Bayesian Analysis (2008) **3**, Number 1, pp. 101–120

How many clusters?

Peter McCullagh* and Jie Yang†

Abstract. The title poses a deceptively simple question that must be addressed by any statistical model or computational algorithm for the clustering of points. Two distinct interpretations are possible, one connected with the number of clusters in the sample and one with the number in the population. Under suitable conditions, these questions may have essentially the same answer, but it is logically possible for one answer to be finite and the other infinite. This paper reformulates the standard Dirichlet allocation model as a cluster process in such a way that these and related questions can be addressed directly. Our conclusion is that the data are sometimes informative for clustering points in the sample, but they seldom contain much information about parameters such as the number of clusters in the population.

Keywords: Cluster process; Dirichlet partition; Gauss-Ewens process; Random sub-clusters; Species-counting model

1 Gaussian mixtures

The basic problem of cluster analysis is to identify subsets or clusters in a finite set of points $y_1, \ldots, y_n$ in $\mathcal{R}^d$, with the idea that a cluster might plausibly represent an identifiable homogeneous sub-population. No external information in the form of covariates or relationships among the units is available to assist in the formation of clusters. One way to formulate this exercise as a statistical problem is to assume that the points $Y_1, Y_2, \ldots$ are independent and identically distributed with distribution f, which is a mixture of k Gaussian components

$$f(y) = \sum_{r=1}^{k} \pi_r \, \phi(y - \xi_r, \, \Sigma_0),$$

in which $\phi(y, \Sigma)$ is the Gaussian density at $y \in \mathcal{R}^d$ with covariance Σ. The mixture proportions are $\pi = \{\pi_1, \ldots, \pi_k\}$, and ξ_r is the mean of the rth component. For a good summary of finite mixture models, see Titterington, Smith, and Makov (1985), chapters 1 & 2. This paper considers only the simplest form of the mixture model in which each component has the same covariance matrix. However, the effect of variable cluster shape is achieved by the simple modification of the Dirichlet cluster process described in section 5.

Technically speaking π is an unordered set of non-negative numbers adding to one, and ξ is a parallel set of points in $\mathcal{R}^d$. Equivalently, the unordered set of ordered

*Department of Statistics, University of Chicago, Chicago, IL, http://galton.uchicago.edu/~pmcc
†Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL, http://www.math.uic.edu/~jyang06

pairs $\{(\pi_1, \xi_1), \ldots, (\pi_k, \xi_k)\}$, along with Σ_0, is sufficient to determine f. In practice, the elements of π are listed in some definite order, and the elements of ξ in the corresponding order. Since a simultaneous permutation of the components of π and ξ has no effect on the density f, it is evident that the individual components such as (ξ_1, π_1) or π_k are not identifiable. Lack of identifiability can be evaded but not entirely avoided by the imposition of order constraints on π or on a component of ξ (Richardson and Green 1997). One logical difficulty with constraints is that a component with low weight might not occur in the sample, which makes it difficult to match up the ordered sample values with the ordered components of the population parameter. Stephens (2000a, section 3) also argues against the imposition of constraints, but for more concrete reasons.

The Gaussian mixture model can be obtained from several different routes. One method is to begin with a list of labels $\{x_i\}$ chosen randomly and independently from a finite set of labels, and to assume that the observed values Y_i are independent Gaussian with mean $\xi(x_i)$ depending on the label (Scott and Symons 1971; Binder 1978; Banfield and Raftery 1993). The covariance matrix may also depend on the label, and the conditional distribution need not be Gaussian, but this level of generality is not used here.

Since the problem is unaffected by permutation of mixture components, it is natural to exploit this additional symmetry by using an exchangeable model for the mixture components, and most authors do so. However, it is desirable to go further by removing labels entirely (MacEachern 1994; Dahl 2005; McCullagh and Yang 2006; Booth, Casella, and Hobert 2007). We formulate the problem as an exchangeable cluster process in such a way that the mixture components occur as unlabelled blocks. Two problems arise in the Bayesian analysis of Gaussian mixtures, one conceptual connected with label-switching (Stephens 2000a,b), and one computational connected with the variable dimension of the parameter space (Richardson and Green 1997). The formulation as a cluster process rather than a mixture model avoids both problems at once without imposing constraints on the parameter space. The effect is to make the model simpler and the desired inferences more direct, at least in principle.

2 Cluster processes

2.1 Random partition

Consider a set $U = \{1, 2, 3, 4\}$ consisting of four units. A function or vector $x\colon U \to \{a, b, c\}$ with components (x_1, x_2, x_3, x_4) determines a partition of the units into three disjoint labelled classes. For example, if $x = (a, b, b, a)$, the classes are

$$x^{-1}(a) = \{1, 4\}, \quad x^{-1}(b) = \{2, 3\}, \quad x^{-1}(c) = \emptyset,$$

while the function $x' = (b, c, c, b)$ gives the same classes with permuted labels. All told, there are $3^4 = 81$ labelled partitions $x\colon U \to \{a, b, c\}$.

For certain purposes, it is more natural to focus on the partition, disregarding the

labels, and this is certainly true for cluster analysis problems in which the labelling of clusters is purely arbitrary. In the example shown above, the functions x and x' are regarded as equivalent because they induce the same unlabelled partition. For $n \geq 1$, a partition B of the set $[n] = \{1, , \ldots, n\}$ is a set of disjoint non-empty subsets, called blocks, whose union is $[n]$. The set $\mathcal{B}_n$ of partitions of $[n]$, called the partition lattice, arises naturally in connection with moments and cumulants (McCullagh 1984). For $n \leq 4$ the sets $\mathcal{B}_n$ are as follows

$$\begin{aligned}
\mathcal{B}_2&: \quad 12, \quad 1|2 \\
\mathcal{B}_3&: \quad 123, \quad 12|3\,[3], \quad 1|2|3 \\
\mathcal{B}_4&: \quad 1234, \quad 123|4\,[4], \quad 12|34\,[3], \quad 12|3|4\,[6], \quad 1|2|3|4
\end{aligned}$$

where $12|34$ is an abbreviation for the partition $\{\{1,2\},\{3,4\}\}$, and $12|34\,[3]$ is an abbreviation for the three partitions

$$12|34\,[3] = \{12|34,\ 13|24,\ 14|23\},$$

each having two blocks of size two. Thus $\mathcal{B}_3$ has 5 elements and $\mathcal{B}_4$ has 15.

Every function $x\colon [n] \to \mathcal{C}$ determines an equivalence relation $B\colon [n] \times [n] \to \{0,1\}$ by the label-forgetting transformation

$$B(i,j) = \begin{cases} 1 & if \mathrm{x(i)} = \mathrm{x(j)} \\ 0 & otherwise. \end{cases}$$

Note that x determines B, but not conversely. No distinction is made in the notation between B as an equivalence relation, B as a set of subsets, and B as a symmetric binary matrix. Thus $\#B$ is both the number of blocks and the rank of the matrix.

