Bayesian Analysis (2011) **6**, Number 4, pp. 665–690

Bayesian Outlier Detection with Dirichlet Process Mixtures

Matthew S. Shotwell* and Elizabeth H. Slate†

Abstract. We introduce a Bayesian inference mechanism for outlier detection using the augmented Dirichlet process mixture. Outliers are detected by forming a maximum *a posteriori* (MAP) estimate of the data partition. Observations that comprise small or singleton clusters in the estimated partition are considered outliers. We offer a novel interpretation of the Dirichlet process precision parameter, and demonstrate its utility in outlier detection problems. The precision parameter is used to form an outlier detection criterion based on the Bayes factor for an outlier partition versus a class of partitions with fewer or no outliers. We further introduce a computational method for MAP estimation that is free of posterior sampling, and guaranteed to find a MAP estimate in finite time. The novel methods are compared with several established strategies in a yeast microarray time series.

Keywords: partition, optimization, Bayes factor

1 Introduction

Outliers are often accommodated at the expense of model complexity. For example, Box and Tiao (1968) formulate separate models for 'good' and 'bad' observations. The outlier detection task is to evaluate the evidence favoring a complex model over a simpler one that does not accommodate outliers. Bayesian model selection techniques involving the Bayes factor and related quantities have been utilized in this context by Petit (1992), Hoeting et al. (1996), and Bayarri and Morales (2003).

A common outlier paradigm dictates that most observations in an experiment arise uniformly from a single stochastic process, and a small number of outliers are generated

*Department of Biostatistics, School of Medicine, Vanderbilt University, Nashville, TN, Matt.Shotwell@Vanderbilt.edu

†Division of Biostatistics and Epidemiology, Department of Medicine, Medical University of South Carolina, Charleston, SC slate@musc.edu

© 2011 International Society for Bayesian Analysis DOI:10.1214/11-BA625

by a different process. Alternatively, each observation may be thought to arise from one of a multitude of heterogeneous processes, and observations arising from infrequently realized processes are considered outliers. The latter notion of outlying observations is naturally approached from a clustering, or partitioning perspective, where the goal is to distinguish observations generated by different processes.

Several clustering methods have been considered in the context of outlier detection, including the Partitioning Around Medoids (PAM), Clustering Large Applications (CLARA), and k-means algorithms (Kaufman and Rousseeuw 1990; Al-Zoubi 2009; Hautamaki et al. 2004). In these methods, the number of clusters is fixed. However, selecting the number of clusters is not trivial. Clustering methods used in this manner are criticized for lack of clear outlier detection criteria. That is, the criteria are not probabilistic, or not easily deduced from the procedure (Ben-Gal 2005).

The Dirichlet process mixture (DPM) model has enjoyed popularity in methodological and applied research. The tendency for DPM models to agglomerate similar observations is an exploited feature. Their utility in outlier detection problems was noted previously by Quintana and Iglesias (2003), and Quintana (2004), who develop a sophisticated decision theory in order to balance model complexity and optimal parameter estimation. Their method is illustrated in the context of outlier detection.

The DPM induces a prior distribution over the data partition. Hence, inference on statistics of the data partition, such as the number of clusters, is a consequence of the posterior distribution. Manipulation of the Dirichlet process precision parameter yields a simple and intuitive criterion for outlier detection with DPM models, which is the principal contribution of this article.

Markov chain Monte Carlo (MCMC) methods are the primary means for summarizing posterior quantities in DPM models (MacEachern 1994; Escobar and West 1995; Ishwaran and James 2001; Green and Richardson 2001). In applications where the DPM is used for partitioning or outlier detection, it may be unnecessary to draw representative samples from a posterior distribution. In these cases, simpler and more computationally efficient methods are available.

This paper describes a simple mechanism for outlier detection using Dirichlet process mixtures. Since DPMs may be constructed from a broad class of statistical models, these methods establish a uniform outlier detection criterion for all such models. We motivate the MAP estimator for the data partition and develop an explicit outlier detection criterion in Section 2. In Section 3, we discuss the literature on computational methods

for DPMs and describe a novel stochastic alternative for MAP estimation. The methods are illustrated using a microarray gene expression dataset in Section 4. The remainder of this section provides additional background and motivation for the Dirichlet process mixture model.

1.1 Dirichlet Process Mixtures

The Dirichlet process, also known as the Ferguson distribution was developed as a probability distribution on the space of probability distributions (Ferguson 1973). The notion of a distribution over distributions was motivated by problems in Bayesian density estimation. Coincidentally, Ferguson (1961) had considered the problem of outlier detection earlier, though not in the context of the namesake distribution.

Suppose G_0 is a probability distribution, and G is a random probability distribution, both with support in space $\mathcal{B}$. Then G is distributed according to a Dirichlet process with base distribution G_0, and precision parameter $\alpha > 0$, if for all finite $r = 1, 2, \ldots, m < \infty$ and measurable partitions $\{B_1, \ldots, B_r\}$ of $\mathcal{B}$, the vector $\{G(B_1), \ldots, G(B_r)\}$ has a Dirichlet distribution with parameter $\{\alpha G_0(B_1), \ldots, \alpha G_0(B_r)\}$. The precision parameter α determines how precisely $\{G(B_1), \ldots, G(B_r)\}$ vary about the Dirichlet mean $\{\alpha G_0(B_1), \ldots, \alpha G_0(B_r)\}$. In notation, $G \sim \mathrm{DP}(\alpha, G_0)$.

Consider a sequence of random variables $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^n$, drawn independently from a DP distributed probability distribution such that

$$\begin{aligned} \theta_j &\sim G \\ G &\sim \mathrm{DP}(\alpha, G_0). \end{aligned}$$

Then the posterior distribution of G is a Dirichlet process with precision $\alpha + n$ and base distribution

$$\frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{j=1}^{n} \delta_{\theta_j},$$

where δ_θ is the Dirac probability mass function, placing all mass at θ. In this way, α is interpreted as a 'prior sample size', or the strength of prior belief in the base measure G_0.

The Polya urn scheme (Blackwell and MacQueen 1973) is a generative construction for $\{\theta_j\}_{j=1}^n$, identified by marginalizing with respect to the DP-distributed measure G.

The Polya urn yields the following sequence of conditional density functions:

$$p(\theta_j|\boldsymbol{\theta}_{1:(j-1)}) \propto \alpha G_0(\theta_j) + \sum_{k=1}^{j-1} \delta_{\theta_k}, \tag{1}$$

where $\boldsymbol{\theta}_{1:(j-1)} = \{\theta_1, \ldots, \theta_{j-1}\}$. Formula (1) captures the essence of the Polya urn scheme and motivates how clustering occurs among draws from a DP-distributed probability distribution. That is, equation (1) assigns positive probability where θ_j is identical to θ_k for some $\theta_k \in \boldsymbol{\theta}_{1:(j-1)}$. We say that θ_j and θ_k are *clustered* when they take identical values.

Antoniak (1974) used the Polya urn construction to further characterize the DP precision parameter α through the expected number of distinct values among $\{\theta_j\}_{j=1}^n$, denoted r. Antoniak gives the expression:

$$E[r] = \sum_{j=1}^{n} \frac{\alpha}{(\alpha + j - 1)}. \tag{2}$$

The expected number of clusters is monotonic in α. Hence, from the Polya urn perspective, α regulates how often distinct values arise in the sequence $\{\theta_j\}_{j=1}^n$.

