MIXTURE MODELS WITH A PRIOR ON THE NUMBER OF COMPONENTS

JEFFREY W. MILLER AND MATTHEW T. HARRISON

Abstract. A natural Bayesian approach for mixture models with an unknown number of components is to take the usual finite mixture model with Dirichlet weights, and put a prior on the number of components—that is, to use a mixture of finite mixtures (MFM). While inference in MFMs can be done with methods such as reversible jump Markov chain Monte Carlo, it is much more common to use Dirichlet process mixture (DPM) models because of the relative ease and generality with which DPM samplers can be applied. In this paper, we show that, in fact, many of the attractive mathematical properties of DPMs are also exhibited by MFMs—a simple exchangeable partition distribution, restaurant process, random measure representation, and in certain cases, a stick-breaking representation. Consequently, the powerful methods developed for inference in DPMs can be directly applied to MFMs as well. We illustrate with simulated and real data, including high-dimensional gene expression data.

1. Introduction

Mixture models are used in a wide range of applications, including population structure (Pritchard et al., 2000), document modeling (Blei et al., 2003), speaker recognition (Reynolds et al., 2000), computer vision (Stauffer and Grimson, 1999), phylogenetics (Pagel and Meade, 2004), and gene expression profiling (Yeung et al., 2001), to name a few prominent examples. A common issue with finite mixtures is that it can be difficult to choose an appropriate number of mixture components, and many methods have been proposed for making this choice (e.g., Henna, 1985; Keribin, 2000; Leroux, 1992; Ishwaran et al., 2001; James et al., 2001).

From a Bayesian perspective, perhaps the most natural approach is to treat the number of components like any other unknown parameter and put a prior on it. For short, we refer to such a model as a mixture of finite mixtures (MFM). Several inference methods have been proposed for this type of model (Nobile, 1994; Phillips and Smith, 1996; Richardson and Green, 1997; Stephens, 2000; Nobile and Fearnside, 2007), the most commonly-used method being reversible jump Markov chain Monte Carlo (Green, 1995; Richardson and Green, 1997). Reversible jump is a very general technique, and has been successfully applied in many contexts, but it is perceived to be difficult to use, and applying it to new situations requires one to design good reversible jump moves, which can be nontrivial, particularly in high-dimensional parameter spaces.

Meanwhile, infinite mixture models such as Dirichlet process mixtures (DPMs) have become popular, partly due to the existence of generic Markov chain Monte Carlo (MCMC) algorithms that can easily be adapted to new applications (Neal, 1992, 2000; MacEachern, 1994, 1998; MacEachern and Müller, 1998; Bush and MacEachern, 1996; West, 1992; West et al., 1994; Escobar and West, 1995; Liu, 1994; Dahl, 2003, 2005; Jain and Neal, 2004, 2007). These algorithms are made possible by the fact that the Dirichlet process has a variety of elegant mathematical properties—an exchangeable partition distribution, the

Blackwell–MacQueen urn process (a.k.a. the Chinese restaurant process), a random discrete measure formulation, and the Sethuraman–Tiwari stick-breaking representation (Ferguson, 1973; Antoniak, 1974; Blackwell and MacQueen, 1973; Aldous, 1985; Pitman, 1995, 1996; Sethuraman, 1994; Sethuraman and Tiwari, 1981).

The purpose of this paper is to show that in fact, MFMs typically exhibit many of these same appealing properties—an exchangeable partition distribution, urn/restaurant process, random discrete measure formulation, and in certain cases, a simple stick-breaking representation—and consequently, that many of the inference techniques developed for DPMs can be directly applied to MFMs. In particular, these properties enable one to do inference in MFMs without using reversible jump. Interestingly, the key properties of MFMs hold for any choice of prior distribution on the number of components.

There has been a large amount of research on efficient inference methods for DPMs, and an immediate consequence of the present work is that most of these methods can also be used for MFMs. Since many DPM sampling algorithms (for both conjugate and non-conjugate priors) are designed to have good mixing properties across a wide range of applications—for instance, the Jain–Neal split-merge samplers (Jain and Neal, 2004, 2007), coupled with incremental Gibbs moves (MacEachern, 1994; Neal, 1992, 2000)—this greatly simplifies the use of MFMs in new applications.

This work resolves an open problem discussed by Green and Richardson (2001), who noted that it would be interesting to be able to apply DPM samplers to MFMs:

> *"In view of the intimate correspondence between DP and [MFM] models discussed above, it is interesting to examine the possibilities of using either class of MCMC methods for the other model class. We have been unsuccessful in our search for incremental Gibbs samplers for the [MFM] models, but it turns out to be reasonably straightforward to implement reversible jump split/merge methods for DP models."*

The paper is organized as follows. In the remainder of this section, we motivate this work with an overview of the similarities and differences between MFMs and DPMs, illustrated by a simulation example. In Sections 2 and 3, we formally define the MFM and show that it gives rise to a simple exchangeable partition distribution closely paralleling that of the Dirichlet process. In Section 4, we present the Pólya urn scheme (restaurant process), random discrete measure formulation, and stick-breaking representation for the MFM. In Section 5, we establish some asymptotic results for MFMs. In Section 6, we show how the properties in Sections 3 and 4 lead to efficient inference algorithms for the MFM, and in Section 7, we apply the model to the galaxy dataset (a standard benchmark) and to high-dimensional gene expression data used to discriminate cancer subtypes. We close with a brief discussion.

1.1. Background: Similarities and differences between MFMs and DPMs.

In many respects, MFMs and DPMs are quite similar, but there are some important differences. In general, we would not say that one model is uniformly better than the other; rather, one should choose the model which is best suited to the application at hand.

Density estimation. For certain nonparametric density estimation problems, both models have been shown to exhibit posterior consistency at the minimax optimal rate, up to logarithmic factors (Kruijer et al., 2010; Ghosal and Van der Vaart, 2007). Even for small sample sizes, we observe empirically that density estimates under the two models are remarkably