A permutation $\sigma\colon [n] \to [n]$ acts on partitions $B \mapsto B^\sigma$ in the obvious way by permuting rows and columns of the matrix $B^\sigma(i,j) = B(\sigma_i, \sigma_j)$. The number of blocks and the block sizes are unaffected. A probability distribution P_n on the set $\mathcal{B}_n$ is said to be symmetric if, for each permutation σ, $P_n(B^\sigma) = P_n(B)$ for all $B \in \mathcal{B}_n$. Symmetry implies that two partitions having the same block sizes also have the same probability.

To each partition $B' \in \mathcal{B}_{n+1}$ there corresponds a partition $B \in \mathcal{B}_n$ obtained by deleting the element $n+1$, i.e. by deleting the last row and column from the matrix B'. In this way, every distribution on $\mathcal{B}_{n+1}$ induces a marginal distribution on $\mathcal{B}_n$. A partition process is a sequence of distributions $\{P_n\}$ on $\mathcal{B}_n$ in which P_n is the marginal distribution of P_{n+1}, and an exchangeable partition process is one in which each distribution P_n is also invariant under permutation of units.

Examples of exchangeable partition processes are given in the next section. It suffices for the moment to observe that the distribution induced from the uniform distribution on $\mathcal{B}_3$ is not uniform on $\mathcal{B}_2$. The uniform distributions are symmetric for each n, but they do not determine a partition process. Another case is Hartigan's product partition model (Hartigan 1990; Crowley 1997; Quintana and Iglesias 2003) determined

by a cohesion function $c(\cdot)$ defined on subsets. The product partition model is infinitely exchangeable only if $c(b) = \lambda\Gamma(\#b)$ which leads to the Ewens family.

2.2 Dirichlet cluster process

As a model for cluster analysis, the Gaussian mixture formulation is a natural place to begin, but it is not entirely satisfactory because it fails to account for the symmetries that are usually present in clustering problems. For example, the labelling of clusters is unnecessary and in most respects undesirable. One way to avoid labels is to construct an exchangeable cluster process consisting of an infinite sequence $Y_1, Y_2, \ldots$ of points in $\mathcal{R}^d$, together with a random partition of the integers into k blocks. The simplest way to generate the leading sequence of length n from such a process is to select the value of k and proceed as follows.

1. Generate the cluster proportions $\pi = (\pi_1, \ldots, \pi_k)$ from the exchangeable Dirichlet distribution $\mathrm{Dir}(\lambda/k, \ldots, \lambda/k)$, where $\lambda > 0$.

2. Given π, generate the sequence of labels independently from the multinomial distribution with proportions π. For a set of n units, the probability of observing the label sequence $x = (x_1, \ldots, x_n)$ is $\pi_1^{n_1} \cdots \pi_k^{n_k}$, where $n_r \geq 0$ is the number of occurrences of label r. The unconditional probability is

$$P_n(x) = \frac{\Gamma(\lambda) \prod_r \Gamma(n_r + \lambda/k)}{\Gamma(n+\lambda)\,(\Gamma(\lambda/k))^k} \ .$$

3. Now forget the labels and let B be the random partition of $[n]$ induced by x. The distribution is

$$P_n(B; \lambda, k) = \frac{k!}{(k - \#B)!} \frac{\Gamma(\lambda) \prod_{b \in B} \Gamma(\#b + \lambda/k)}{\Gamma(n+\lambda)\,(\Gamma(\lambda/k))^{\#B}} \ . \tag{1}$$

   In this context, $\#B \leq k$ is the number of blocks in B, and for each block $b \in B$, the number of elements is $\#b \geq 1$.

4. For the same set of n units, the conditional distribution of $Y = (Y_1, \ldots, Y_n)$ given the sequence of n labels $x_1, \ldots, x_n$, depends only on the partition B of the given set of n units. The conditional distribution is Gaussian with constant mean vector $\mathbf{1}\mu$, and covariance matrix $\Sigma_B = I_n \otimes \Sigma_0 + B \otimes \Sigma_1$ whose components are

$$\mathrm{cov}(Y_{ir}, Y_{js} \,|\, B) = \delta_{ij}\Sigma_{0\,rs} + B_{ij}\Sigma_{1\,rs} \ ,$$

   where Σ_0, Σ_1 are the within- and between-cluster covariance matrices of order d. In other words, $Y(u) = \mu + \epsilon(u) + \xi(x(u))$ where ξ and ϵ are independent processes, both with independent Gaussian components. Consequently the joint distribution of (Y, B) is

$$p_n(y, B) = \phi(y - \mathbf{1}\mu,\, \Sigma_B) \times P_n(B; \lambda, k) \tag{2}$$

where $\phi(\cdot, \cdot)$ denotes the normal density in $\mathcal{R}^{nd}$.

5. For clustering problems in which only Y is observed, the marginal density at $y \in \mathcal{R}^{nd}$ is

$$p_n(y) = \sum_{B \in \mathcal{B}_n} \phi(y - \mathbf{1}\mu,\, I_n \otimes \Sigma_0 + B \otimes \Sigma_1)\; P_n(B; \lambda, k)\;. \tag{3}$$

The density (3) determines an exchangeable process and serves as the likelihood function for cluster analysis. In a partially supervised design where B is observed for some but not all units, the likelihood has an additional factor (2) for the supervised points. In practice, we often work with the marginal likelihood function (Tunnicliffe Wilson 1989; McCullagh 2008) based on the configuration statistic $(y - \mathbf{1}\overline{y})S^{-1/2}$, where y is the list of points arranged as a matrix of order $n \times d$, $\overline{y}$ is the mean vector in $\mathcal{R}^d$, and S is the sample covariance matrix. The main advantage is that the marginal distribution depends only on $(\Sigma_0^{-1}\Sigma_1, \lambda, k)$, and the conclusions are unaffected by affine transformation of points in $\mathcal{R}^d$. The marginal likelihood is effective in clustering problems where d is small relative to n, say $d < n/2$. Its effectiveness diminishes if $d > n/2$, and it is completely uninformative if $d \geq n$ (McCullagh 2008).

The parameter space for the cluster model (2) consists of the components $(k, \lambda, \mu, \Sigma_0, \Sigma_1)$, which is a union of manifolds, one for each positive integer k. Each of these manifolds has the same dimension regardless of k, so the problem of variable dimension does not arise. One minor complication arises due to the fact that the parameter is not identifiable: for $k = 1$ the distribution does not depend on λ. Otherwise the model is regular for $k \geq 2$, which is assumed where necessary.