Consider a random sample $\mathbf{y} = \{y_1, \ldots, y_n\}$, where $f(y_j|\theta_j)$ is a probability density function indexed by θ_j, then

$$\begin{aligned} y_j|\theta_j &\sim f(y_j|\theta_j) \\ \theta_j &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

constitutes a *Dirichlet Process Mixture* model. Conditional on θ_j, y_j are independent from the remaining sample observations. The posterior density function for $\theta_j|\boldsymbol{\theta}_{1:(j-1)}$ is then

$$p(\theta_j|\boldsymbol{\theta}_{1:(j-1)}, y_j) \propto \alpha G_0(\theta_j) f(y_j|\theta_j) + \sum_{k=1}^{j-1} \delta_{\theta_k}(\theta_j) f(y_j|\theta_k). \tag{3}$$

Escobar and West (1995) proposed a posterior sampling algorithm based on this form of the conditional posterior distribution.

1.2 Augmented Dirichlet Process Mixtures

The posterior mass function for $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ is a product of the n terms given by expression (3). For even moderate n this product is intractable. The augmented construction of the DPM yields a more tractable conditional posterior mass function. The

parameter $\boldsymbol{\theta}$ decomposes into two components, $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_r\}$ and $\mathbf{z} = \{z_1, \ldots, z_n\}$, where $\boldsymbol{\phi}$ is the set of r unique values in $\boldsymbol{\theta}$ and $\mathbf{z}$ is the set of n cluster membership variables such that $z_j = k$ if and only if $\theta_j = \phi_k$. The variable $\mathbf{z}$ represents the partition of n observations into r clusters and is later referenced as the data partition variable. In practice, $\mathbf{z}$ is usually written and coded as a vector of labels, often integers. However, the partition $\mathbf{z}$ is invariant under permutations of the cluster labels.

Conditioning on $\mathbf{z}$, the posterior density function for $\boldsymbol{\phi}$ is given by

$$p(\phi_1, \ldots, \phi_r | \mathbf{z}, \mathbf{y}) \propto \prod_{k=1}^{r} G_0(\phi_k) L(\phi_k | \mathbf{y}^{(k)}),$$

where $L(\phi_k|\mathbf{y}^{(k)})$ is the likelihood function and $\mathbf{y}^{(k)} = \{y_j : z_j = k,\ j = 1, \ldots, n\}$ is the set of observations assigned to the k^{th} cluster. The number of observations in $\mathbf{y}^{(k)}$ is later denoted $n^{(k)}$. Conditional on $\mathbf{z}$, the distinct $\{\phi_1, \ldots, \phi_r\}$ are independent *a posteriori*. In hierarchical notation, the augmented DPM may be written

$$\begin{aligned} y_j | z_j = k &\sim f(y_j|\phi_k) \\ \phi_k &\sim G_0 \\ p(\mathbf{z}) &\propto \alpha^r \prod_{k=1}^{r} \Gamma(n^{(k)}), \end{aligned} \tag{4}$$

where the prior mass function $p(\mathbf{z})$ is obtained from the Polya urn representation of the DPM. The augmented DPM is a special type of product partition model (Hartigan 1990, PPM).

Variants of the augmented Dirichlet process mixture have been exploited in the Gibbs sampling algorithms of MacEachern (1994), Bush and MacEachern (1996), and MacEachern and Müller (1998), among others. The posterior distribution over the data partition variable $\mathbf{z}$ may be approximated by sequentially sampling from the full conditional distributions with mass functions

$$p(z_j | \mathbf{z}_{-j}, \mathbf{y}) \propto \begin{cases} \int \alpha f(y_j|\phi) G_0(\phi) d\phi & z_j \neq z_i \text{ for all } z_i \in \mathbf{z}_{-j} \\ \int n_{-j}^{(k)} f(y_j|\phi) L(\phi|\mathbf{y}_{-j}^{(k)}) G_0(\phi) d\phi & z_j = z_i = k \text{ for some } z_i \in \mathbf{z}_{-j} \end{cases}, \tag{5}$$

where notation with subscript $-j$ indicates all but the j^{th} observation. This method is later referenced as the Polya urn Gibbs sampler.

The augmented DPM formulation makes the data partition explicit and simplifies the notions of splitting and merging clusters. A single cluster whose member observations

have been redistributed into exactly two new clusters is said to have undergone a 'split' operation. Two clusters whose observations are reassigned to a single new cluster are said to have undergone a 'merge' operation.

2 Outlier Detection

Consider a data partition $\mathbf{z}_o$ consisting of r_o clusters, of which zero or more are *outlier clusters*, where an outlier cluster contains $n_o \ll n$ or fewer observations. Let $\mathcal{M}_o$ be the union of all partitions formed by any sequence of merge operations on the clusters of the partition $\mathbf{z}_o$. Hence, each partition $\mathbf{z}_m \in \mathcal{M}_o$ consists of fewer clusters, and the size of each cluster is the sum of one or more cluster sizes of $\mathbf{z}_o$. Conversely, $\mathbf{z}_o$ may be formed by a sequence of split operations on the clusters of a partition $\mathbf{z}_m \in \mathcal{M}_o$. Since $\mathcal{M}_o$ contains all the partitions that result from a merge operation on the outlier clusters of $\mathbf{z}_o$, the outlier detection problem may be cast as a decision between $\mathbf{z}_o$ and the members of $\mathcal{M}_o$.

The cost of making a poor decision about $\mathbf{z}_o$, given the true partition $\mathbf{z}_T$, is encoded by a loss function. In the terminology of Berger (1985), $\mathbf{z}_T$ is the true state of nature at the time of an experiment. The conditional Bayes decision principle prescribes that the *Bayes estimate*, or *Bayes decision* minimizes the expected value of the loss function with respect to the posterior distribution of the unobserved state of nature $\mathbf{z}$ (Berger 1985; Hogg et al. 2005).

The zero-one loss function

$$L(\mathbf{z}, \mathbf{z}_T) = \begin{cases} 1 & \mathbf{z} \neq \mathbf{z}_T \\ 0 & \mathbf{z} = \mathbf{z}_T \end{cases},$$

yields the decision principle of largest posterior mass, and the Bayes estimator $\hat{\mathbf{z}}$ that maximizes the marginal posterior distribution over $\mathbf{z}$. Hence, $\hat{\mathbf{z}}$ is the maximum *a posteriori* (MAP) estimate of $\mathbf{z}_T$.

Statistical decision making on the basis of largest posterior mass is independently intuitive and useful. Hence, the zero-one loss function is often used implicitly. However, Lau and Green (2007) argue for an alternative Bayes decision principle designed to recover the true partition. The corresponding loss function is a weighted count of all observation pairs that cluster discordantly between the estimated and true partition. For equal weights, this is equivalent to maximizing the expected Rand (1971) index between the estimated and true partition. The following discussions assume the decision

principle of largest posterior mass. Alternative decision principles may also yield useful criteria for outlier detection.