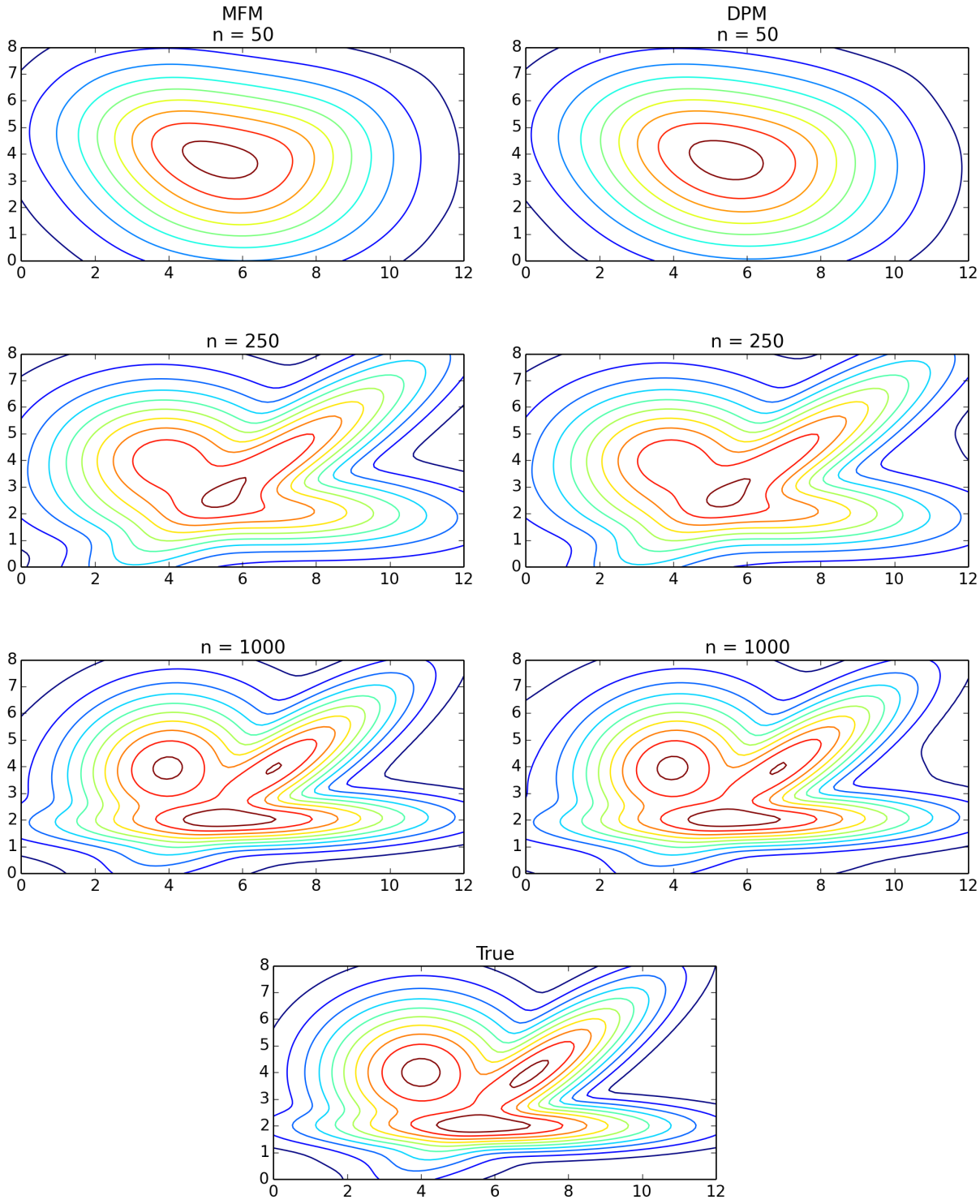

FIGURE 1. Density estimates for MFM (left) and DPM (right) on increasing amounts of data from a three-component Gaussian mixture (bottom). As n increases, the estimates appear to be converging to the true density, as expected. See Section 7.1 for details.

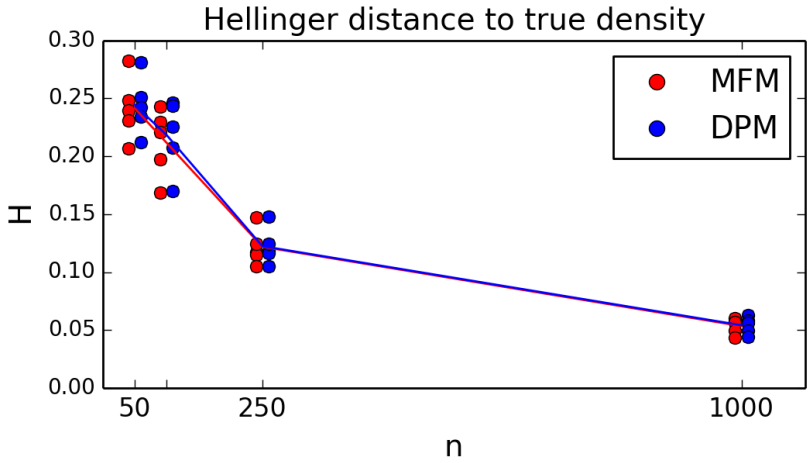

FIGURE 2. Hellinger distance to the true density, for MFM (red, left) and DPM (blue, right) density estimates, on data from a three-component Gaussian mixture. For each $n \in \{50, 100, 250, 1000\}$, five independent datasets of size n were used, and the lines connect the averages of the distances for each n. See Section 7.1 for details.

similar. As a toy example, Figure 1 compares their density estimates on data from a bivariate Gaussian mixture with three components. As the amount of data increases, these density estimates appear to be converging to the true density, as expected; indeed, Figure 2 indicates that the Hellinger distance to the true density is going to zero.

Clustering. It seems that more often, mixture models are used for clustering and latent class discovery rather than density estimation. MFMs have a partition distribution that takes a very similar form to that of the Dirichlet process; see Section 3. However, despite this similarity, the MFM partition distribution differs in two fundamental respects. While the first is widely known, the second is far less often appreciated, and yet is perhaps even more important.

(1) The prior on the number of clusters t is very different. In an MFM, one has complete control over the prior on the number of components k, which in turn provides control over the prior on t. As the sample size n grows, in an MFM the prior on t converges to the prior on k (in fact, t converges to k almost surely). In contrast, in a Dirichlet process, the prior on t takes a particular parametric form and diverges at a $\log n$ rate.
(2) Given t, the prior on the cluster sizes is very different. In an MFM, most of the prior mass is on partitions in which the sizes of the clusters are all the same order of magnitude, while in a Dirichlet process, most of the prior mass is on partitions in which the sizes vary widely, with a few large clusters and many very small clusters.

See Section 5 for more precise descriptions of (1) and (2) in mathematical terms. (Note that while some authors use the terms "cluster" and "component" interchangeably, we use *cluster* to refer to a group of data points, and *component* to refer to one of the probability distributions in a mixture model.)