Although λ is identifiable for $k \geq 2$, it is not consistently estimable in (1) unless $k = \infty$, and even then the rate of convergence is such that $\mathrm{var}(\widehat{\lambda}) = O(1/\log(n))$. If there was a compelling need to estimate λ accurately, this rate would be a serious drawback. However, the reason that the parameter is effectively unidentifiable is that its effect on distributions is slight, and this remark applies to both λ and k provided that k is not too small. Consequently the value has only a modest effect on conditional distributions. Consider for example, the partition B having five blocks of size 20. For $\lambda = 1$ the likelihood has a maximum at $k = 8$, but the ratio $P(B; 8)/P(B; \infty)$ is finite, in fact only 1.78. For certain purposes such as classification it may be sufficient to set $k = \infty$ and $\lambda = \#B/\log(n)$ if B is observed, leaving only $\Sigma_0^{-1}\Sigma_1$ to be estimated.

The partition distribution P_n in (1) depends only on the block sizes, so it is symmetric. In addition P_n is equal to the marginal distribution of P_{n+1}, so these distributions determine an exchangeable partition process. The limit as $k \to \infty$ is the Ewens process (Ewens 1972; Ishwaran and Zarepour 2002; Pitman 2006), also called the Chinese restaurant process (Aldous 1985; Pitman 2006, section 3.1). Likewise, each distribution p_n in (2) is invariant under coordinate permutation, $p_n(y^\sigma, B^\sigma) = p_n(y, B)$, so each distribution is symmetric. In addition, p_n is the marginal distribution of p_{n+1}, so these distributions determine an exchangeable cluster process.

For cluster analysis purposes, the Gauss-Ewens process is a special case of the Dirichlet process mixture models (MacEachern 1994; Neal 2000; Blei and Jordan 2006). The latter is an infinite mixture model using the Dirichlet process prior (Ferguson 1973; Antoniak 1974) for cluster centroids. In the Bayesian literature, most authors construct their hierarchical models using cluster labels. Exceptions include MacEachern (1994) who explicitly uses partitions with a conjugate style Dirichlet process prior, and Dahl (2005) who provides samplers for updating partitions with nonconjugate prior.

The first three steps of our construction are essentially the same as the model suggested by Fisher, Corbet, and Williams (1943) for estimating the number of species in a population, a model subsequently developed by Good and Toulmin (1956). Richardson and Green (1997) allow within-cluster covariance matrices to vary from cluster to cluster, but otherwise their construction follows the same lines and is formally equivalent for fixed k. Apart from our emphasis on the cluster process (2), and the distribution (3) as a model for the observations, there are other differences that have a substantial effect on conclusions. Richardson and Green use a parameterization in which $\delta = \lambda/k$ is held fixed, so the relation between their process for k and $k+1$ is different from ours. This difference is quite substantial, so much so that the partition model has a non-trivial limit as $k \to \infty$ for fixed λ, but there is no similar limit as $k \to \infty$ for fixed δ. For that reason, a Bayesian model in which (k, δ) are a priori independent may be very different from a model in which (k, λ) are independent.

3 Cluster analysis

3.1 Aims and objectives

The key idea is to use the family of Dirichlet cluster processes as a statistical model to address the sorts of questions posed in cluster analysis and related problems that are often addressed by Gaussian mixture models. In other words, given that $(y_1, \ldots, y_n)$ is observed from the marginal process with distribution (3), what can be said about the clusters? With a suitable prior distribution on the parameters $\theta = (k, \lambda, \mu, \Sigma_0, \Sigma_1)$, specific issues that may be addressed include the following.

1. Find the posterior distribution for k.
2. Find the posterior conditional distribution $p_n(B \mid y)$ for the clustering B of the sampled units.
3. Find the posterior conditional distribution for $\#B$, the number of clusters that occur among the sampled units.
4. Find the posterior conditional mean $E(B \mid y)$ for the sampled units.
5. Find the posterior modal clustering relative to a suitable baseline, either uniform or (1).
6. Predict the response value for a subsequent unit by computing the conditional density $p_{n+1}\bigl(y_{n+1} \mid y_1, \ldots, y_n\bigr)$ for the process (3).

If the Gauss-Ewens process is employed as a model, the answer to question 1 is $k = \infty$ with probability one, whereas the answer to question 3 is evidently finite. For large n, the unconditional distribution of $\#B$ implied by the Ewens process is approximately Poisson with parameter $\lambda \log(n)$, so the number of sample clusters increases rather slowly with the sample size (Arratia, Barbour, and Tavaré 2003, chapter 4; Pitman 2006, section 3.3).

In ecological applications, most authors make a strong distinction between the number of species in the population and the number that occur in a sample of individuals (Fisher, Corbet, and Williams 1943; Good and Toulmin 1956). However, few papers on mixture models and cluster analysis emphasize this distinction, or discuss which question is relevant for what purpose. For example, Tibshirani, Walther, and Hastie (2001) avoid formal models, so questions such as 1 or 6 cannot easily be addressed. Instead, they use a gap statistic for estimating 'the number of clusters in a set of data' making it clear that the gap statistic aims to answer question 3. Most proponents of formal models for cluster analysis appear to take a different view of the matter because question 3 is seldom considered. Banfield and Raftery (1993) use a Bayesian model in which the number of components is the number in the population, so their posterior distribution for k clearly addresses question 1. Similar remarks apply to Binder (1978) and to Richardson and Green (1997). Our experience is that undifferentiated data without class information can sometimes be mildly informative for question 3 and other matters related to the clustering of the sampled units. But question 1 is much more difficult. Even with the advantage of strong parametric assumptions embedded in the Dirichlet cluster process, the data seldom contain much information to address the matter.

The emphasis on question 1 over question 3 is defensible if k is small relative to n, and the model is such that $n \min\{\pi_r\}$ is large with high probability. This implies that the number of blocks in the sample is small, and the smallest block contains an appreciable number of units. But the Dirichlet allocation scheme does not guarantee this, so there could be numerous small blocks. In specific applications, it may be feasible to set a finite upper bound on the number of clusters based on physical or biological considerations, and it is then reasonable to restrict attention to prior distributions such that $\mathrm{pr}(k < \infty) = 1$. But in general, if there is substantial uncertainty about the number of clusters, it is mathematically more natural to allocate non-zero prior mass to the event that k is very large. Fisher, Corbet, and Williams (1943, pg. 54) favours $k = \infty$ for entomological applications, and the same assumption is widely used in connection with alleles in population genetics (Ewens 1972; Kingman 1978).

For the cluster model (2), the difference between the number of clusters in the population and the number that occur in a large sample is typically rather large. For the model considered by Richardson and Green (1997), the difference is not entirely negligible even for fairly large samples unless $\delta = \lambda/k$ is large. For example, if $k = 10$, the expected number of clusters occurring in a sample of size $n = 200$ is around 4.4 if $\lambda = 1$, and around 9.6 if $\lambda/k = 1$. Even if the cluster membership information is available for the sampled units, it is often difficult to say much about k other than $k \geq \#B$, without knowing λ.