Under the decision principle of largest posterior mass, outliers are detected only when an outlier partition $\mathbf{z}_o$ satisfies the posterior condition $p(\mathbf{z}_o)p(\mathbf{y}|\mathbf{z}_o) > p(\mathbf{z}_m)p(\mathbf{y}|\mathbf{z}_m)$ for all $\mathbf{z}_m \in \mathcal{M}_o$. Substituting (4) for $p(\mathbf{z})$ and rearranging gives the equivalent condition,

$$\frac{p(\mathbf{y}|\mathbf{z}_o)}{p(\mathbf{y}|\mathbf{z}_m)} > \frac{1}{\alpha^v} \frac{\prod_{k=1}^{r_m} \Gamma(n_m^{(k)})}{\prod_{k=1}^{r_o} \Gamma(n_o^{(k)})}, \tag{6}$$

where v is the difference in the number of clusters between $\mathbf{z}_o$ and $\mathbf{z}_m$, $v = r_o - r_m$.

The left-hand side of (6) is the Bayes factor for an outlier model $\mathbf{z}_o$ versus a model $\mathbf{z}_m \in \mathcal{M}_o$, denoted $BF_{o/m}$. A criterion for outlier detection is made by imposing a minimum value on $BF_{o/m}$. That is, outliers are detected only when the Bayes factor favoring an outlier partition exceeds the lower bound.

The second ratio on the right-hand side of (6) takes a minimum value of one for all $\mathbf{z}_m \in \mathcal{M}_o$. For proof, consider that each $n_m^{(\cdot)}$ is the sum of one or more $n_o^{(k)}$ for $k = 1 \dots r_o$. The ratio is then a product of r_m multinomial coefficients, each taking a minimum value of one. Hence, the quantity $1/\alpha^v$ forms a lower bound on the Bayes factor favoring an outlier partition $\mathbf{z}_o$ versus any partition $\mathbf{z}_m \in \mathcal{M}_o$. The lower bound is conservative because the second ratio on the right-hand side of (6) is often much greater than one. An exact criterion may be computed for any pair of partitions $\mathbf{z}_o$ and $\mathbf{z}_m \in \mathcal{M}_o$ by substitution in inequality (6).

Fixing the value of α ensures that outliers are detected only when the associated (integrated) likelihood is at least $1/\alpha^v$ times that of any partition formed by a sequence of v merge operations on the outlier partition. In this context, the inverse precision parameter $1/\alpha$ is interpreted as the minimum Bayes factor required to detect a partition with outliers. We recommend Jeffreys' scale of evidence for Bayes factors (Jeffreys 1961; Efron and Gous 2001) as a guide for selecting and interpreting the value of $1/\alpha$.

For the criterion to be valid, an estimated partition $\hat{\mathbf{z}}$ need only have greater posterior mass than any partition formed by a sequence of merge operations on the estimate. Note that a MAP estimate automatically satisfies this requirement. However, for many outlier detection problems, satisfying this requirement is much less complex. In addition, this property implies a certain robustness under poorly computed MAP estimates. Hence, an estimate that does not fully maximize the marginal posterior mass function may still satisfy the outlier detection criterion.

2.1 Finite Mixture Comparison

Consider the marginal likelihood $p(\mathbf{y}|\mathbf{z})$ as a *classification likelihood* in the finite mixture model framework. Fraley and Raftery (2002) propose an outlier detection strategy by optimizing the associated Bayesian Information Criterion (BIC). The BIC is given by $-\frac{1}{2}BIC(\mathbf{y}|\mathbf{z}) = \log p(\mathbf{y}|\mathbf{z}) - \frac{\rho}{2}r\log(n)$, where r is the number of distinct clusters (*i.e.* mixture components), and ρ is the fraction of free parameters per cluster. The model-based clustering strategy identifies outlier clusters by maximizing $-\frac{1}{2}BIC(\mathbf{y}|\mathbf{z})$, or equivalently $p(\mathbf{y}|\mathbf{z})n^{-\frac{\rho}{2}r}$. Hence, for an outlier partition $\mathbf{z}_o$ and a partition $\mathbf{z}_m \in \mathcal{M}_o$ (defined in the preceding discussion), outliers are detected in the finite mixture framework when

$$\frac{p(\mathbf{y}|\mathbf{z}_o)}{p(\mathbf{y}|\mathbf{z}_m)} > n^{\frac{\rho}{2}v}. \tag{7}$$

Fraley and Raftery (2002) point to results supporting the appropriateness of optimizing the BIC in classification problems, including consistency (in the classical sense) for the number of clusters. Note that optimizing the BIC also imposes a lower bound on the Bayes factor favoring the outlier partition. Indeed, by fixing the DPM precision parameter such that

$$\frac{1}{\alpha^v} = n^{\frac{\rho}{2}v},$$

the evidence required to detect outliers using the DPM method is at least that imposed by the BIC.

By changing the prior distribution over data partitions from that given in (4) to $p(\mathbf{z}) \propto \alpha^r$ and taking $\alpha = n^{-\rho/2}$, the posterior mass function is identical to the BIC penalized classification likelihood. Hence, the corresponding MAP estimate is identical to the estimate obtained by optimizing the BIC. This property is used in Section 4 to draw comparison between the proposed DPM method, and the finite mixture/BIC strategy.

2.2 Marginal Likelihood Concerns

Recent work in Bayesian nonparametrics has cast doubt on the utility of some traditional parametric modeling strategies in semiparametric and nonparametric models. The findings of Wang and Dunson (2011) suggest that partition estimation for nonparametric applications is sensitive to over-fitting under marginal likelihood models. The authors recommend a pseudo-marginal likelihood (PML), or a product of the conditional pre-

dictive ordinates (Geisser 1980, CPO), which imposes a leave-one-out cross-validation strategy to avoid over-fitting. Although we develop an outlier detection criterion in the context of a marginal likelihood $p(\mathbf{y}|\mathbf{z})$, a pseudo-marginal likelihood and the associated pseudo Bayes factor may be substituted. However, the value of $1/\alpha$ should be carefully selected to reflect comparison of PMLs.

Bush et al. (2010) further warn that strategies to address prior sensitivity in parametric models may not be appropriate in semiparametric models. MacEachern and Guha (2011) recently explained a related paradox, where a posterior distribution under a semiparametric model is more concentrated than that of a corresponding parametric model, even when the semiparametric prior distribution is less concentrated. These concerns are directed toward semiparametric and nonparametric models where the likelihoods of interest are marginal with respect to the data partition. In contrast, the likelihood terms of inequality (6) are conditional on a data partition, and marginal with respect to all other parameters. In this sense, the outlier detection criterion draws comparison between two nested parametric models, rather than semiparametric or nonparametric models.

For the Bayes factor of inequality (6), the numerator likelihood always has more terms, with equal or fewer observations contributing to each cluster-specific likelihood. Consequently, the contribution of prior information is weighted more strongly in the numerator than in the denominator. Where prior information is informative, the imbalance of prior contribution may be a useful shrinkage mechanism in clusters of the numerator partition. When the prior is not intended to be informative, the effect may be weakened by selecting a diffuse or improper prior. In specific cases, especially in conjugate models, the contribution of the prior to the marginal likelihood may be examined directly, and suitably adjusted to reflect prior belief. We return to this point in an applied context in Section 4.

3 Estimation

Computational methods for summarizing posterior quantities in Dirichlet process mixtures are varied in design and purpose. The data partition induced by the DPM is often a nuisance in nonparametric applications. Estimates of the partition need not be optimal in order to yield good nonparametric predictions. A variety of computational strategies exploit this property to reduce the computational burden associated with nonparametric and semiparametric inference in DPMs.