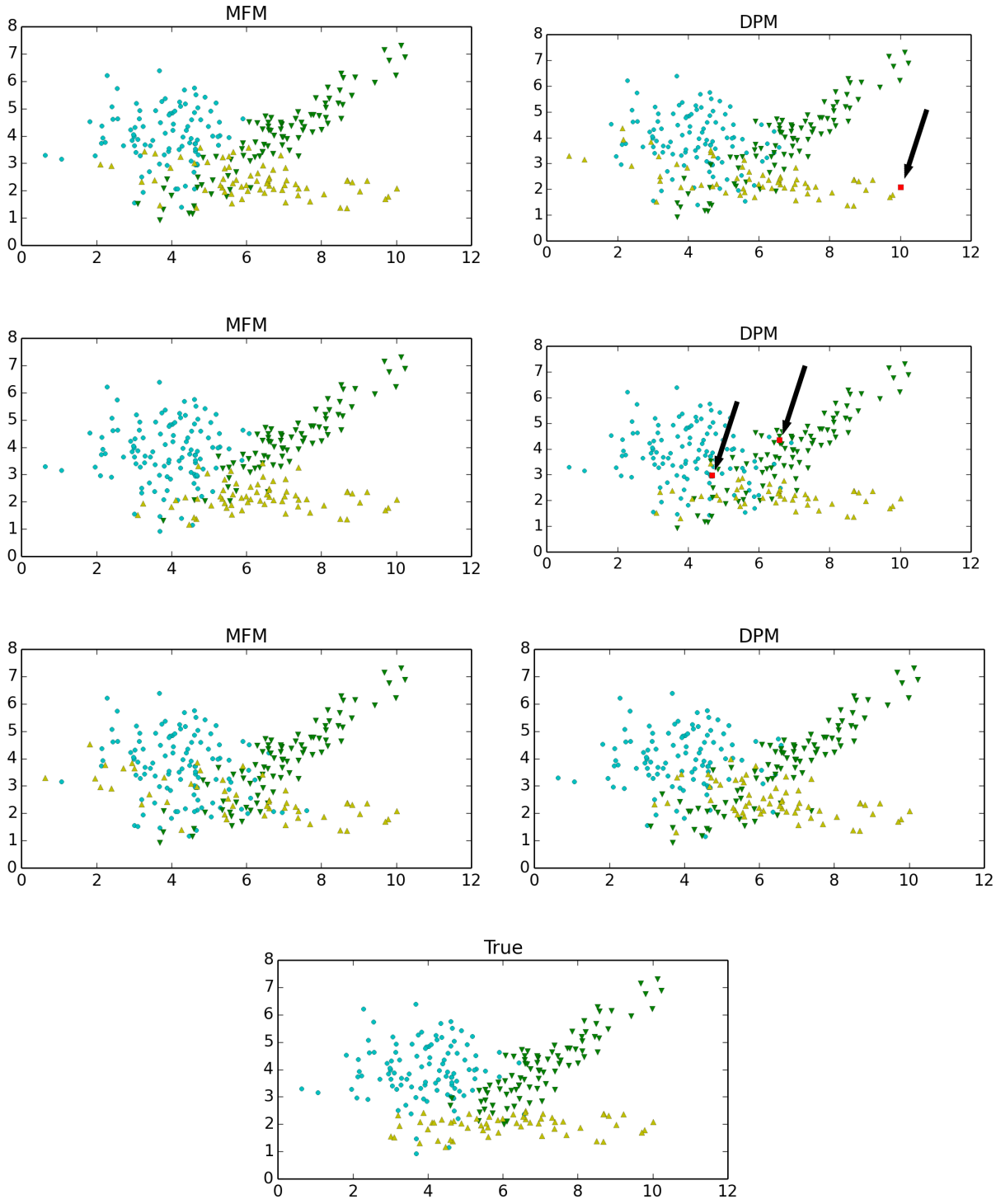

FIGURE 3. Typical sample clusterings from the posterior for the MFM (left) and DPM (right), on $n = 250$ data points from a three-component Gaussian mixture; the bottom plot shows the true component assignments. Note the small extra clusters in the DPM samples (red squares). Best viewed in color. See Section 7.1 for details.

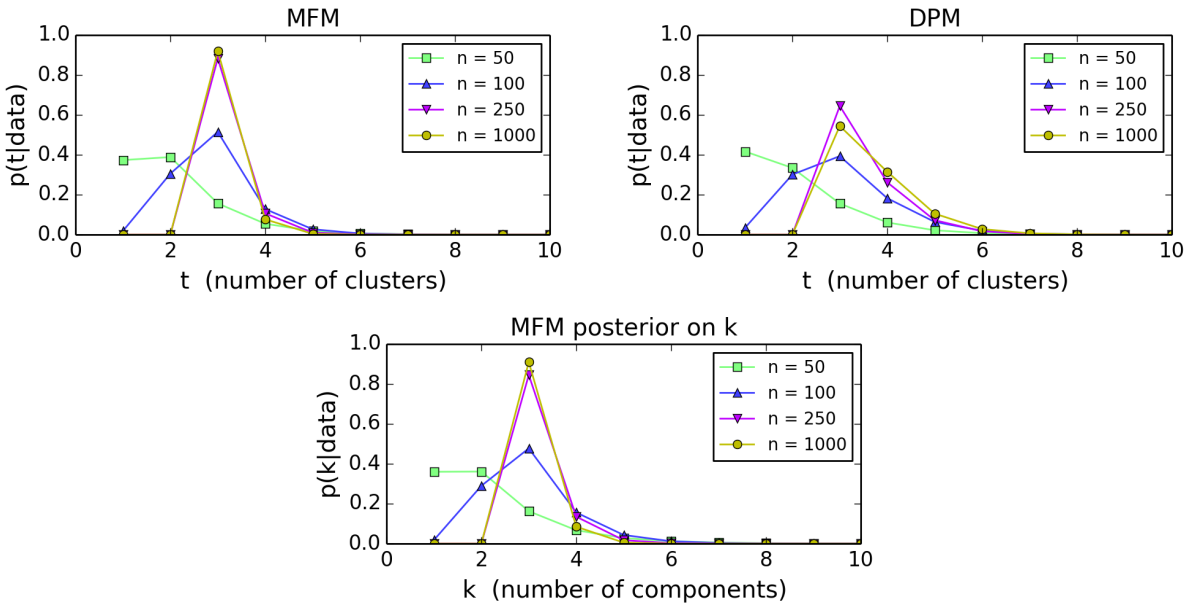

FIGURE 4. Posterior on the number of clusters t for the MFM (top left) and DPM (top right), and the posterior on the number of components k for the MFM (bottom), on increasing amounts of data from a three-component Gaussian mixture. See Section 7.1 for details.

These prior differences carry over to noticeably different posterior clustering behavior. For instance, Figure 3 displays typical clusterings sampled from the posterior, illustrating that DPM samples tend to have tiny "extra" clusters, while MFM samples do not.

It should also be mentioned that due to (2), MFMs "dislike" partitions with very small clusters, causing incremental Gibbs samplers to mix more slowly (empirically) when n is large, however, this is easily remedied by using split-merge samplers (Jain and Neal, 2004, 2007); see Section 6.

Mixing distribution and the number of components. Assuming that the data is from a finite mixture, it is also sometimes of interest to infer the mixing distribution or the number of components, subject to the caveat that these inferences are meaningful only to the extent that the component distributions are correctly specified and the model is mixture identifiable. While Nguyen (2013) has shown that under certain conditions, DPMs are consistent for the mixing distribution (in the Wasserstein metric), Miller and Harrison (2014) have shown that the posterior on the number of clusters in a DPM is typically not consistent for the number of components. On the other hand, MFMs are consistent for the mixing distribution and the number of components (for Lebesgue almost-all parameter values) under very general conditions; this is a straightforward consequence of Doob's theorem (Nobile, 1994). The relative ease with which this consistency can be established for MFMs is due to the fact that in an MFM, the parameter space is a countable union of finite-dimensional spaces, rather than an infinite-dimensional space.