Although the clusters are unlabelled, it may sometimes be necessary to make inferences about the mean of the cluster that contains a specific unit. Questions of this sort are best addressed directly in the following manner without recourse to labels. To each sample unit u there corresponds a block $b(u) = \{u' : B(u, u') = 1\}$ consisting of all units in the population belonging to the same block. The Dirichlet cluster model implies that $Y(u) = \epsilon(u) + \xi(b(u))$ is the sum of two independent Gaussian processes, each with independent and identically distributed components. In principle, the block mean $\xi(b(u)) = E(Y(u') \,|\, B(u, u') = 1)$ can be estimated from the data by weighting each sample unit u' in proportion to the estimate of $B(u, u')$. However, it would be naive to think that the block mean can be estimated accurately unless that particular block is well separated from the others that occur in the sample.

3.2 Identification of clusters

We consider in this section the problem of identifying clusters in a given sample. For this purpose, we suppose that the points in $\mathcal{R}^d$ are in fact generated independently from two normal populations, both with covariance matrix I_d. The two samples determine the true partition B^* having two blocks of equal size, one with mean $\Delta/2$ the other with mean $-\Delta/2$. The true partition is not observed, but we look to the conditional distribution $p_n(B \,|\, y)$ to see whether B^* has appreciable conditional probability. Even if n is large, we should not expect B^* to be the modal partition, but we might expect it and nearby partitions to have greater probability than the one-block partition. In the Dirichlet cluster model (2) we proceed as if $\Sigma_0 = I_d$, and $\Sigma_1 = \theta I_d$ with θ arbitrary but known.

For any partition B, the weighted sum of squares for blocks is

$$S_B^2 = \sum_{b \in B} \frac{n_b^2 \theta |\overline{y}_b|^2}{1 + n_b \theta}$$

where $\overline{y}_b \in \mathcal{R}^d$ is the block mean, n_b is the block size, and $|\overline{y}_b|$ is the usual norm in $\mathcal{R}^d$. Thus $S_1^2 = n^2 \theta |\overline{y}|^2/(1 + n\theta)$ for the one-block partition, and $S_B^2 - S_1^2$ is approximately the conventional between-blocks sum of squares when $d = 1$. The Gaussian density with covariance matrix $I_n + \theta B$ can be simplified so that the marginal density (3) of the observations at $y \in \mathcal{R}^{nd}$ satisfies

$$p_n(y; \lambda, k) = \phi_{nd}(y; 0, 1) \times \sum_{B \in \mathcal{B}_n} \frac{e^{S_B^2/2}}{\prod_{b \in B} (1 + n_b \theta)^{d/2}} P_n(B; \lambda, k). \tag{4}$$

The factor $(1 + n_b\theta)^{d/2}$ comes from the determinant of the covariance matrix, and $\phi_{nd}(y; 0, 1)$ is the spherical Gaussian density. In other words, the likelihood for (λ, k) is a linear combination of Dirichlet partition probabilities $P_n(B; \lambda, k)$ with coefficients depending on the weighted sum of squares for blocks.

The conditional distribution on sample partitions

$$p_n(B \,|\, y) \propto e^{S_B^2/2} \, \frac{k!}{(k - \#B)!} \prod_{b \in B} \frac{\Gamma(n_b + \lambda/k)}{\Gamma(\lambda/k)\,(1 + n_b\theta)^{d/2}} \tag{5}$$

is governed partly by the Dirichlet distribution (1) and partly by the between-blocks sum of squares. For $k = \infty$, the one-block partition has conditional probability proportional to $e^{S_1^2/2}\lambda\Gamma(n)/(1 + n\theta)^{d/2}$ and a two-block partition has conditional probability proportional to

$$\frac{\lambda^2\,\Gamma(n_1)\Gamma(n_2)\,e^{S_B^2/2}}{(1 + n_1\theta)^{d/2}(1 + n_2\theta)^{d/2}} \,.$$

Thus, the conditional probability of B exceeds that of the one-block partition if the between-blocks sum of squares is sufficiently large, i.e. if

$$e^{(S_B^2 - S_1^2)/2} \geq \frac{\Gamma(n)\,(1 + n_1\theta)^{d/2}(1 + n_2\theta)^{d/2}}{\lambda\Gamma(n_1)\Gamma(n_2)\,(1 + n\theta)^{d/2}} \,.$$

If $n_1 = n_2 = n/2$ are both large, this condition is satisfied if

$$e^{(S_B^2 - S_1^2)/2} \geq \frac{n^{(d+1)/2}\,2^{n-d-1}\,\theta}{\lambda\sqrt{2\pi}}$$

or $S_B^2 - S_1^2 > 2n\log(2) + (d+1)\log(n) + O(1)$. For a three-block partition with blocks of equal size, the critical value is $S_B^2 - S_1^2 > 2n\log(3) + 2(d+1)\log(n)$.

For the true partition B^*, the between-blocks sum of squares is $n|\Delta|^2/4 + O_p(1)$, so $p_n(B^* \,|\, y) \geq p_n(1 \,|\, y)$ if $|\Delta|^2 > 8\log(2)$ or $|\Delta| > 2.355$ regardless of the parameters. Note that the mixture density is bimodal if $|\Delta| > 2$ (Helguero 1904; Konstantellos 1980), so bimodality is not enough to guarantee that the two-block partition B^* has greater posterior probability than the one-block partition. The ratio $p_n(B^* \,|\, y)/p_n(1 \,|\, y)$ increases with n if $|\Delta|$ exceeds the critical value 2.355; otherwise it decreases. Even if $|\Delta|$ exceeds the critical value, B^* is usually not the modal partition. Accordingly, consistent identification of clusters is not feasible unless the clusters are well separated. Even then ambiguous points are inevitable. In this respect, the problem of cluster identification is fundamentally different from the problem of distribution estimation in a finite-dimensional Gaussian mixture model because the mixture model does not determine the clusters.

A more realistic target allows a small fraction of points to remain unclassified, recognizing that any point roughly equi-distant from two cluster centers cannot be assigned with certainty to either cluster. For $n - m$ points unambiguously classified into two blocks, and the remaining m assigned to one or other block, there are 2^m partitions to be considered, all having roughly the same value of S_B^2. The total probability of this set exceeds that of the one-block partition if $|\Delta|^2 > 8(1 - m/n)\log(2)$, so the conclusion is not greatly affected.

