}_{0(\gamma)}\Bigg)\\
&\quad\times\Bigg(\sum_{\mathbf{x}_{i(\gamma)}\in C_k}\mathbf{x}_{i(\gamma)}+h_1^{-1}\boldsymbol{\mu}_{0(\gamma)}\Bigg)^T\\
&=\sum_{\mathbf{x}_{i(\gamma)}\in C_k}\left(\mathbf{x}_{i(\gamma)}-\bar{\mathbf{x}}_{k(\gamma)}\right)\left(\mathbf{x}_{i(\gamma)}-\bar{\mathbf{x}}_{k(\gamma)}\right)^T\\
&\quad+\frac{n_k}{h_1 n_k+1}\left(\boldsymbol{\mu}_{0(\gamma)}-\bar{\mathbf{x}}_{k(\gamma)}\right)\left(\boldsymbol{\mu}_{0(\gamma)}-\bar{\mathbf{x}}_{k(\gamma)}\right)^T
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{S}_{0(\gamma)}&=\sum_{i=1}^{n}\mathbf{x}_{i(\gamma^c)}\mathbf{x}_{i(\gamma^c)}^T+h_0^{-1}\boldsymbol{\mu}_{0(\gamma^c)}\boldsymbol{\mu}_{0(\gamma^c)}^T\\
&\quad-(n+h_0^{-1})^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_{i(\gamma^c)}+h_0^{-1}\boldsymbol{\mu}_{0(\gamma^c)}\right)\\
&\quad\times\left(\sum_{i=1}^{n}\mathbf{x}_{i(\gamma^c)}+h_0^{-1}\boldsymbol{\mu}_{0(\gamma^c)}\right)^T\\
&=\sum_{i=1}^{n}\left(\mathbf{x}_{i(\gamma^c)}-\bar{\mathbf{x}}_{(\gamma^c)}\right)\left(\mathbf{x}_{i(\gamma^c)}-\bar{\mathbf{x}}_{(\gamma^c)}\right)^T\\
&\quad+\frac{n}{h_0 n+1}\left(\boldsymbol{\mu}_{0(\gamma^c)}-\bar{\mathbf{x}}_{(\gamma^c)}\right)\left(\boldsymbol{\mu}_{0(\gamma^c)}-\bar{\mathbf{x}}_{(\gamma^c)}\right)^T,
\end{aligned}
$$

with $\bar{\mathbf{x}}_{k(\gamma)}=\frac{1}{n_k}\sum_{\mathbf{x}_{i(\gamma)}\in C_k}\mathbf{x}_{i(\gamma)}$ and $\bar{\mathbf{x}}_{(\gamma^c)}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i(\gamma^c)}$.