These consistency/inconsistency properties are readily observed empirically—they are not simply large-sample phenomena. As seen in Figure 4, the tendency of DPM samples to have tiny extra clusters causes the number of clusters t to be somewhat inflated, apparently making

the DPM posterior on t fail to concentrate, while the MFM posterior on t concentrates at the true value (see Section 5.2). In addition to the number of clusters t, the MFM also permits inference for the number of components k in a natural way (Figure 4), while in the DPM the number of components is always infinite. See Section 8 for discussion regarding issues with estimating the number of components.

2. Model

We consider the following well-known model:

$$\begin{aligned} &K \sim p_K, \text{ where } p_K \text{ is a p.m.f. on } \{1, 2, \dots\} \\ &(\pi_1, \dots, \pi_k) \sim \text{Dirichlet}_k(\gamma, \dots, \gamma), \text{ given } K = k \\ &Z_1, \dots, Z_n \overset{\text{iid}}{\sim} \pi, \text{ given } \pi \\ &\theta_1, \dots, \theta_k \overset{\text{iid}}{\sim} H, \text{ given } K = k \\ &X_j \sim f_{\theta_{Z_j}} \text{ independently for } j = 1, \dots, n, \text{ given } \theta_{1:K}, Z_{1:n}. \end{aligned} \tag{2.1}$$

Here, H is a prior or "base measure" on $\Theta \subset \mathbb{R}^\ell$, and $\{f_\theta : \theta \in \Theta\}$ is a family of probability densities with respect to a sigma-finite measure ζ on $\mathcal{X} \subset \mathbb{R}^d$. (As usual, we give Θ and $\mathcal{X}$ the Borel sigma-algebra, and assume $(x, \theta) \mapsto f_\theta(x)$ is measurable.) We denote $x_{1:n} = (x_1, \dots, x_n)$. Typically, the values $X_1, \dots, X_n$ would be observed, and all other variables would be hidden/latent. We refer to this as a *mixture of finite mixtures* (MFM) model.

It is important to note that we assume a symmetric Dirichlet with a single parameter γ not depending on k. This assumption is key to deriving a simple form for the partition distribution and the other resulting properties. Assuming symmetry in the distribution of π is quite natural, since the distribution of $X_1, \dots, X_n$ under any asymmetric distribution on π would be the same as if this were replaced by its symmetrized version, i.e., if the entries of π were uniformly permuted (although this would no longer necessarily be a Dirichlet distribution). Assuming the same γ for all k is a genuine restriction, albeit a fairly natural one, often made in such models even when not strictly necessary (Nobile, 1994; Phillips and Smith, 1996; Richardson and Green, 1997; Green and Richardson, 2001; Stephens, 2000; Nobile and Fearnside, 2007). Note that prior information about the relative sizes of the mixing weights $\pi_1, \dots, \pi_k$ can be introduced through γ—roughly speaking, small γ favors lower entropy π's, while large γ favors higher entropy π's.

Meanwhile, we put very few restrictions on p_K, the distribution of the number of components. For practical purposes, we need the infinite series $\sum_{k=1}^\infty p_K(k)$ to converge to 1 reasonably quickly, but any choice of p_K arising in practice should not be a problem. For certain theoretical purposes—in particular, consistency for the number of components—it is desirable to have $p_K(k) > 0$ for all $k \in \{1, 2, \dots\}$.

For comparison, the Dirichlet process mixture (DPM) model with concentration parameter $\alpha > 0$ and base measure H is defined as follows, using Sethuraman's (1994) representation:

$$\begin{aligned} &B_1, B_2, \dots \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha) \\ &Z_1, \dots, Z_n \overset{\text{iid}}{\sim} \pi, \text{ given } \pi = (\pi_1, \pi_2, \dots) \text{ where } \pi_i = B_i \textstyle\prod_{j=1}^{i-1}(1 - B_j) \\ &\theta_1, \theta_2, \dots \overset{\text{iid}}{\sim} H \\ &X_j \sim f_{\theta_{Z_j}} \text{ independently for } j = 1, \dots, n, \text{ given } \theta_{1:\infty}, Z_{1:n}. \end{aligned}$$

3. Exchangeable partition distribution

The primary observation on which our development relies is that the distribution on partitions induced by an MFM takes a form which is simple enough that it can be easily computed. Let $\mathcal{C}$ denote the unordered partition of $[n] := \{1, \dots, n\}$ induced by $Z_1, \dots, Z_n$; in other words, $\mathcal{C} = \{E_i : |E_i| > 0\}$ where $E_i = \{j : Z_j = i\}$ for $i \in \{1, 2, \dots\}$.

Theorem 3.1. *Under the MFM (Equation 2.1), the probability mass function of $\mathcal{C}$ is*

$$p(\mathcal{C}) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \tag{3.1}$$

where $t = |\mathcal{C}|$ is the number of parts in the partition, and

$$V_n(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} p_K(k). \tag{3.2}$$

All proofs have been collected in Appendix B. Here, $x^{(m)} = x(x+1)\cdots(x+m-1)$ and $x_{(m)} = x(x-1)\cdots(x-m+1)$, with $x^{(0)} = 1$ and $x_{(0)} = 1$ by convention. We discuss computation of $V_n(t)$ in Section 3.2. For comparison, under the DPM, the partition distribution induced by $Z_1, \dots, Z_n$ is $p_{\text{DPM}}(\mathcal{C}) = \frac{\alpha^t}{\alpha^{(n)}} \prod_{c \in \mathcal{C}} (|c| - 1)!$ (Antoniak, 1974).

Viewed as a function of the part sizes $(|c| : c \in \mathcal{C})$, Equation 3.1 is an *exchangeable partition probability function* (EPPF) in the terminology of Pitman (1995, 2006), since it is a symmetric function of the part sizes. Consequently, $\mathcal{C}$ is an *exchangeable random partition* of $[n]$; that is, its distribution is invariant under permutations of $[n]$ (alternatively, this can be seen directly from the definition of the model, since $Z_1, \dots, Z_n$ are exchangeable).

More specifically, we observe that Equation 3.1 is a member of the family of Gibbs partition distributions (Pitman, 2006); this is also implied by the results of Gnedin and Pitman (2006) characterizing the extreme points of the space of Gibbs partition distributions. Further results on Gibbs partitions are provided by Ho et al. (2007), Lijoi et al. (2008), Cerquetti (2008, 2011), Gnedin (2010), and Lijoi and Prünster (2010). However, the utility of this representation for inference in mixture models with a prior on the number of components does not seem to have been previously explored in the literature.