These calculations are based on the assumption that both covariance matrices are known. In the more realistic model with these as unknown parameters, the posterior conditional distribution gives a more honest assessment of the information available about clusters in the sample. Although consistent identification of clusters is clearly a hopeless task, the conditional distribution is sometimes quite informative, depending on the configuration of points. In some cases most elements of the matrix $E(B \,|\, y)$ are close to either zero or one, so the status of most pairs is well determined.

If $\Delta = 0$, the observations come from a single cluster with distribution $N(0, I_d)$, S_B^2 is a weighted sum of χ_d^2 random variables with weights $n_b\theta/(1 + n_b\theta)$ strictly less than one, and $E(\exp(S_B^2/2)) = \prod_{b\in B}(1 + n_b\theta)^{d/2}$ for each fixed partition B. The quadratic form $S_B^2(y)$ is not a symmetric function of y, so $S_B^2(y) \neq S_B^2(\sigma y)$ although they have the same expectation. Averaging over permutations suggests the approximation

$$\operatorname*{ave}_{\sigma} e^{S_B^2(\sigma y)} \simeq \prod_{b\in B}(1 + n_b\theta)^{d/2}$$

for large n. The accuracy of this approximation deteriorates as $\theta \to \infty$. Using this approximation in (5), we find that the conditional distribution of the block sizes is approximately equal to the unconditional distribution, i.e. the distribution on integer partitions implied by (1):

$$Q_n(1^{m_1}2^{m_2}\cdots n^{m_n}; \lambda, k) = P_n(B; \lambda, k) \times \frac{n!}{\prod_{j=1}^{n}(j!)^{m_j}\, m_j!}$$

where B is any partition having m_1 blocks of size one, m_2 blocks of size two, and so on. For a large sample from a homogeneous population, this calculation implies that the conditional distribution of the number of sample clusters does not converge to one as might have been expected. Instead, the conditional probability of the one-block partition is approximately $Q_n(n^1) \simeq n^{-\lambda(1-1/k)}$, i.e. negligible for large n. This large-sample theoretical calculation ignores the normalizing constant in (5). However, the conclusions have been confirmed by simulation, and the phenomenon persists for moderate values of Δ. This failure is not the result of a deficiency in the Dirichlet model; it is an honest recognition of the difficulty of the task.

3.3 Application to classification

Although the clusters in (2) are unlabelled, the model is simple and effective for classification or supervised learning in which (y, B) are both observed for the sampled units (Blei, Ng, and Jordan 2003). Any reasonable estimate of the parameters suffices, for example maximum likelihood or residual maximum likelihood. If there is appreciable uncertainty regarding k, an effective remedy is to set $k = \infty$ even if the number of classes is known to be finite. Suppose that (y, B) is observed on an initial set of n units, and that we wish to classify a subsequent out-of-sample unit u' with feature value $y(u')$. The conditional distribution $p_{n+1}(\cdot \,|\, \text{data})$ is determined by the probabilities assigned

to the events $u' \mapsto b$ for $b \in B$ and $b = \emptyset$:

$$p_{n+1}(u' \mapsto b \,|\, \text{data}) \propto \begin{cases} (\#b + \lambda/k)\, \phi_{n+1}(y' - \mathbf{1}\mu, \Sigma_{B_b}) & b \in B, \\ \lambda(1 - \#B/k)\, \phi_{n+1}(y' - \mathbf{1}\mu, \Sigma_{B_\emptyset}) & b = \emptyset, \end{cases}$$

where ϕ_{n+1} is the Gaussian density in $\mathcal{R}^{(n+1)d}$ and y' is the complete list of features. The notation B_b denotes the partition of order $n+1$ in which the observed partition B is the leading sub-matrix, and the last element belongs to block b.

If the matrices Σ_0, Σ_1 are proportional, the conditional distribution can be simplified using properties of the normal density. In that case, the probability assigned to the new class is small unless $y(u')$ is sufficiently far from the observed cluster means. Apart from a small shrinkage factor for the cluster means, the conditional probabilities are similar to those obtained from the classical Fisher discriminant model.

4 Numerical illustrations

4.1 Best-case scenario

The most optimistic scenario for estimating k is one in which the observed points fall into distinct clusters sufficiently well separated in $\mathcal{R}^d$ that the partition B can be determined with negligible error. In the calculations that follow, it is assumed that B is observed without error for the sampled units. The likelihood (2) has two factors, only one of which includes the target parameter k. Given B, the y-values are irrelevant for estimating k, and the likelihood function for (λ, k) is given by the the Dirichlet partition model (1). We aim to compute the posterior distribution for k under a range of assumptions about λ.

For numerical illustration we take $n = 100$ with two partitions into five blocks, the first uniform with five blocks of 20 points each, and the second with block sizes $\{50, 30, 15, 4, 1\}$. Figure 1 shows the contour plot of the log likelihood relative to the value at $(k = \infty, \lambda = 0.948)$. The log likelihood is plotted for two parameterizations $(\log k, \log \lambda)$ in the top row, and $(\log k, \log(\lambda/k))$ in the second row.

It is helpful for present purposes to distinguish between normal partitions whose block sizes are over-dispersed, and exceptional partitions whose block sizes are under-dispersed. Over-dispersion means that the sample variance of the block sizes exceeds the sample mean $n/\#B$. From a range of simulations using over-dispersed partitions it is invariably observed that the likelihood has an infinite ridge oriented horizontally or diagonally as shown in the right panels of Figure 1. A unique maximum occurs along the ridge, frequently at $k = \#B$ or at $k = \infty$ depending on the number and size of the small blocks. If the smallest block is sufficiently large, the profile likelihood decreases sharply from $k = \#B$, and is usually fairly flat over the remainder of the range. Over-dispersed partitions exist for which the likelihood has a maximum at an interior point, e.g. $\{60, 30, 5, 4, 1\}$, but the profile likelihood for k in such cases is usually flat over the entire range. For the uneven partition shown in Figure 1, the profile likelihood for k decreases monotonically, but the total decrease is less than one log likelihood