Most recently, Wang and Dunson (2011) propose a sequential update and greedy search (SUGS), and compare their method with others in this paradigm, including predictive recursion (Newton 2002), variational approximation (Blei and Jordan 2006), and sequential importance sampling (MacEachern et al. 1999). Wang and Dunson (2011) evaluate the method on the criteria of computing speed and the Kullback-Leibler divergence between a density used to simulate data, and a predictive density computed conditional on the partition estimate. The authors demonstrate that SUGS gives fast 'reasonable' partition estimates for use in nonparametric inference. Direct comparison of SUGS on other criteria unfairly ignores the purpose of its design.

The authors of the SUGS method suggest multiple random initializations to combat order dependence in the partition estimate. As a minor adaptation of the SUGS method, we propose to substitute the multiple random initialization steps with Polya urn Gibbs updates. We refer to the resulting optimization strategy as SUGS++. In order to emphasize the variability in computational strategies, we return to the SUGS and SUGS++ methods in Section 4.

The outlier detection criterion presented in Section 2 requires the estimated data partition to have greater posterior mass than any partition formed by a sequence of merge operations on its clusters. This condition may not hold when a MAP estimate is poorly computed and subsequent inferences may be invalid.

A MAP estimator for the data partition variable $\mathbf{z}$ is generally not available in closed form. The data partition in a DPM has support over the number of possible partitions of n observations, or the n^{th} Bell number (Bell 1934; Rota 1964). Hence, enumerative optimization is not currently feasible for n much larger than ten. Reasonable approximate solutions are the subject of active research.

Heard et al. (2006) extend the simple agglomerative method of Ward (1963) in order to compute the MAP estimate in an augmented DPM. The method initially partitions all observations into distinct clusters. At each subsequent step, two clusters are merged such that the posterior mass of the resulting partition is largest. This process is repeated until only one cluster remains. Of the partitions considered during the merging process, that with greatest posterior mass is taken as the MAP estimate. The Bayesian hierarchical clustering method (Heller and Ghahramani 2005; Xu et al. 2009) is another extension of the agglomerative method where the criterion for merging is based on a statistical hypothesis test rather than the largest increase in posterior mass. Fraley and Raftery (2002) recommend hierarchical agglomeration for initial model-based classification.

The Polya urn Gibbs sampler sequentially samples from the full conditional distributions of the cluster membership variables $\{z_1, \ldots, z_n\}$. However, the sequential nature of the Gibbs sampler makes it somewhat slow to mix in the space of partitions. The Metropolis-Hastings methods of Green and Richardson (2001), and Jain and Neal (2004, 2007) are designed to improve MCMC mixing by traversing the space of data partitions using 'split' and 'merge' operations.

Posterior sampling methods generate a consistent sequence of MAP estimates, but are somewhat wasteful in this context because they approximate the entire posterior distribution. In other words, an MCMC posterior sampling strategy will unnecessarily explore regions of lower density in order to satisfy a detailed balance with the posterior distribution. The stochastic method presented below utilizes the concept of 'split' and 'merge' operations to sequentially approximate the MAP estimate without the added complexity and computational expense of ensuring the chain is ergodic. However, care is taken to guarantee the MAP estimate is found in a finite number of iterations.

An iterative partition estimator may be initialized using the SUGS or agglomerative methods. However, the agglomerative method always requires $O(n^2)$ evaluations of the posterior mass function.

Updating the stochastic algorithm involves repeated 'explode' and 'merge' steps. At the 'explode' step, a subsample is selected uniformly at random from all observations. The subsample observations are then each distributed uniformly at random to an existing or new cluster. If the 'explode' step does not result in an increase in posterior probability, the subsample observations are 'merged' with one of the remaining clusters in a sequentially optimal manner. That is, a merge occurs where the resulting change in posterior probability is largest. If the 'merge' step does not result in an increase in posterior probability, the subsample observations are returned to their original clusters. The algorithm pseudocode at iteration t is

set $\mathbf{z}' = \mathbf{z}^{(t)}$
draw n_J from $\{1, \ldots, n\}$
draw vector J of length n_J from $\{1, \ldots, n\}$ w/o replacement
for j in J **do**
    draw z'_j from $\{1, \ldots, n\}$ (explode step)
end for
if $p(\mathbf{z}'|\mathbf{y}) > p(\mathbf{z}^{(t)}|\mathbf{y})$ **then**
    return $\mathbf{z}'$

end if
for j in J **do**
  set z'_j to maximize $p(z'_j|\mathbf{z}'_{-j}, \mathbf{y})$ (merge step)
end for
if $p(\mathbf{z}'|\mathbf{y}) > p(\mathbf{z}^{(t)}|\mathbf{y})$ **then**
  return $\mathbf{z}'$
else
  return $\mathbf{z}^{(t)}$
end if

where random draws are made uniformly from the appropriate support. The algorithm is terminated when a designated number of iterations has failed to improve the value of $p(\mathbf{z}^{(t)}|\mathbf{y})$. The 'explode' step of this update algorithm guarantees the MAP estimate will be obtained in a finite number of iterations (see Appendix 6.1). Hence, alternative 'merge' steps may be considered without loss of this property. In addition, this approach to computing the MAP estimate is amenable to simple but efficient schemes for parallel computing. In particular, mutually exclusive subspaces of the partition space may be searched in parallel.

The stochastic algorithm assumes the marginal posterior mass function $p(\mathbf{z}^{(t)}|\mathbf{y})$ is computable. However, computationally efficient use of the stochastic method requires an analytical solution for $p(\mathbf{z}^{(t)}|\mathbf{y})$. In practice, this requires conjugacy in the data likelihood parameter $\boldsymbol{\phi}$. Monte Carlo approximation of $p(\mathbf{z}^{(t)}|\mathbf{y})$ significantly reduces performance of the stochastic method versus other methods.

4 Illustration

The yeast cell cycle dataset was collected and analyzed by Spellman et al. (1998) in a time-series of microarray experiments with synchronized cultures of the baker's yeast *Saccharomyces cerevisiae*. Spellman *et al.* identified 800 'cell cycle-regulated' yeast genes. Each of these was then assigned to one of 5 cell cycle phases based on the similarity of its expression profile to that of other genes known to be associated with a particular cell cycle phase.

The yeast cell cycle data have been reanalyzed by Luan and Li (2003), Ng et al. (2006), and Ray and Mallick (2006) among others. In particular, Ray and Mallick used a Dirichlet process mixture of wavelet models, where the goal was to capture the

cluster structure in a manner similar to the original analysis of Spellman. They used posterior sampling methods to summarize the model posterior distribution and form a Monte Carlo estimate for the conditional likelihood of the observations, given the data partition. The data partition they present is that which maximizes the conditional likelihood estimate, rather than a MAP estimate.

The goal of the present analysis is not to reproduce the results of other analyses of these data, but to identify outliers in the clusters produced by the original analysis of Spellman et al. (1998). The results of this analysis might be useful for refining hypotheses concerning the gene regulatory mechanisms, or for discovering novel regulatory mechanisms common to a cell cycle phase.