Due to Theorem 3.1, we have the following equivalent representation of the model:

$$\begin{aligned} &\mathcal{C} \sim p(\mathcal{C}), \text{ with } p(\mathcal{C}) \text{ as in Equation 3.1} \\ &\phi_c \overset{\text{iid}}{\sim} H \text{ for } c \in \mathcal{C}, \text{ given } \mathcal{C} \\ &X_j \sim f_{\phi_c} \text{ independently for } j \in c,\ c \in \mathcal{C}, \text{ given } \phi, \mathcal{C}, \end{aligned} \tag{3.3}$$

where $\phi = (\phi_c : c \in \mathcal{C})$ is a tuple of $t = |\mathcal{C}|$ parameters $\phi_c \in \Theta$, one for each part $c \in \mathcal{C}$.

This representation is particularly useful for doing inference, since one does not have to deal with cluster labels or empty components. The formulation of models starting from a partition distribution has been a fruitful approach, exemplified by the development of product partition models (Hartigan, 1990; Barry and Hartigan, 1992; Quintana and Iglesias, 2003; Dahl, 2009; Park and Dunson, 2010; Müller and Quintana, 2010; Müller et al., 2011).

3.1. Basic properties.

We list here some basic properties of the MFM model. See Appendix B for proofs. Denoting $x_c = (x_j : j \in c)$ and $m(x_c) = \int_\Theta \left[\prod_{j \in c} f_\theta(x_j)\right] H(d\theta)$ (with the

convention that $m(x_\varnothing) = 1$), we have

$$p(x_{1:n}|\mathcal{C}) = \prod_{c\in\mathcal{C}} m(x_c). \tag{3.4}$$

The number of components K and the number of clusters $T = |\mathcal{C}|$ are related by

$$p(t|k) = \frac{k_{(t)}}{(\gamma k)^{(n)}} \sum_{\mathcal{C}:|\mathcal{C}|=t} \prod_{c\in\mathcal{C}} \gamma^{(|c|)}, \tag{3.5}$$

$$p(k|t) = \frac{1}{V_n(t)} \frac{k_{(t)}}{(\gamma k)^{(n)}} p_K(k), \tag{3.6}$$

where in Equation 3.5, the sum is over partitions $\mathcal{C}$ of $[n]$ such that $|\mathcal{C}| = t$. The formula for $p(k|t)$ is required for doing inference about the number of components K based on posterior samples of $\mathcal{C}$; fortunately, it is easy to compute. We have the conditional independence relations

$$\mathcal{C} \perp K \mid T, \tag{3.7}$$

$$X_{1:n} \perp K \mid T. \tag{3.8}$$

3.2. **The coefficients $V_n(t)$.** The following recursion is a special case of a more general result for Gibbs partitions (Gnedin and Pitman, 2006).

Proposition 3.2. *The numbers $V_n(t)$ (Equation 3.2) satisfy the recursion*

$$V_{n+1}(t+1) = V_n(t)/\gamma - (n/\gamma + t)V_{n+1}(t) \tag{3.9}$$

for any $0 \le t \le n$ and $\gamma > 0$.

This is easily seen by plugging the identity

$$k_{(t+1)} = (\gamma k + n)k_{(t)}/\gamma - (n/\gamma + t)k_{(t)}$$

into the expression for $V_{n+1}(t+1)$. In the case of $\gamma = 1$, Gnedin (2010) has discovered a beautiful example of a distribution on K for which both $p_K(k)$ and $V_n(t)$ have closed-form expressions.

In previous work on the MFM model, it has been common for p_K to be chosen to be proportional to a Poisson distribution restricted to the positive integers or a subset thereof (Phillips and Smith, 1996; Stephens, 2000; Nobile and Fearnside, 2007), and Nobile (2005) has proposed a theoretical justification for this choice. Interestingly, the model has some nice mathematical properties if one instead chooses $K-1$ to be given a Poisson distribution, that is, $p_K(k) = \text{Poisson}(k-1|\lambda)$ for some $\lambda > 0$. One example of this arises here (for another example, see Section 4.3): it turns out that if $p_K(k) = \text{Poisson}(k-1|\lambda)$ and $\gamma = 1$ then

$$V_n(0) = \frac{1}{\lambda^n}\Big(1 - \sum_{k=1}^{n} p_K(k)\Big). \tag{3.10}$$

However, to do inference, it is not necessary to choose p_K to have any particular form. To do inference, we just need to be able to compute $p(\mathcal{C})$, and in turn, we need to be able to compute $V_n(t)$. To this end, note that $k_{(t)}/(\gamma k)^{(n)} \le k^t/(\gamma k)^n$, and thus the infinite series for $V_n(t)$ converges rapidly when $t \ll n$. It always converges to a finite value when $1 \le t \le n$; this is clear from the fact that $p(\mathcal{C}) \in [0, 1]$. This finiteness can also be seen directly from the series since $k^t/(\gamma k)^n \le 1/\gamma^n$, and in fact, this shows that the series for $V_n(t)$ converges at

least as rapidly (up to a constant) as the series $\sum_{k=1}^{\infty} p_K(k)$ converges to 1. Hence, for any reasonable choice of p_K (i.e., not having an extraordinarily heavy tail), $V_n(t)$ can easily be numerically approximated to a high level of precision. In practice, computing the required values of $V_n(t)$ takes a negligible amount of time.

3.3. **Self-consistent marginals.** For each $n = 1, 2, \ldots$, let $q_n(\mathcal{C})$ denote the distribution on partitions of $[n]$ as defined above (Equation 3.1). This family of partition distributions is preserved under marginalization, in the following sense.

Proposition 3.3. *If $m < n$ then q_m coincides with the marginal distribution on partitions of $[m]$ induced by q_n.*

In other words, drawing a sample from q_n and removing elements $m+1, \ldots, n$ from it yields a sample from q_m. This can be seen directly from the model definition (Equation 2.1), since $\mathcal{C}$ is the partition induced by the Z's, and the distribution of $Z_{1:m}$ is the same when the model is defined with any $n \geq m$. This property is sometimes referred to as *consistency in distribution* (Pitman, 2006).

By Kolmogorov's extension theorem (e.g., Durrett, 1996), it is well-known that this implies the existence of a unique probability distribution on partitions of the positive integers $\mathbb{Z}_{>0} = \{1, 2, \ldots\}$ such that the marginal distribution on partitions of $[n]$ is q_n for all $n \in \{1, 2, \ldots\}$. A random partition of $\mathbb{Z}_{>0}$ from such a distribution is a *combinatorial stochastic process*; for background, see Pitman (2006).

4. Restaurant process, stick-breaking, and random measure representations

4.1. **Pólya urn scheme / Restaurant process.** Pitman (1996) considered a general class of urn schemes, or restaurant processes, corresponding to exchangeable partition probability functions (EPPFs). The following scheme for the MFM falls into this general class.

Theorem 4.1. *The following process generates partitions $\mathcal{C}_1, \mathcal{C}_2, \ldots$ such that for any $n \in \{1, 2, \ldots\}$, the probability mass function of $\mathcal{C}_n$ is given by Equation 3.1.*

- *Initialize with a single cluster consisting of element 1 alone: $\mathcal{C}_1 = \{\{1\}\}$.*
- *For $n = 2, 3, \ldots$, element n is placed in ...*
  - *an existing cluster $c \in \mathcal{C}_{n-1}$ with probability $\propto |c| + \gamma$*
  - *a new cluster with probability $\propto \dfrac{V_n(t+1)}{V_n(t)}\gamma$*

  where $t = |\mathcal{C}_{n-1}|$.