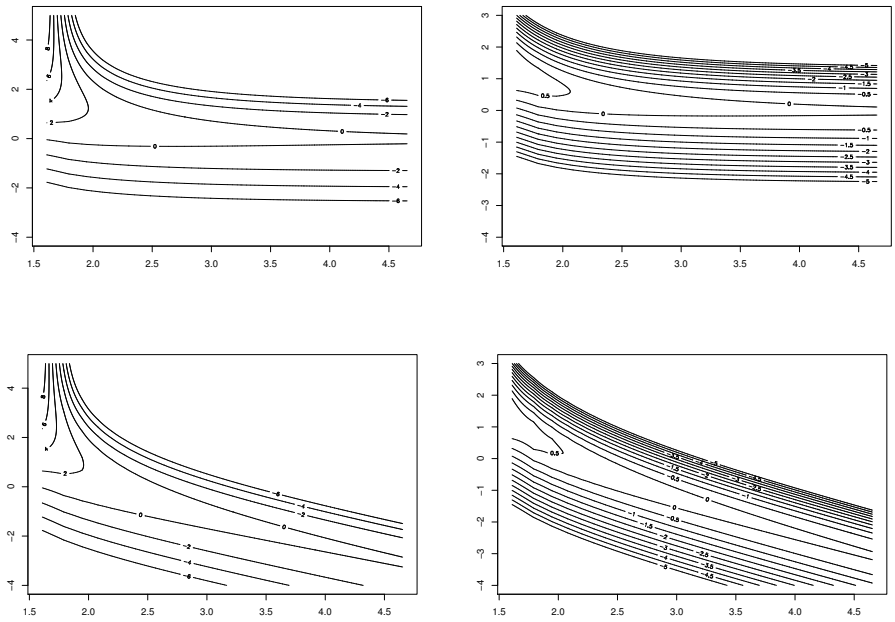

Figure 1: Contour plots of the log likelihood for two parameterizations $(\log k, \log \lambda)$ in the top row and $(\log k, \log(\lambda/k))$ in the bottom row. The configurations consist of 100 points in five blocks of equal size (left), and unequal sizes $\{50, 30, 15, 4, 1\}$ (right).

unit. The implication is that such sample partitions contain little information about the population parameter.

If the block sizes, considered as a set of size $k \geq \#B$ with $k - \#B$ zeros, are under-dispersed, the likelihood for fixed k has a maximum at $\widehat{\lambda}_k = \infty$. In such cases, the overall maximum usually occurs at $\widehat{k} = \#B$, in which case the profile likelihood for k decreases monotonically as illustrated in the left panel of Figure 1. But if the number of blocks is very large, e.g. 33 blocks of size 3, the maximum may occur at a finite value $\widehat{k} > \#B$. For the under-dispersed partition of $n = 10$ with one block of size four and six of size one, the maximum occurs at $\widehat{k} = \infty$, and the profile likelihood increases monotonically.

A pronounced infinite ridge in the likelihood function has a number of consequences for Bayesian inference. Consider first a prior distribution such that $\log \lambda$ is independent of k, say standard Cauchy. It is evident from the upper panels of Figure 1 that the marginal likelihood for k after integrating out λ is non-negligible as $k \to \infty$. If there are sufficiently many small blocks, the marginal likelihood is approximately constant for $k \geq \#B$. Consider now a second prior distribution such that $\delta = \lambda/k$ is independent of k. It is evident from the lower panels in Figure 1, that the marginal likelihood for k

after integrating out λ is such that large values of k have negligible marginal likelihood. Regardless of the observation, a large value of k having substantial prior probability has negligible posterior probability. In particular, the value $k = \infty$ has zero marginal likelihood whatever the observed partition. In the usual circumstance where the sample partition includes a few small blocks, the conclusion that k is finite is seldom supported by the likelihood function alone, but this conclusion is an inevitable consequence of the assumption of prior independence of δ and k. The difficulty cannot be evaded by the use of improper priors because the likelihood function for given finite k is not integrable.

All aspects of a stochastic model are arbitrary to some degree, and most compelling arguments are based on notions of symmetry whose relevance to the application must be gauged on a case-by-case basis. The arguments leading to (1) and (2) are based on exchangeability (permutation of units and irrelevance of block labels), so the model is reasonably firmly grounded in symmetry. In the absence of further symmetry arguments, it is difficult to make an equally compelling argument for one prior over another. However, it seems ill-advised to use a prior guaranteeing a conclusion that may not be supported by the likelihood. Since the Dirichlet partition process (1) has a non-degenerate limit for each λ as $k \to \infty$, this argument suggests that the conditional prior for λ given k should also have a non-degenerate limit. Prior information about the magnitude of k can be incorporated into the marginal prior where its effect is more readily apparent.

The problem of estimating k based on an observed partition B is formally equivalent to the classical problem of estimating the number of unseen species. The solution due to Fisher, Corbet, and Williams (1943) is essentially the Dirichlet partition model described in section 2. The set partition B induces a partition $1^{m_1}2^{m_2}, \ldots, n^{m_n}$ of the integer n in which m_r is the number of blocks of size r, and the integer partition is the sufficient statistic. The Dirichlet partition model implies that the expected frequencies decrease according to a negative binomial distribution. For a literary application, see Efron and Thisted (1976), who set out to estimate the number of words that Shakespeare knew based on the frequency of usage in the Shakespearean canon. The negative binomial model fits the observed frequencies exceptionally well, but even with $n \simeq 10^6$, the target parameter is extraordinarily difficult to estimate accurately and considerable ingenuity is required to obtain a finite estimate.

4.2 Counting sample clusters

Figure 2 shows four datasets of 60 points each from the Dirichlet cluster model with $\lambda = 1$, $\mu = 0$, $\Sigma_0 = I_2$, $\Sigma_1 = 9I_2$, and $k = 1, 2, 3, 4$. For illustrative purposes, these were selected so that the number of sample clusters is equal to k: the block sizes are $\{60\}$, $\{35, 25\}$, $\{35, 17, 8\}$, and $\{25, 19, 11, 5\}$.

We proceed as if B is not observed, aiming to infer B from the point configuration alone. For illustrative purposes, we assume that the true values of μ, and Σ_0 are known, and B is a Ewens partition with $\lambda = 1$. The choice of λ is not critical, but it is worth bearing in mind that two distinct units from the Ewens process belong to the same

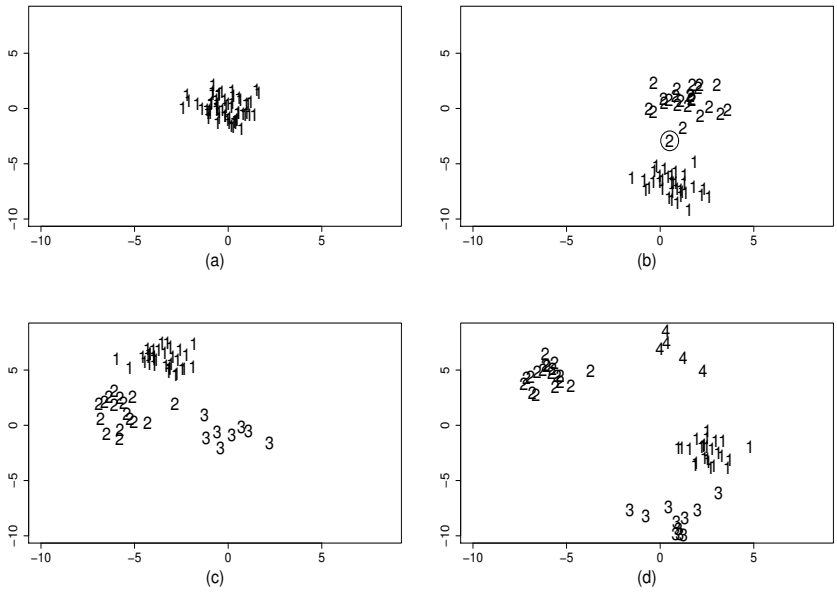

Figure 2: Simulated data sets of 60 points in 1–4 clusters. Points are simulated from the Dirichlet cluster model with $\lambda = 1$, $\mu = 0$, $\Sigma_0 = I_2$, $\Sigma_1 = 9I_2$ and $k = 1, 2, 3, 4$ as described in section 2.2.

cluster with probability $1/(\lambda+1)$. Finally, $\Sigma_1 = \theta\Sigma_0$ for some scalar θ with prior density $1/(1+\theta)^2$ chosen to be proper but minimally informative.