For most cell cycle-regulated yeast genes, expression is periodic. In fact, Spellman et al. (1998) were primarily concerned with clustering the phases of expression profiles under a Fourier transform. Alternatively, the Dirichlet process mixture of linear models may be modified to accommodate a nonlinear or periodic profile by projecting the time covariate onto the $w-$dimensional space generated by a set of nonlinear and periodic basis functions. Notationally, let $\mathbf{\Phi}(x_j) = [\Phi_1(x_j), \Phi_2(x_j), \ldots, \Phi_w(x_j)]$ be the set of w linearly independent basis functions evaluated at time x_j. For the yeast cell cycle dataset, the DPM of these transformed linear models is given by

$$\begin{aligned} y_{ij} &\sim N(\mathbf{\Phi}(x_j)\boldsymbol{\beta}_i, \kappa_i) \\ (\boldsymbol{\beta}_i, \kappa_i) &\sim G \\ G &\sim DP(\alpha, G_0) \\ G_0 &= Ng_q(\mathbf{m}_0, s_0\mathbf{I}_q, a_0/2, 2/b_0), \end{aligned} \tag{8}$$

for $i = 1 \ldots n$ and $j = 1 \ldots q$, where y_{ij} is a gene expression value for the i^{th} gene collected at the j^{th} time point x_j in minutes, and Ng_q is the multivariate normal-gamma distribution (Bernardo and Smith 1994, pp. 118, 136, and 140). In order to capture the nonlinear and periodic nature of the expression profiles, the set of basis functions was selected from the power and sine functions. The first five power functions with non-negative integer exponents were selected to account for non-periodic trends in the expression profile. Five sine functions were selected so that their wavelengths were equal to the estimated cell cycle period (66 minutes) from Spellman et al. (1998), and whose phases were offset by one-fifth the range of the time covariate.

Let $\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r\}$ and $\{\tau_1, \ldots, \tau_r\}$ be the r unique values among $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n\}$ and $\{\kappa_1, \ldots, \kappa_n\}$ respectively. Conditional on the data partition variable $\mathbf{z}$ and the observations $\mathbf{y} = \{y_1, \ldots, y_n\}$, the unique pairs $\{(\boldsymbol{\gamma}_1, \tau_1), \ldots, (\boldsymbol{\gamma}_r, \tau_r)\}$ are independent

and multivariate normal-gamma distributed. The conditional posterior distribution for $(\boldsymbol{\gamma}_k, \tau_k)$ is given by

$$\boldsymbol{\gamma}_k, \tau_k | \mathbf{z}, \mathbf{y} \quad \sim \quad Ng_q(\mathbf{m}_k, \mathbf{S}_k, a_k/2, 2/b_k),$$

where $\mathbf{m}_k$ and $\tau_k \mathbf{S}_k$ are the mean vector and precision matrix of the multivariate normal component, and $a_k/2$ and $2/b_k$ are the shape and scale parameters for the gamma component. The posterior density function $p(\mathbf{z}|\mathbf{y})$ is proportional to

$$p(\mathbf{z}|\mathbf{y}) \propto \alpha^r \prod_{k=1}^{r} \frac{\Gamma(n^{(k)})\Gamma(a_k/2)}{(b_k/2)^{(a_k/2)}|\mathbf{S}_k|^{(1/2)}}.$$

The quantities $\mathbf{m}_k$, $\mathbf{S}_k$, $a_k/2$, and $2/b_k$ are posterior statistics of $\mathbf{z}$ and $\mathbf{y}$ given by

$$\begin{aligned} \mathbf{S}_k &= s_0 \mathbf{I}_q + \mathbf{X}^{(k)\prime}\mathbf{X}^{(k)} \\ \mathbf{m}_k &= \mathbf{S}_k^{-1}(s_0 \mathbf{m}_0 + \mathbf{X}^{(k)\prime}\mathbf{y}^{(k)}) \\ a_k &= a_0 + qn^{(k)} \\ b_k &= b_0 + \mathbf{y}^{(k)\prime}\mathbf{y}^{(k)} + s_0 \mathbf{m}_0'\mathbf{m}_0 - \mathbf{m}_k'\mathbf{S}_k\mathbf{m}_k, \end{aligned} \tag{9}$$

for $i = 1 \ldots n$ and $j = 1 \ldots q$, where $\mathbf{y}^{(k)}$ is the vector formed by concatenating all of the y_{ij} in the k^{th} cluster, $\mathbf{X}^{(k)}$ is the matrix formed by joining as rows all the $\mathbf{x}_{ij}$ of the k^{th} cluster, and $n^{(k)}$ is the number of observations assigned to the k^{th} cluster.

The explicit formulations in equation block (9) are a consequence of the conjugacy in the data likelihood and base prior G_0, and make clear the contribution of hyperparameters $\mathbf{m}_0$, s_0, a_0, and b_0. Clearly, the contribution of the prior is greater for clusters with few observations. In models where many clusters are possible, partition estimates may be highly sensitive to the value of hyperparameters. Within the confines of conjugacy, the relative contribution of G_0 may be reduced arbitrarily by setting each hyperparameter close to zero. There are no such obvious countermeasures for nonconjugate DPMs, and we recommend careful consideration of prior sensitivity in these cases.

The `R` (R Development Core Team 2011) package `profdpm` was used to analyze the gene expression profiles within each of the five original clusters produced by Spellman et al. (1998). The `R` code used to transform the time covariate is available in Appendix 6.2. The prior parameters were set as $a_0 = 0.001$, $b_0 = 0.001$, $\mathbf{m}_0 = [0, \ldots, 0]$, and $s_0 = 1$. The precision parameter α was set to $1/150$, which imposes the requirement of 'very strong' evidence in favor of a split operation that results in an outlier cluster (Efron and Gous 2001).

Fixing the value of α and n induces a prior distribution over the number of clusters r, with mean given by expression (2). For $\alpha = 1/150$, the prior expected number of clusters ranges from 1.03 in cell cycle phase S ($n = 69$), to 1.04 in phase G1 ($n = 297$). Drawing further from Antoniak (1974), we note that for fixed α and n, multiple clusters ($r > 1$) arise with probability

$$\begin{aligned} P(r > 1) &= 1 - P(r = 1) \\ &= 1 - \prod_{j=2}^{n} \left(1 - \frac{\alpha}{\alpha + j - 1}\right). \end{aligned}$$

For $\alpha = 1/150$, the prior probability that $r > 1$ ranges from 0.03 in phase S, to 0.04 in phase G1. Hence, requiring 'very strong' evidence to detect outlier clusters is sufficient for expressing prior belief that the number of clusters is concentrated near unity.

The agglomerative method (Heard et al. 2006; Ward 1963) was initially used to compute the MAP estimate for the data partition in each phase. The SUGS (Wang and Dunson 2011), SUGS++, Polya urn Gibbs sampler (MacEachern 1994), and stochastic methods detailed in Section 3 were each used independently to compute a MAP estimate. In the Gibbs approach, the MAP estimate was taken to be the sampled partition with largest posterior mass. The Gibbs and stochastic algorithms were each initialized using the agglomerative method, and evaluated for at most 50000 iterations. The SUGS method was applied in random observation order per the recommendation of Wang and Dunson (2011). The SUGS++ implementation consisted of a single SUGS evaluation, followed by 30 Polya urn Gibbs updates. An outlier cluster was taken to be any cluster consisting of three or fewer observations.

The finite mixture/BIC strategy was emulated by using the prior distribution over data partitions as described in section 2. Note that Fraley and Raftery (2002) recommend an expectation-maximization technique for model-based clustering. However, because the choice of optimization strategy may be confounding, the stochastic method proposed in Section 3 was used to compute the corresponding partition estimate.