Clearly, this bears a close resemblance to the Chinese restaurant process (i.e., the Blackwell–MacQueen urn process), in which the nth element is placed in an existing cluster c with probability $\propto |c|$ or a new cluster with probability $\propto \alpha$ (the concentration parameter) (Blackwell and MacQueen, 1973; Aldous, 1985).

4.2. **Random discrete measures.** The MFM can also be formulated starting from a distribution on discrete measures that is analogous to the Dirichlet process. With K, π, and $\theta_{1:K}$ as in Equation 2.1, let

$$G = \sum_{i=1}^{K} \pi_i \delta_{\theta_i}$$

where δ_θ is the unit point mass at θ. Let us denote the distribution of G by $\mathcal{M}(p_K, \gamma, H)$. Note that G is a random discrete measure over Θ. If H is continuous (i.e., $H(\{\theta\}) = 0$ for all $\theta \in \Theta$), then with probability 1, the number of atoms is K; otherwise, there may be fewer than K atoms. If we take $X_1, \dots, X_n | G$ i.i.d. from the resulting mixture, namely,

$$f_G(x) := \int f_\theta(x) G(d\theta) = \sum_{i=1}^{K} \pi_i f_{\theta_i}(x),$$

then the distribution of $X_{1:n}$ is the same as before. So, in this notation, the MFM model is:

$$G \sim \mathcal{M}(p_K, \gamma, H)$$
$$X_1, \dots, X_n \overset{\text{iid}}{\sim} f_G, \text{ given } G.$$

This random discrete measure perspective is connected to work on *species sampling models* (Pitman, 1996) in the following way. When H is continuous, we can construct a species sampling model by letting $G \sim \mathcal{M}(p_K, \gamma, H)$ and modeling the observed data as $\beta_1, \dots, \beta_n \sim G$. We refer to Pitman (1996), Hansen and Pitman (2000), Ishwaran and James (2003), Lijoi et al. (2005, 2007), and Lijoi et al. (2008) for more background on species sampling models and further examples. The following posterior predictive rule for this particular model follows the form of Pitman's general rule; note the close relationship to the restaurant process (Theorem 4.1 above).

Theorem 4.2. *If H is continuous, then $\beta_1 \sim H$ and the distribution of β_n given $\beta_1, \dots, \beta_{n-1}$ is proportional to*

$$\frac{V_n(t+1)}{V_n(t)} \gamma H + \sum_{i=1}^{t} (n_i + \gamma) \delta_{\beta_i^*}, \tag{4.1}$$

where $\beta_1^*, \dots, \beta_t^*$ are the distinct values taken by $\beta_1, \dots, \beta_{n-1}$, and $n_i = \#\{j \in [n-1] : \beta_j = \beta_i^*\}$.

For comparison, when $G \sim \mathrm{DP}(\alpha H)$ instead, the distribution of β_n given $\beta_1, \dots, \beta_{n-1}$ is proportional to $\alpha H + \sum_{j=1}^{n-1} \delta_{\beta_j} = \alpha H + \sum_{i=1}^{t} n_i \beta_i^*$, since $G | \beta_1, \dots, \beta_{n-1} \sim \mathrm{DP}(\alpha H + \sum_{j=1}^{n-1} \delta_{\beta_j})$ (Ferguson, 1973; Blackwell and MacQueen, 1973).

4.3. **Stick-breaking representation.** The Dirichlet process has an elegant stick-breaking representation for the mixture weights $\pi_1, \pi_2, \dots$ (Sethuraman, 1994; Sethuraman and Tiwari, 1981). This extraordinarily clarifying perspective has inspired a number of nonparametric models (MacEachern, 1999, 2000; Hjort, 2000; Ishwaran and Zarepour, 2000; Ishwaran and James, 2001; Griffin and Steel, 2006; Dunson and Park, 2008; Chung and Dunson, 2009; Rodriguez and Dunson, 2011; Broderick et al., 2012), has provided insight into the properties of related models (Favaro et al., 2012; Teh et al., 2007; Thibaux and Jordan, 2007; Paisley et al., 2010), and has been used to develop efficient inference algorithms (Ishwaran and James, 2001; Blei and Jordan, 2006; Papaspiliopoulos and Roberts, 2008; Walker, 2007; Kalli et al., 2011).

In a certain special case—namely, when $p_K(k) = \mathrm{Poisson}(k-1|\lambda)$ and $\gamma = 1$—we have noticed that the MFM also has an interesting representation that we describe using the stick-breaking analogy, although it is somewhat different in nature. This is another example of the nice mathematical properties resulting from this choice of p_K and γ. Consider the following procedure:

Take a unit-length stick, and break off i.i.d. Exponential(λ) *pieces until you run out of stick.*

In other words, let $\epsilon_1, \epsilon_2, \ldots \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$, define $\tilde{K} = \min\{j : \sum_{i=1}^{j} \epsilon_i \geq 1\}$, and set $\tilde{\pi}_i = \epsilon_i$ for $i = 1, \ldots, \tilde{K} - 1$ and $\tilde{\pi}_{\tilde{K}} = 1 - \sum_{i=1}^{\tilde{K}-1} \tilde{\pi}_i$.

Proposition 4.3. *The stick lengths $\tilde{\pi}$ have the same distribution as the mixture weights π in the MFM model when $p_K(k) = \text{Poisson}(k-1|\lambda)$ and $\gamma = 1$.*

This is a consequence of a standard construction for Poisson processes. This suggests a way of generalizing the MFM model: take any sequence of nonnegative random variables $(\epsilon_1, \epsilon_2, \ldots)$ (not necessarily independent or identically distributed) such that $\sum_{i=1}^{\infty} \epsilon_i > 1$ with probability 1, and define $\tilde{K}$ and $\tilde{\pi}$ as above. Although the distribution of $\tilde{K}$ and $\tilde{\pi}$ may be complicated, in some cases it might still be possible to do inference based on the stick-breaking representation. This might be an interesting way to introduce different kinds of prior information on the mixture weights, however, we have not explored this possibility.

5. Asymptotics

In this section, we consider the asymptotics of $V_n(t)$, the asymptotic relationship between the number of components and the number of clusters, and the approximate form of the conditional distribution on cluster sizes given the number of clusters.

5.1. **Asymptotics of $V_n(t)$.** Recall that $V_n(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} p_K(k)$ (Equation 3.2) for $1 \leq t \leq n$, with $\gamma > 0$ and p_K a p.m.f. on $\{1, 2, \ldots\}$.

Theorem 5.1. *For any $t \in \{1, 2, \ldots\}$, if $p_K(t) > 0$ then*

$$V_n(t) \sim \frac{t_{(t)}}{(\gamma t)^{(n)}} p_K(t) \sim \frac{t!}{n!} \frac{\Gamma(\gamma t)}{n^{\gamma t - 1}} p_K(t) \tag{5.1}$$

as $n \to \infty$.