### Heterogeneous Covariances

$$
\begin{aligned}
&f(\mathbf{X},\mathbf{y}|G,\mathbf{w},\boldsymbol{\gamma})\\
&\quad=\int f(\mathbf{X},\mathbf{y}|G,\mathbf{w},\boldsymbol{\mu},\mathbf{\Sigma},\boldsymbol{\eta},\mathbf{\Omega},\boldsymbol{\gamma})f(\boldsymbol{\mu}|G,\mathbf{\Sigma},\boldsymbol{\gamma})f(\mathbf{\Sigma}|G,\boldsymbol{\gamma})\\
&\qquad\times f(\boldsymbol{\eta}|\mathbf{\Omega},\boldsymbol{\gamma})f(\mathbf{\Omega}|\boldsymbol{\gamma})\,d\boldsymbol{\mu}\,d\mathbf{\Sigma}\,d\boldsymbol{\eta}\,d\mathbf{\Omega}\\
&\quad=\int\prod_{k=1}^{G}\Bigg\{w_k^{n_k}\left|\mathbf{\Sigma}_{k(\gamma)}\right|^{-(n_k+1)/2}(2\pi)^{-(n_k+1)p_\gamma/2}h_1^{-p_\gamma/2}\\
&\qquad\times 2^{-p_\gamma(\delta+p_\gamma-1)/2}\pi^{-p_\gamma(p_\gamma-1)/4}\left(\prod_{j=1}^{p_\gamma}\Gamma\left(\frac{\delta+p_\gamma-j}{2}\right)\right)^{-1}\Bigg\}\\
&\qquad\times\exp\left\{-\frac{1}{2}\sum_{k=1}^{G}\sum_{\mathbf{x}_{i(\gamma)}\in C_k}\left(\mathbf{x}_{i(\gamma)}-\boldsymbol{\mu}_k\right)^T\mathbf{\Sigma}_{k(\gamma)}^{-1}\left(\mathbf{x}_{i(\gamma)}-\boldsymbol{\mu}_k\right)\right\}\\
&\qquad\times\exp\Bigg\{-\frac{1}{2}\sum_{k=1}^{G}\left(\boldsymbol{\mu}_{k(\gamma)}-\boldsymbol{\mu}_{0(\gamma)}\right)^T\left(h_1\mathbf{\Sigma}_{k(\gamma)}\right)^{-1}\\
&\qquad\times\left(\boldsymbol{\mu}_{k(\gamma)}-\boldsymbol{\mu}_{0(\gamma)}\right)\Bigg\}\\
&\qquad\times\prod_{k=1}^{G}\Bigg\{\left|\mathbf{Q}_{1(\gamma)}\right|^{(\delta+p_\gamma-1)/2}\left|\mathbf{\Sigma}_{k(\gamma)}\right|^{-(\delta+2p_\gamma)/2}\\
&\qquad\times\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}_{k(\gamma)}^{-1}\mathbf{Q}_{1(\gamma)}\right)\right\}\Bigg\}\,d\boldsymbol{\mu}\,d\mathbf{\Sigma}\\
&\qquad\times\pi^{-n(p-p_\gamma)/2}(h_0 n+1)^{-(p-p_\gamma)/2}\\
&\qquad\times\prod_{j=1}^{p-p_\gamma}\left(\Gamma\left(\frac{\delta+n+p-p_\gamma-j}{2}\right)\Big/\Gamma\left(\frac{\delta+p-p_\gamma-j}{2}\right)\right)\\
&\qquad\times\left|\mathbf{Q}_{0(\gamma^c)}\right|^{(\delta+p-p_\gamma-1)/2}\left|\mathbf{Q}_{0(\gamma^c)}+\mathbf{S}_{0(\gamma^c)}\right|^{-(\delta+n+p-p_\gamma-1)/2}\\
&\quad=\int_{\mathbf{\Sigma}}\int_{\boldsymbol{\mu}}\prod_{k=1}^{G}\Bigg\{(2\pi)^{-(n_k+1)p_\gamma/2}h_1^{-p_\gamma/2}w_k^{n_k}2^{-p_\gamma(\delta+p_\gamma-1)/2}\\
&\qquad\times\pi^{-p_\gamma(p_\gamma-1)/4}\left(\prod_{j=1}^{p_\gamma}\Gamma\left(\frac{\delta+p_\gamma-j}{2}\right)\right)^{-1}\\
&\qquad\times\left|\mathbf{\Sigma}_{k(\gamma)}\right|^{-(n_k+1)/2}\Bigg\}\\
&\qquad\times\exp\Bigg\{-\frac{1}{2}\sum_{k=1}^{G}\Bigg(\boldsymbol{\mu}_{k(\gamma)}-\frac{\sum_{\mathbf{x}_{i(\gamma)}\in C_k}\mathbf{x}_{i(\gamma)}+h_1^{-1}\boldsymbol{\mu}_{0(\gamma)}}{n_k+h_1^{-1}}\Bigg)^T\\
&\qquad\times(n_k+h_1^{-1})\mathbf{\Sigma}_{k(\gamma)}^{-1}\Bigg(\boldsymbol{\mu}_{k(\gamma)}-\frac{\sum_{\mathbf{x}_{i(\gamma)}\in C_k}\mathbf{x}_{i(\gamma)}+h_1^{-1}\boldsymbol{\mu}_{0(\gamma)}}{n_k+h_1^{-1}}\Bigg)\Bigg\}\\
&\qquad\times\prod_{k=1}^{G}\Bigg\{\left|\mathbf{Q}_{1(\gamma)}\right|^{(\delta+p_\gamma-1)/2}\left|\mathbf{\Sigma}_{k(\gamma)}\right|^{-(\delta+2p_\gamma)/2}\\
&\qquad\times\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}_{k(\gamma)}^{-1}\left[\mathbf{Q}_{1(\gamma)}+\mathbf{S}_{k(\gamma)}\right]\right)\right\}\Bigg\}\,d\boldsymbol{\mu}\,d\mathbf{\Sigma}\\
&\qquad\times\pi^{-n(p-p_\gamma)/2}(h_0 n+1)^{-(p-p_\gamma)/2}
\end{aligned}
$$