For each phase, Table 1 lists the number of clusters, number of outlier observations, number of outlier clusters, unnormalized log posterior values for each MAP estimate, and the Rand (Rand 1971) index between the stochastic estimate and each other partition estimate. Larger Rand indices indicate greater agreement between the two partitions, where a Rand index of one indicates perfect agreement. The stochastic method yielded MAP estimates with uniformly greatest posterior mass. However, for most phases the Gibbs method also performs well.

Partitions estimated under the finite mixture/BIC strategy exhibit fair agreement with that estimated using the stochastic DPM method. However, the number of identified outliers is generally greater, illustrating that the evidence required to detect outliers is somewhat less than the 'very strong' requirement imposed by the DPM strategy.

The MAP estimation results illustrate a multi-way balance among utility, computational expense, accuracy, and stochasticity. For instance, the Gibbs method and other MCMC methods yield good posterior estimates and have high utility in applications beyond MAP estimation. However, the utility of MCMC methods comes with greater computational expense. The agglomerative and stochastic methods have less utility, but offer less computational burden. The agglomerative method is deterministic, but does not guarantee accurate MAP estimates. The SUGS method generates partition estimates quickly, but is not competitive with respect to log posterior mass. However, the SUGS++ results indicate that the SUGS strategy may be modified to accommodate MAP estimation with little or no additional computational burden. The fair agreement in Rand index between the SUGS and 'Stochastic' partitions may partially explain why the SUGS method yields reasonable predictions in nonparametric applications, despite having smaller posterior mass.

Figure 1 illustrates the MAP data partition for 297 genes regulated in the G1 phase. The MAP data partition for the G1 phase consisted of 12 outlier clusters accounting for 19 outlier observations. In order to illustrate inference using the outlier detection criterion presented in Section 2, consider the set of data partitions that might be formed by merging the outlier cluster in the upper righthand corner of Figure 1 with one of the remaining clusters. Specifying $\alpha = 1/150$ ensures that the Bayes factor for the MAP partition versus any such partition takes a value greater than 150. Hence, there is 'very strong' evidence for the decision to identify this cluster as an outlier.

Figure 2 summarizes the five MAP estimates by collapsing clusters according to the outlier cluster size rule (*i.e.* clusters having $n^{(k)} \leq 3$ are labeled outliers). Because distinct clusters have dissimilar mean expression profiles, collapsed observations in the 'non-outlier' and 'outlier' panels of Figure 2 may appear inhomogeneous. This reflects the possibility that multiple outlier and non-outlier clusters are present in each panel, and does not indicate poor sensitivity or specificity to detect outliers.

The distinction between outlier and non-outlier clusters is made on the basis of cluster size. That is, small clusters are considered outliers, for some notion of smallness in a particular data problem (*e.g.* n_k smaller than 1% of n). More importantly, when $\alpha = 1/150$, there is 'very strong' evidence that gene expression profiles assigned to dif-

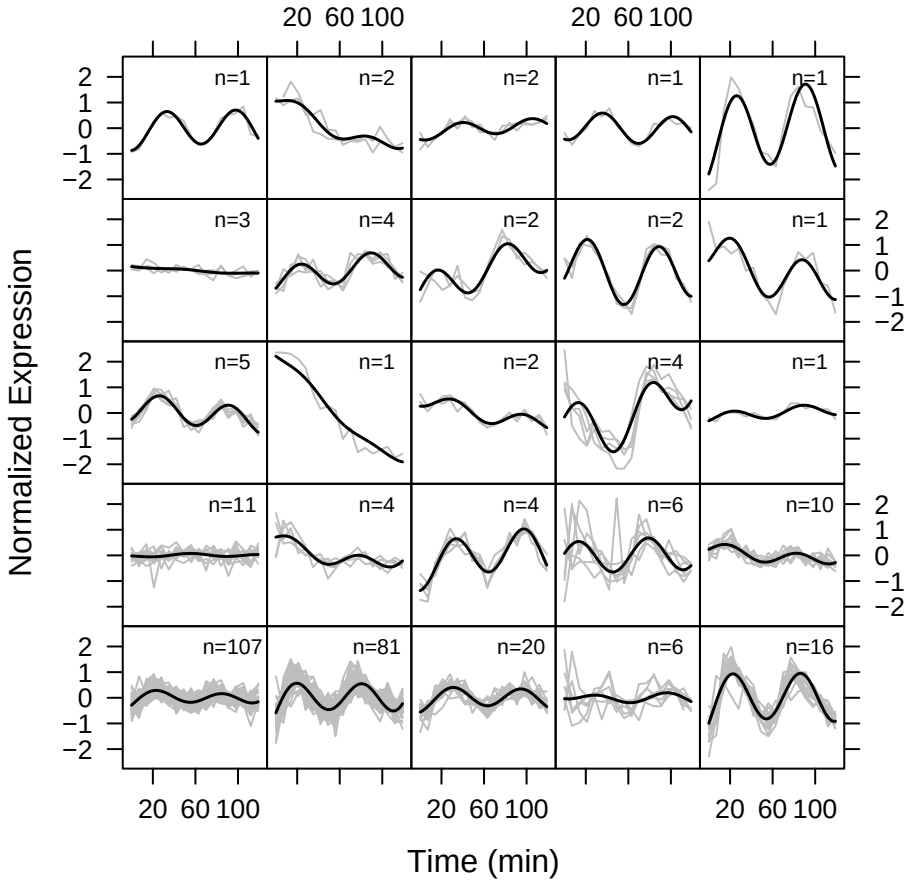

Figure 1: Gene expression profiles from the 'alpha' synchronized experiment for 297 genes regulated in the G1 phase of the cell cycle (Spellman et al. 1998). Each panel represents a cluster of genes identified by the MAP estimated data partition. Light gray lines give the expression profiles for each gene. Black lines give the profile posterior mean function for each cluster. The values of `n` are the number of expression profiles presented in each panel.

ferent clusters in the MAP estimate indeed arose from distinct processes. This property holds regardless of what size cluster is considered small.