In particular, $V_n(t)$ has a simple interpretation, asymptotically—it behaves like the $k = t$ term in the series.

5.2. **Relationship between the number of clusters and number of components.** In the MFM, it is perhaps intuitively clear that, under the prior at least, the number of clusters $T = |\mathcal{C}|$ behaves very similarly to the number of components K when n is large. It turns out that under the posterior they also behave very similarly for large n.

Theorem 5.2. *Let $x_1, x_2, \ldots \in \mathcal{X}$ and $k \in \{1, 2, \ldots\}$. If $p_K(1), \ldots, p_K(k) > 0$ then*

$$\big|p(T = k \mid x_{1:n}) - p(K = k \mid x_{1:n})\big| \longrightarrow 0$$

as $n \to \infty$.

5.3. **Distribution of the cluster sizes under the prior.** Here, we examine one of the major differences between the MFM and DPM priors. Roughly speaking, under the prior, the MFM prefers all clusters to be the same order of magnitude, while the DPM prefers having a few large clusters and many very small clusters. In the following calculations, we quantify the preceding statement more precisely. (See Green and Richardson (2001) for informal observations along these lines.) Interestingly, these prior influences remain visible in certain

aspects of the posterior, even in the limit as n goes to infinity, as shown by the inconsistency of DPMs for the number of components in a finite mixture (Miller and Harrison, 2014).

Let $\mathcal{C}$ be the partition of $[n]$ in the MFM model (Equation 3.1), and let $A = (A_1, \dots, A_T)$ be the ordered partition of $[n]$ obtained by randomly ordering the parts of $\mathcal{C}$, uniformly among the $T!$ possible choices, where $T = |\mathcal{C}|$. Then

$$p(A) = \frac{p(\mathcal{C})}{|\mathcal{C}|!} = \frac{1}{t!} V_n(t) \prod_{i=1}^{t} \gamma^{(|A_i|)},$$

where $t = |\mathcal{C}|$. Now, let $S = (S_1, \dots, S_T)$ be the vector of part sizes of A, that is, $S_i = |A_i|$. Then

$$p(S = s) = \sum_{A : S(A) = s} p(A) = V_n(t) \, \frac{n!}{t!} \prod_{i=1}^{t} \frac{\gamma^{(s_i)}}{s_i!}$$

for $s \in \Delta_t$, $t \in \{1, \dots, n\}$, where $\Delta_t = \{s \in \mathbb{Z}^t : \sum_i s_i = n, \, s_i \geq 1 \, \forall i\}$ (i.e., the t-part compositions of n). For any $x > 0$, writing $x^{(m)}/m! = \Gamma(x+m)/(m! \, \Gamma(x))$ and using Stirling's approximation, we have $x^{(m)}/m! \sim m^{x-1}/\Gamma(x)$ as $m \to \infty$. This yields the approximations

$$p(S = s) \approx \frac{V_n(t)}{\Gamma(\gamma)^t} \, \frac{n!}{t!} \prod_{i=1}^{t} s_i^{\gamma - 1} \approx \frac{p_K(t)}{n^{\gamma t - 1}} \, \frac{\Gamma(\gamma t)}{\Gamma(\gamma)^t} \prod_{i=1}^{t} s_i^{\gamma - 1}$$

(using Theorem 5.1 in the second step), and

$$p(S = s \mid T = t) \approx \kappa \prod_{i=1}^{t} s_i^{\gamma - 1}$$

for $s \in \Delta_t$, where κ is a normalization constant. Thus $p(s|t)$, although a discrete distribution, has approximately the same shape as a symmetric t-dimensional Dirichlet distribution. This would be obvious if we were conditioning on the number of components K, and it makes intuitive sense when conditioning on T, since K and T are essentially the same for large n.

It is very interesting to compare this to the corresponding distributions for Dirichlet process mixtures. In the DPM, we have $p_{\text{DPM}}(\mathcal{C}) = \frac{\alpha^t}{\alpha^{(n)}} \prod_{c \in \mathcal{C}} (|c| - 1)!$, and $p_{\text{DPM}}(A) = p_{\text{DPM}}(\mathcal{C})/|\mathcal{C}|!$ as before, so for $s \in \Delta_t$, $t \in \{1, \dots, n\}$,

$$p_{\text{DPM}}(S = s) = \frac{n!}{\alpha^{(n)}} \, \frac{\alpha^t}{t!} \, s_1^{-1} \cdots s_t^{-1}$$

and

$$p_{\text{DPM}}(S = s \mid T = t) \propto s_1^{-1} \cdots s_t^{-1},$$

which has the same shape as a t-dimensional Dirichlet distribution with all the parameters taken to 0 (noting that this is normalizable since Δ_t is finite). Asymptotically in n, $p_{\text{DPM}}(s|t)$ puts all of its mass in the "corners" of the discrete simplex Δ_t, while under the MFM, $p(s|t)$ remains more evenly dispersed.

6. Inference algorithms

As shown by the results of Sections 3 and 4, MFMs have many of the same properties as DPMs. As a result, much of the extensive body of work on MCMC samplers for DPMs can be directly applied to MFMs, including samplers for conjugate and non-conjugate cases, as well as split-merge samplers.

When H is a conjugate prior for $\{f_\theta\}$, such that the marginal likelihood $m(x_c) = \int_\Theta \left[\prod_{j\in c} f_\theta(x_j)\right] H(d\theta)$ can be easily computed, the following Gibbs sampling algorithm can be used to sample from the posterior on partitions, $p(\mathcal{C}|x_{1:n})$. Given a partition $\mathcal{C}$, let $\mathcal{C} \setminus j$ denote the partition obtained by removing element j from $\mathcal{C}$.

(1) Initialize $\mathcal{C} = \{[n]\}$ (i.e., one cluster).
(2) Repeat the following N times, to obtain N samples.

For $j = 1, \dots, n$: Remove element j from $\mathcal{C}$, and place it ...

in $c \in \mathcal{C} \setminus j$ with probability $\propto (|c| + \gamma)\dfrac{m(x_{c\cup j})}{m(x_c)}$

in a new cluster with probability $\propto \gamma\, \dfrac{V_n(t+1)}{V_n(t)}\, m(x_j)$

where $t = |\mathcal{C} \setminus j|$.

This is a direct adaptation of "Algorithm 3" for DPMs (MacEachern, 1994; Neal, 1992, 2000). The only differences are that in Algorithm 3, $|c| + \gamma$ is replaced by $|c|$, and $\gamma V_n(t+1)/V_n(t)$ is replaced by α (the concentration parameter). Thus, the differences between the MFM and DPM versions of the algorithm are precisely the same as the differences between their respective restaurant processes. Computing the required values of $V_n(t)$ takes a negligible amount of time compared to running the sampler. In order for the algorithm to be valid, the Markov chain needs to be irreducible, and to achieve this it is necessary to have $\{t \in \{1, 2, \dots\} : V_n(t) > 0\}$ be a block of consecutive integers. In fact, it turns out that this is always the case (and this block includes $t = 1$), since for any k such that $p_K(k) > 0$, we have $V_n(t) > 0$ for all $t = 1, \dots, k$.