$$\times \prod_{j=1}^{p-p_\gamma} \left( \Gamma\left( \frac{\delta + n + p - p_\gamma - j}{2} \right) \Big/ \Gamma\left( \frac{\delta + p - p_\gamma - j}{2} \right) \right)$$

$$\times \left| \mathbf{Q}_{0(\gamma^c)} \right|^{(\delta+p-p_\gamma-1)/2} \left| \mathbf{Q}_{0(\gamma^c)} + \mathbf{S}_{0(\gamma^c)} \right|^{-(\delta+n+p-p_\gamma-1)/2}$$

$$= \pi^{-np/2} \prod_{k=1}^{G} \Bigg\{ (hn_k + 1)^{-p_\gamma/2} w_k^{n_k} \left| \mathbf{Q}_{(\gamma)} \right|^{(\delta+p_\gamma-1)/2}$$

$$\times 2^{-p_\gamma(\delta+n_k+p_\gamma-1)/2} \pi^{-p_\gamma(p_\gamma-1)/4}$$

$$\times \left( \prod_{j=1}^{p_\gamma} \Gamma\left( \frac{\delta + p_\gamma - j}{2} \right) \right)^{-1} \Bigg\}$$

$$\times \int_{\mathbf{\Sigma}} \prod_{k=1}^{G} \Bigg\{ \left| \mathbf{\Sigma}_{k(\gamma)} \right|^{-(\delta+n_k+2p_\gamma)/2}$$

$$\times \exp\left\{ -\frac{1}{2} \operatorname{tr}\left( \mathbf{\Sigma}_{k(\gamma)}^{-1} \left[ \mathbf{Q}_{(\gamma)} + \mathbf{S}_{k(\gamma)} \right] \right) \right\} \Bigg\} d\mathbf{\Sigma}$$

$$\times (h_0 n + 1)^{-(p-p_\gamma)/2}$$

$$\times \prod_{j=1}^{p-p_\gamma} \left( \Gamma\left( \frac{\delta + n + p - p_\gamma - j}{2} \right) \Big/ \Gamma\left( \frac{\delta + p - p_\gamma - j}{2} \right) \right)$$

$$\times \left| \mathbf{Q}_{0(\gamma^c)} \right|^{(\delta+p-p_\gamma-1)/2} \left| \mathbf{Q}_{0(\gamma^c)} + \mathbf{S}_{0(\gamma^c)} \right|^{-(\delta+n+p-p_\gamma-1)/2}$$

$$= \pi^{-np/2} \prod_{k=1}^{G} \Bigg\{ (hn_k + 1)^{-p_\gamma/2} w_k^{n_k}$$

$$\times \prod_{j=1}^{p_\gamma} \left( \Gamma\left( \frac{\delta + n_k + p_\gamma - 1}{2} \right) \Big/ \left( \Gamma\left( \frac{\delta + p_\gamma - 1}{2} \right) \right) \right) \Bigg\}$$