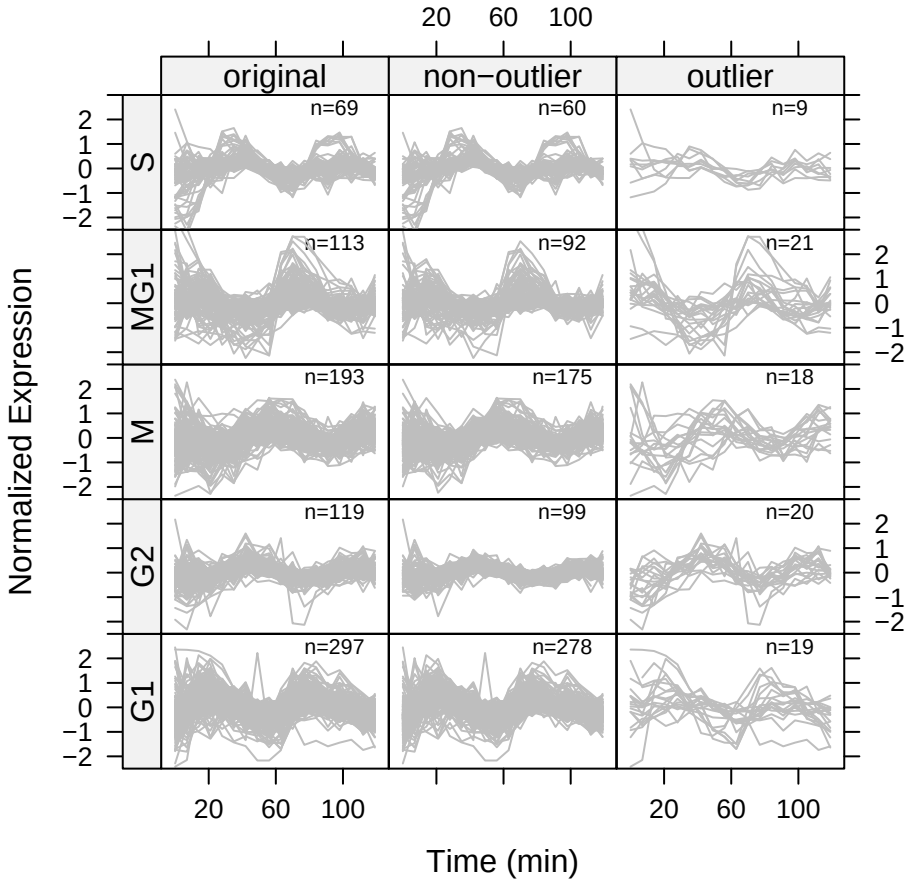

Figure 2: Gene expression profiles from the 'alpha' synchronized experiment for 791 'cell cycle-regulated' genes identified by Spellman et al. (1998). Light gray lines give the expression profiles for each gene. The values of `n` are the number of expression profiles presented in each panel. Panels in the leftmost column represent the original clusters of Spellman *et al.* The center and rightmost panels are collated from multiple distinct clusters of the estimated data partition according to the outlier cluster size rule ($n \leq 3$).

5 Conclusion

The outlier detection criterion of Section 2 offers an inference mechanism that evaluates the Bayes factor between an estimated (MAP) outlier partition, and a broad class of partitions formed by merging outlier clusters. However, the criterion makes no comparison with partitions outside this class. In particular, there is no guarantee that partitions formed by a sequence of merge *and* split operations will satisfy the criterion.

The stochastic MAP estimation algorithm of Section 3 avoids the complexity and expense of posterior sampling. As a result, this method has computational efficiency on par with popular non-Bayesian optimization methods. Still, more tailored strategies are welcome. The work of Dahl (2009) is a pioneering example, where the MAP estimate in a restricted class of augmented DPMs may be found in only $n(n+1)/2$ posterior evaluations. Adapting methods from the computer science literature, such as branching and bounding (Laursen 1993), and simulated annealing (Kirkpatrick et al. 1983) may also be fruitful. Because parallel computeing is gaining favor over serial computing, there is a growing need for computational strategies that exploit sophisticated parallelization in multicore and cluster computing environments. Wilkinson (2008) identifies concurrency and simplicity as important qualities in the future of statistical computing.

6 Appendix

6.1 Algorithm Consistency

Suppose $\{\mathbf{z}^{(t)}\}$ is a sequence of estimates for the data partition MAP estimator $\mathbf{z}_{\text{MAP}}$. Let q_t be the nonzero probability that $\mathbf{z}^{(t)}$ takes on the value $\mathbf{z}_{\text{MAP}}$. That is, with probability q_t, $p(\mathbf{z}^{(t)}|\mathbf{y}) \geq p(\mathbf{z}'|\mathbf{y})$ for all $\mathbf{z}'$. However, the algorithm of Section 3 stipulates that $p(\mathbf{z}^{(t)}|\mathbf{y}) \geq p(\mathbf{z}^{(t-1)}|\mathbf{y})$. This implies that $q_t \geq q_{t-1}$. Let T be the smallest value of t such that $\mathbf{z}^{(t)} = \mathbf{z}_{\text{MAP}}$. It is shown that $\lim_{T' \to \infty} p(T < T') = 1$. This probability is expressed

$$\begin{aligned} p(T < T') &= 1 - p(T \geq T') \\ &= 1 - p(T = T') - p(T > T') \\ &= 1 - q_{T'} \prod_{t=0}^{T'-1} (1 - q_t) - \prod_{t=0}^{T'} (1 - q_t). \end{aligned}$$

Since q_t is a nonzero probability and $q_t \geq q_{t-1}$ for all t, the limit of $p(T < T')$ as T' approaches infinity is one.

6.2 Covariate Transformation

The following `R` code was used to transform a time covariate `x` onto the space spanned by a collection of power and sine functions.

```
transform <- function( x, wl = 66/119 ) {
  cbind( rep(1, length(x)),
    x, x^2, x^3, x^4,
    sin(x/wl*2*pi),
    sin(x/wl*2*pi+pi/5),
    sin(x/wl*2*pi+2*pi/5),
    sin(x/wl*2*pi+3*pi/5),
    sin(x/wl*2*pi+4*pi/5) )
}
```

References

Al-Zoubi, M. (2009). "An Effective Clustering-Based Approach for Outlier Detection." *European Journal of Scientific Research*, 28: 310–316.

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *The Annals of Statistics*, 2: 1152–1174.

Bayarri, M. J. and Morales, J. (2003). "Bayesian measures of surprise for outlier detection." *Journal of Statistical Planning and Inference*, 111(1-2): 3 – 22.

Bell, E. T. (1934). "Exponential Numbers." *The American Mathematical Monthly*, 41: 411–419.

Ben-Gal, I. (2005). "Outlier Detection." In Maimon, O. and Rockach, L. (eds.), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*". Kulwer Academic Publishers.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York Inc., 2 edition.

Bernardo, J. and Smith, A. (1994). *Bayesian theory*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions Via Polya Urn Schemes." *The Annals of Statistics*, 1(2): 353–355.

Blei, D. and Jordan, M. (2006). "Variational Inference for Dirichlet Process Mixtures." *Bayesian Analysis*, 1(1): 121–144.

Box, G. E. P. and Tiao, G. C. (1968). "A Bayesian Approach to Some Outlier Problems." *Biometrika*, 55: 119–129.

Bush, C. A., Lee, J., and MacEachern, S. N. (2010). "Minimally informative prior distributions for non-parametric Bayesian analysis." *Journal of the Royal Statistical Society, Series B*, 72: 253–268.

Bush, C. A. and MacEachern, S. N. (1996). "A Semiparametric Bayesian Model for Randomised Block Designs." *Biometrika*, 83: 275–285.

Dahl, D. B. (2009). "Modal Clustering in a Class of Product Partition Models." *Bayesian Analysis*, 4: 243–264.

Efron, B. and Gous, A. (2001). "Scales of Evidence for Model Selection: Fisher versus Jeffreys." In Lahiri, P. (ed.), *Model Selection: Lecture Notes–Monograph Series*, volume 38, 208–246. Institute of Mathematical Statistics, Beachwood, OH.

Escobar, M. D. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90(430): 577–588.

Ferguson, T. S. (1961). "Rules for Rejection of Outliers." *Review of the International Statistical Institute*, 29(3): 29–43.

— (1973). "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1(2): 209–230.

Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Descriminant Analysis, and Density Estimation." *Journal of the American Statistical Association*, 97: 611–631.

Geisser, S. (1980). "Discussion of 'Sampling and Bayes Inference in Scientific Modeling and Robustness', by G.E.P. Box." *Journal of the Royal Statistical Society, Ser. A*, 143: 416–417.