Table 1: $p_n(\#B|y) \times 1000$

$\#B$	1	2	3	4	5	6	7	8	9	10	11	$E(\#B\|y)$
Case(a)	18	74	158	217	213	157	92	44	18	6	2	4.76
Case(b)	0	87	336	327	175	58	15	3	0	0	0	3.84
Case(c)	0	0	213	366	272	113	30	5	1	0	0	4.40
Case(d)	0	0	0	376	392	177	46	7	1	0	0	4.92
Ewens	17	78	168	225	213	152	86	40	15	5	1	4.68

Table 2: Pairwise probabilities $p_n(i \sim j|y) \times 100$ for case (b)

$i\backslash j$	1	7	14	21	28	35	42	47	49	56	60
1	100	99	99	99	99	99	0	10	0	0	0
7	99	100	99	99	99	99	0	10	0	0	0
14	99	99	100	99	99	99	0	10	0	0	0
21	99	99	99	100	99	99	0	10	0	0	0
28	99	99	99	99	100	99	0	10	0	0	0
35	99	99	99	99	99	100	0	10	0	0	0
42	0	0	0	0	0	0	100	14	75	73	76
47	10	10	10	10	10	10	14	100	12	14	14
49	0	0	0	0	0	0	75	12	100	84	84
56	0	0	0	0	0	0	73	14	84	100	85
60	0	0	0	0	0	0	76	14	84	85	100

Markov chain Monte Carlo with Metropolis-Hastings updates (Hastings 1970; Neal 2000) was used to approximate the posterior conditional distribution $p_n(B \,|\, y)$, from which the marginal distribution $p_n(\#B \,|\, y)$ was obtained. Note that the state space of the Markov chain consists only of partitions of $\{1, 2, \ldots, 60\}$ as in the collapsed Gibbs sampler proposed by MacEachern (1994). In our case, the prior is not conjugate.

From the posterior distribution of B, we can compute the distribution of averages such as $E(B_{ij} \,|\, y) = p_n(i \sim j \,|\, y)$, the conditional probability that two units belong to the same cluster. Table 1 shows the results for the number of blocks based on the average of 5 independent chains. The small part of the matrix $E(B \,|\, y)$ shown in Table 2 demonstrates that the conditional distribution of B given y is very different from the unconditional distribution in which all off-diagonal elements are equal to $1/(\lambda+1)$.

For the homogeneous case (a), the posterior conditional distribution $p_n(\#B|y)$ is fairly close to the Ewens distribution shown in the last row of Table 1. This surprising result is explained by the argument in section 3.2. For the other configurations, the posterior distribution is quite different from the Ewens distribution, though it is not nearly so concentrated on the true value as might be expected. However, the posterior distribution establishes a clear minimum for the number of clusters in non-homogeneous configurations.

An alternative analysis uses the marginal likelihood based on the residual configura-

tion statistic or maximal invariant under affine transformation of points in $\mathcal{R}^2$, thereby avoiding the need for a prior on μ or Σ_0. This analysis is preferred because the conclusions are unaffected by affine transformation of the points in $\mathcal{R}^2$. However, qualitatively similar conclusions are obtained under the assumption that $\theta = \Sigma_0^{-1}\Sigma_1$ is a scalar with prior density $1/(1+\theta)^2$. The main difference is that for the two, three and four-cluster datasets, the posterior conditional distribution of the number of clusters is a little more diffuse in both tails.

If we change the prior for θ from the original $1/(1+\theta)^2$ with median 1 to $1/[9(1+\theta/9)^2]$ with median 9, the posterior conditional distribution for $\#B$ is not greatly affected. However, it would be a mistake to deduce that the conclusions are robust to the choice of prior. A prior that puts negligible mass on small values of θ, say zero for $\theta < 9$ and $2/[9(1+\theta/9)^2]$ for $\theta > 9$, implies that clusters are unlikely to have much overlap. For such a prior, the upper tail of the conditional distribution of $\#B$ is greatly reduced, and the conclusions are much tighter for all configurations. A sharply peaked posterior distribution for the number of sample clusters requires an informative prior.

To understand why $p_n(\#B = 3|y)$ is so much bigger than $p_n(\#B = 2|y)$ in case (b), we list part of the matrix $E(B\,|\,y)$ in Table 2. Point number 47 from cluster 2, which is circled in Figure 2(b), lies equi-distant between two clusters but, as indicated by the marginal posterior $p_n(47 \sim j\,|\,y)$, it is an outlier from both clusters. It could belong to either cluster, but it could equally plausibly belong to a new cluster. In fact, the true clustering B^* has less posterior probability than the three-block partition B' in which point 47 comprises a separate block. The ratio $p_n(B'\,|\,y)/p_n(B^*\,|\,y)$ is equal to 2.27, so the three-block partition is preferred. In large samples, this phenomenon is not uncommon.

5 Extensions

The Gaussian Dirichlet model (2) has the property that the clusters are geometrically congruent, all having the same within-cluster covariance matrix. If the application demands non-congruent clusters, the conventional modification is to associate with each cluster an independent random covariance matrix (Banfield and Raftery 1993; Richardson and Green 1997). A simpler solution is to formulate a model in which each cluster is a microcosm of the population, consisting of an independent random configuration of sub-clusters. The primary clusters are determined by a random partition B_1, and the sub-clusters by a random sub-partition $B_2 \leq B_1$ in which each block of B_2 is a subset of some block of B_1. For simplicity we consider the case $k = \infty$ in which the distribution of the primary clusters is

$$P_n(B_1;\lambda_1) = \frac{\lambda_1^{\#B_1}\Gamma(\lambda_1)}{\Gamma(n+\lambda_1)} \prod_{b\in B_1} \Gamma(\#b).$$

Given B_1, the distribution on sub-clusters is

$$P_n(B_2 \,|\, B_1, \lambda_2) = \lambda_2^{\#B_2} \prod_{b \in B_1} \frac{\Gamma(\lambda_2)}{\Gamma(\#b + \lambda_2)} \times \prod_{b' \in B_2} \Gamma(\#b').$$

In the population, i.e. in the limit as $n \to \infty$, each primary cluster has an infinite number of sub-clusters in a distinct random configuration. For finite n, it is possible that $B_2 = B_1$, in which case no primary cluster contains a proper sub-cluster. In any event, the larger the primary cluster the more likely it is to be split into proper sub-clusters.