$$\times \prod_{k=1}^{G} \left\{ \left| \mathbf{Q}_{(\gamma)} \right|^{(\delta+p_\gamma-1)/2} \left| \mathbf{Q}_{(\gamma)} + \mathbf{S}_{k(\gamma)} \right|^{-(\delta+n_k+p_\gamma-1)/2} \right\}$$

$$\times (h_0 n + 1)^{-(p-p_\gamma)/2}$$

$$\times \prod_{j=1}^{p-p_\gamma} \left( \Gamma\left( \frac{\delta + n + p - p_\gamma - j}{2} \right) \Big/ \left( \Gamma\left( \frac{\delta + p - p_\gamma - j}{2} \right) \right) \right)$$

$$\times \left| \mathbf{Q}_{0(\gamma^c)} \right|^{(\delta+p-p_\gamma-1)/2} \left| \mathbf{Q}_{0(\gamma^c)} + \mathbf{S}_{0(\gamma^c)} \right|^{-(\delta+n+p-p_\gamma-1)/2}.$$

*[Received May 2003. Revised August 2004.]*

## REFERENCES


Anderson, E. (1935), "The Irises of the Gaspe Peninsula," *Bulletin of the American Iris Society*, 59, 2–5.

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821.

Brown, P. J. (1993), *Measurement, Regression, and Calibration*, Oxford, U.K.: Clarendon Press.

Brown, P. J., Vannucci, M., and Fearn, T. (1998a), "Bayesian Wavelength Selection in Multicomponent Analysis," *Chemometrics*, 12, 173–182.

——— (1998b), "Multivariate Bayesian Variable Selection and Prediction," *Journal of the Royal Statistical Society*, Ser. B, 60, 627–641.

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

Chang, W.-C. (1983), "On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions," *Applied Statistics*, 32, 267–275.

Diebolt, J., and Robert, C. P. (1994), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society*, Ser. B, 56, 363–375.

Fowlkes, E. B., Gnanadesikan, R., and Kettering, J. R. (1988), "Variable Selection in Clustering," *Journal of Classification*, 5, 205–228.

Friedman, J. H., and Meulman, J. J. (2003), "Clustering Objects on Subsets of Attributes," technical report, Stanford University, Dept. of Statistics and Stanford Linear Accelerator Center.

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Ghosh, D., and Chinnaiyan, A. M. (2002), "Mixture Modelling of Gene Expression Data From Microarray Experiments," *Bioinformatics*, 18, 275–286.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Gnanadesikan, R., Kettering, J. R., and Tao, S. L. (1995), "Weighting and Selection of Variables for Cluster Analysis," *Journal of Classification*, 12, 113–136.

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

Madigan, D., and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002), "A Mixture Model-Based Approach to the Clustering of Microarray Expression Data," *Bioinformatics*, 18, 413–422.

Medvedovic, M., and Sivaganesan, S. (2002), "Bayesian Infinite Mixture Model-Based Clustering of Gene Expression Profiles," *Bioinformatics*, 18, 1194–1206.

Milligan, G. W. (1989), "A Validation Study of a Variable Weighting Algorithm for Cluster Analysis," *Journal of Classification*, 6, 53–71.

Murtagh, F., and Hernández-Pajares, M. (1995), "The Kohonen Self-Organizing Map Method: An Assessment," *Journal of Classification*, 12, 165–190.

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*, Ser. B, 59, 731–792.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., 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.

Stephens, M. (2000a), "Bayesian Analysis of Mixture Models With an Unknown Number of Components—An Alternative to Reversible-Jump Methods," *The Annals of Statistics*, 28, 40–74.

——— (2000b), "Dealing With Label Switching in Mixture Models," *Journal of the Royal Statistical Society*, Ser. B, 62, 795–809.

Tadesse, M. G., Ibrahim, J. G., and Mutter, G. L. (2003), "Identification of Differentially Expressed Genes in High-Density Oligonucleotide Arrays Accounting for the Quantification Limits of the Technology," *Biometrics*, 59, 542–554.