Green, P. J. and Richardson, S. (2001). "Modelling Heterogeneity With and Without the Dirichlet Process." *Scandinavian Journal of Statistics*, 28: 355–375.

Hartigan, J. A. (1990). "Partition Models." *Communications in Statistics, Theory and Methods*, 19: 2745–2756.

Hautamaki, V., Karkkainen, I., and Franti, P. (2004). "Outlier Detection Using K-nearest Neighbor Graph." In *Proceedings of the 17th International Conference on Pattern Recognition*, 430–433.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). "A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves." *Journal of the American Statistical Association*, 101(473): 18–28.

Heller, K. A. and Ghahramani, Z. (2005). "Bayesian Hierarchical Clustering." In *Twenty-second International Conference on Machine Learning*.

Hoeting, J., Raftery, A. E., and Madigan, D. (1996). "A method for simultaneous variable selection and outlier identification in linear regression." *Computational Statistics & Data Analysis*, 22(3): 251 – 270.

Hogg, R. V., McKean, J. W., and Craig, A. T. (2005). *Introduction to Mathematical Statistics*. Pearson Prentice Hall, 6 edition.

Ishwaran, H. and James, L. F. (2001). "Gibbs Sampling Methods for Stick-breaking Priors." *Journal of the American Statistical Association*, 96: 161–174.

Jain, S. and Neal, R. M. (2004). "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *Journal of Computational and Graphical Statistics*, 13(1): 158–182.

— (2007). "Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model." *Bayesian Analysis*, 2(3): 445–472.

Jeffreys, H. (1961). *Theory of Probability*. Oxford, UK: Oxford University Press, 3 edition.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). "Optimization by Simulated Annealing." *Science*, 220: 671–680.

Lau, J. W. and Green, P. J. (2007). "Bayesian Model Based Clustering Procedures." *Journal of Computational and Graphical Statistics*, 16: 526–558.

Laursen, P. S. (1993). "Simple Approaches to Parallel Branch and Bound." *Parallel Computing*, 19: 143–152.

Luan, Y. and Li, H. (2003). "Clustering of Time-course Gene Expression Data Using a Mixed-effects Model with B-splines." *Bioinformatics*, 19: 474–482.

MacEachern, S. N. (1994). "Estimating Normal Means with a Conjugate Style Dirichlet Process Prior." *Communications in Statistics B*, 23: 727–741.

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999). "Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation." *Canadian Journal of Statistics*, 27: 251–267.

MacEachern, S. N. and Guha, S. (2011). "Parametric and semiparametric hypotheses in the linear model." *The Canadian Journal of Statistics*, 39: 165–180.

MacEachern, S. N. and Müller, P. (1998). "Estimating Mixture of Dirichlet Process Models." *Journal of Computational and Graphical Statistics*, 7: 223–238.

Newton, M. A. (2002). "On a Nonparametric Recursive Estimator of the Mixing Distribution." *Sankhyā, Series A*, 64: 306–322.

Ng, S., McLachlan, G., Wang, K., Ben-Tovim Jones, L., and Ng, S. (2006). "A Mixture Model with Random-effects Components for Clustering Correlated Gene-expression Profiles." *Bioinformatics*, 22: 1745–1752.

Petit, L. I. (1992). "Bayes Factors for Outlier Models using the Device of Imaginary Observations." *Journal of the American Statistical Association*, 87: 541–545.

Quintana, F. A. (2004). "A Predictive View of Bayesian Clustering." *Journal of Statistical Planning and Inference*, 136: 2407–2429.

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian Clustering and Product Partition Models." *Journal of the Royal Statistical Society, Series B*, 65: 557–574.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL http://www.R-project.org/

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, 66: 846–850.

Ray, S. and Mallick, B. (2006). "Functional Clustering by Bayesian Wavelet Methods." *Journal of the Royal Statistical Society, Series B*, 68: 305–332.

Rota, G. (1964). "The Number of Partitions of a Set." *The American Mathematical Monthly*, 71: 498–504.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization." *Molecular Biology of the Cell*, 9: 3273–3297.

Wang, L. and Dunson, D. B. (2011). "Fast Bayesian Inference in Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 20: 196–216.

Ward, J. H. (1963). "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association*, 58(301): 236–244.

Wilkinson, L. (2008). "The Future of Statistical Computing." *Technometrics*, 50(4): 418–435.

Xu, Y., Heller, K. A., and Ghahramani, Z. (2009). "Tree-Based Inference for Dirichlet Process Mixtures." In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 615–622.

Acknowledgments

The authors wish to thank the Editor, Associate Editor, and Referee for their thoughtful and productive comments. This research was partly funded by the following National Institutes of Health, National Institute of General Medical Sciences, and National Science Foundation funding projects: NIH 1T32GM074934, NIH R03CA137805, NSF DMS0604666, NIH P20RR017696.

Phase	Method	n	r	n_o	r_o	$\log p(\hat{\mathbf{z}}\|\mathbf{y})$	Rand
M	SUGS	193	163	187	162	2455	0.87
M	SUGS++	193	28	24	13	3056	0.88
M	Agglo	193	25	18	11	3075	0.97
M	Sampler	193	25	18	11	3075	0.97
M	Stochastic	193	26	18	12	3083	–
M	Finite/BIC	193	34	28	13	–	0.89
MG1	SUGS	113	96	105	94	1035	0.90
MG1	SUGS++	113	23	20	13	1355	0.93
MG1	Agglo	113	21	21	12	1366	0.98
MG1	Sampler	113	21	21	12	1366	0.98
MG1	Stochastic	113	21	21	12	1371	–
MG1	Finite/BIC	113	29	24	16	–	0.94
G1	SUGS	297	224	268	217	3206	0.79
G1	SUGS++	297	30	32	18	4570	0.88
G1	Agglo	297	26	27	15	4515	0.79
G1	Sampler	297	29	29	17	4600	0.89
G1	Stochastic	297	25	19	12	4643	–
G1	Finite/BIC	297	35	25	13	–	0.83
G2	SUGS	119	91	101	87	1837	0.80
G2	SUGS++	119	17	10	6	2131	0.84
G2	Agglo	119	18	20	12	2178	0.99
G2	Sampler	119	18	20	12	2178	0.99
G2	Stochastic	119	18	20	12	2179	–
G2	Finite/BIC	119	24	20	11	–	0.83
S	SUGS	69	58	69	58	904	0.85
S	SUGS++	69	13	7	6	1092	0.95
S	Agglo	69	14	12	9	1099	0.96
S	Sampler	69	13	8	6	1100	0.95
S	Stochastic	69	13	9	8	1102	–
S	Finite/BIC	69	21	16	11	–	0.90

Table 1: Optimization results by cell cycle phase and optimization/modeling method. The 'Agglo' method is the agglomerative method of Ward (1963). The 'Sampler' method is the Polya urn Gibbs sampler. The 'Stochastic' method is the method presented in Section 3. The values under n_o and r_o are the numbers of outlier observations and clusters respectively. The second column from the right contains the unnormalized log posterior mass at the estimate $\hat{\mathbf{z}}$ rounded to the nearest integer. The rightmost column holds the Rand index between the stochastic MAP estimate and each other partition estimate. Rows corresponding to the 'Finite/BIC' method describe the partition estimated using the model-based clustering method of Fraley and Raftery (2002).