The two-level Gaussian cluster process is such that the conditional distribution of Y given the pair B_1, B_2 is Gaussian with constant mean and covariance

$$\operatorname{cov}(Y_{ir}, Y_{js} \,|\, B_1, B_2) = \delta_{ij}\Sigma_{0\,rs} + B_{1\,ij}\Sigma_{1\,rs} + B_{2\,ij}\Sigma_{2\,rs}.$$

Variability between units in the same sub-cluster is determined by Σ_0, and between units in different sub-clusters of the same primary cluster by $\Sigma_0 + \Sigma_2$. Evidently, the sequence of clusters and sub-clusters can be extended indefinitely by recursive partitioning. Each of these processes is exchangeable.

6 Conclusions

In Bayesian calculations connected with the Dirichlet partition model, careful attention to the prior is required. Regardless of the marginal prior for the number of population clusters, a prior in which $k, \lambda/k$ are independent effectively guarantees the conclusion that k is not much larger than the number of sample clusters. It is not unreasonable in certain applications to expect that the difference between k and $\#B$ might be small, but there is ample evidence in other applications that the difference is sometimes large. It is best if the information in support of this conclusion comes primarily from the configuration of sample clusters in the data, not from a property of the prior distribution introduced for convenience of computation. On balance, a prior in which k, λ are independent seems preferable for inferences concerning k.

The variance ratio parameter $\theta = \Sigma_0^{-1}\Sigma_1$ is a critical component of the Dirichlet cluster process, and conclusions about the number and configuration of sample clusters can be substantially altered by changing the prior distribution. If the prior puts appreciable mass on small values, say $\theta < 4$, a sample configuration y that appears to be homogeneous has as much chance of occurring as the superposition of two or more coincident clusters as it does from a single cluster: $p_n(y \,|\, \#B = 1) \simeq p_n(y \,|\, \#B = 2)$. Accordingly, if θ is small with appreciable probability, a homogeneous configuration of points conveys little information about the number of sample clusters. If the model is to be used for counting sample clusters, this phenomenon is best avoided, and to do so the prior for θ must put negligible mass on small values.

References

Aldous, D. J. (1985). "Exchangeability and related topics." *École d'Été de Probabilités de Saint-Flour XIII*. Springer. 105

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to nonparametric problems." *The Annals of Statistics*, 2:1152–1174. 106

Arratia, R., Barbour, A. D., and Tavaré, S. (2003). *Logarithmic Combinatorial Structures: A Probabilistic Approach*. European Mathematical Society. 107

Banfield, J. D. and Raftery, A. E. (1993). "Model-based Gaussian and non-Gaussian clustering." *Biometrics*, 49:803–821. 102, 107, 116

Binder, D. A. (1978). "Bayesian cluster analysis." *Biometrika*, 65:31–38. 102, 107

Blei, D., Ng, A., and Jordan, M. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3:993–1022. 110

Blei, D. and Jordan, M. (2006). "Variational inference for Dirichlet process mixtures." *Bayesian Analysis*, 1:121–144. 106

Booth, J., Casella, G., and Hobert, J. (2007). "Clustering using objective functions and stochastic search." Technical Report, Department of Statistics, University of Florida. 102

Crowley, E. M. (1997). "Product partition models for normal means." *Journal of the American Statistical Association*, 92:192–198. 103

Dahl, D. B. (2005). "Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models." Technical Report, Department of Statistics, Texas A&M University. 102, 106

Efron, B. and Thisted, R. (1976). "Estimating the number of unseen species: how many words did Shakespeare know?" *Biometrika*, 63:435–447. 113

Ewens, W. J. (1972). "The sampling theory of selectively neutral alleles." *Theoretical Population Biology*, 3:87–112. 105, 107

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1:209–230. 106

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). "The relation between the number of species and the number of individuals in a random sample of an animal population." *The Journal of Animal Ecology*, 12:42–58. 106, 107, 113

Good, I. J. and Toulmin, G. H. (1956). "The number of new species, and the increase in population coverage, when a sample is increased." *Biometrika*, 43:45–63. 106, 107

Hartigan, J. A. (1990). "Partition models." *Communications in Statistics, Part A - Theory and Methods*, 19:2745-2756. 103

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57:97–109. 115

Helguero, F. (1904). "Sui massimi delle curve dimorfiche." *Biometrika*, 3:84–98. 109

Ishwaran, H. and Zarepour, M. (2002). "Dirichlet prior sieves in finite normal mixtures." *Statistica Sinica*, 12:941–963. 105

Kingman, J. F. C. (1978). "Random partitions in population genetics." *Proceedings of the Royal Society of London: Series A*, 361:1–20. 107

Konstantellos, A. C. (1980). "Unimodality conditions for Gaussian sums." *IEEE Transactions on Automatic Control*, AC-25:838–839. 109

MacEachern, S. N. (1994). "Estimating normal means with a conjugate-style Dirichlet process prior." *Communication in Statistics: Simulation and Computation*, 23:727–741. 102, 106, 115

McCullagh, P. (1984). "Tensor notation and cumulants of polynomials." *Biometrika*, 71:461–476. 103

McCullagh, P. (2008). "Marginal likelihood for parallel series." *Bernoulli*, (to appear). 105

McCullagh, P. and Yang, J. (2006). "Stochastic classification models." *Proceedings of the International Congress of Mathematicians (Madrid, 2006)*, III:669–686. 102

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9:249–265. 106, 115

Pitman, J. (2006). *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002*, J. Picard (ed.). Springer. 105, 107

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. 103

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components." *Journal of the Royal Statistical Society: Series B*, 59:731–792. 102, 106, 107, 116

Scott, A. J. and Symons, M. J. (1971). "Clustering methods based on likelihood ratio criteria." *Biometrics*, 27:387–397. 102

Stephens, M. (2000a). "Dealing with label-switching in mixture models." *Journal of the Royal Statistical Society: Series B*, 62:795–809. 102

— (2000b). "Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods." *The Annals of Statistics*, 28:40–74. 102

Tibshirani, R., Walther, G., and Hastie, T. (2001). "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society: Series B*, 63:411–423. 107

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons. 101

Tunnicliffe Wilson, G. (1989). "On the use of marginal likelihood in time series model estimation." *Journal of the Royal Statistical Society: Series B*, 51:15–27. 105

Acknowledgments

Support for this research was provided by NSF Grant DMS-0305009.