When H is a non-conjugate prior (and $m(x_c)$ cannot be easily computed), a clever auxiliary variable technique referred to as "Algorithm 8" can be used for inference in the DPM (Neal, 2000; MacEachern and Müller, 1998). Making the same substitutions as above, we can apply Algorithm 8 to perform inference in the MFM as well; see Miller (2014) for details.

A well-known issue with incremental Gibbs samplers such as these, however, when applied to DPMs, is that the mixing can be somewhat slow, since it may take a long time to create or destroy substantial clusters by moving one element at a time. With MFMs, this issue seems to be exacerbated, since MFMs tend to put small probability (compared with DPMs) on partitions with tiny clusters (see Section 5.3), making it difficult for the sampler to move through these regions of the space.

To deal with this issue, split-merge samplers for DPMs have been developed, in which a large number of elements can be reassigned in a single move (Dahl, 2003, 2005; Jain and Neal, 2004, 2007). In the same way as the incremental samplers, one can directly apply these split-merge samplers (both conjugate and non-conjugate) to MFMs, using the properties described in Sections 3 and 4. More generally, it seems likely that any partition-based MCMC sampler for DPMs could be applied to MFMs as well.

In Section 7, we apply the Jain–Neal split-merge samplers coupled with incremental Gibbs samplers, in both conjugate and non-conjugate settings.

7. Empirical demonstrations

In this section, we demonstrate the MFM on simulated and real datasets. All of the examples below involve Gaussian component densities, but of course our approach is not limited to mixtures of Gaussians.

7.1. **Simulation example.** In the introduction, we presented several figures comparing the MFM and DPM on data from a three-component bivariate Gaussian mixture, illustrating the behavior of the MFM with respect to density estimation, clustering, and inference for the number of components. Here, we provide the details of the data, model, and method of inference for this simulation example.

Data. The data distribution is $\sum_{i=1}^{3} w_i \mathcal{N}(\mu_i, C_i)$ where $w = (0.45, 0.3, 0.25)$, $\mu_1 = \left(\begin{smallmatrix} 4 \\ 4 \end{smallmatrix}\right)$, $\mu_2 = \left(\begin{smallmatrix} 7 \\ 4 \end{smallmatrix}\right)$, $\mu_3 = \left(\begin{smallmatrix} 6 \\ 2 \end{smallmatrix}\right)$, $C_1 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $C_2 = R\left(\begin{smallmatrix} 2.5 & 0 \\ 0 & 0.2 \end{smallmatrix}\right)R^{\mathsf{T}}$ where $R = \left(\begin{smallmatrix} \cos\rho & -\sin\rho \\ \sin\rho & \cos\rho \end{smallmatrix}\right)$ with $\rho = \pi/4$, and $C_3 = \left(\begin{smallmatrix} 3 & 0 \\ 0 & 0.1 \end{smallmatrix}\right)$.

Model. The component densities are multivariate normal, $f_\theta(x) = f_{\mu,\Lambda}(x) = \mathcal{N}(x|\mu, \Lambda^{-1})$ and the base measure (prior) H on $\theta = (\mu, \Lambda)$ is $\mu \sim \mathcal{N}(\widehat{\mu}, \widehat{C})$, $\Lambda \sim \text{Wishart}_d(V, \nu)$ independently, where $\widehat{\mu}$ is the sample mean, $\widehat{C}$ is the sample covariance, $\nu = d = 2$, and $V = \widehat{C}^{-1}/\nu$. Here, $\text{Wishart}_d(\Lambda|V, \nu) \propto |\det \Lambda|^{(\nu-d-1)/2} \exp\big(-\frac{1}{2}\text{tr}(V^{-1}\Lambda)\big)$. Note that this is a data-dependent prior.

For the MFM, we take $K \sim \text{Geometric}(r)$ ($p_K(k) = (1-r)^{k-1}r$ for $k = 1, 2, \dots$) with $r = 0.1$, and we choose $\gamma = 1$ for the finite-dimensional Dirichlet parameters. For the DPM, we put an Exponential(1) prior on the concentration parameter, α.

Note that taking μ and Λ to be independent results in a non-conjugate prior. This prior is appropriate when the location of the data is not informative about the covariance (and vice versa).

Inference. For both the MFM and DPM, we use the non-conjugate split-merge sampler of Jain and Neal (2007), coupled with Algorithm 8 of Neal (2000) (using a single auxiliary variable) for incremental Gibbs updates to the partition. Specifically, following Jain and Neal (2007), we use the (5,1,1,5) scheme: 5 intermediate scans to reach the split launch state, 1 split-merge move per iteration, 1 incremental Gibbs scan per iteration, and 5 intermediate moves to reach the merge launch state. Gibbs updates to the DPM concentration parameter α are made using Metropolis–Hastings moves.

Five independent datasets were used for each $n \in \{50, 100, 250, 1000\}$, and for each model (MFM and DPM), the sampler was run for 5,000 burn-in iterations and 95,000 sample iterations (for a total of 100,000). Judging by traceplots and running averages of various statistics, this appeared to be sufficient for mixing. The cluster sizes were recorded after each iteration, and to reduce memory storage requirements, the full state of the chain was recorded only once every 100 iterations. For each run, the seed of the random number generator was initialized to the same value for both the MFM and DPM.

For a dataset of size n, the sampler used for these experiments took approximately $8 \times 10^{-6}\, n$ seconds per iteration, using a 2.80 GHz processor with 6 GB of RAM.

Results. As described in the introduction, the results of this simulation empirically indicate that on data from a finite mixture, MFMs and DPMs are consistent for the density (Figures 1 and 2), DPM clusterings tend to have small extra clusters while MFM clusterings do not (Figure 3), and MFMs are consistent for the number of components while DPMs are not (Figure 4). This is what we expect from theory (although to be precise, the inconsistency result of Miller and Harrison (2014) only applies to the case of fixed concentration parameter α). These results are not too surprising, since when the data distribution is a finite mixture from the assumed family, the MFM is correctly specified, while the DPM is not. On data from

an infinite mixture, one would expect the DPM to have certain advantages. See Appendix A for formulas for computing the posterior on k and the density estimates.

7.2. **Galaxy dataset.** The galaxy dataset (Roeder, 1990) is a standard benchmark for mixture models, consisting of measurements of the velocities of 82 galaxies in the Corona Borealis region; see Figure 5. The purpose of this example is to demonstrate agreement between our method and published results using reversible jump MCMC with the same model, and also to show that using hyperpriors presents no difficulties.

Model. To enable comparison, we use exactly the same model as Richardson and Green (1997). The component densities are univariate normal, $f_\theta(x) = f_{\mu,\lambda}(x) = \mathcal{N}(x|\mu, \lambda^{-1})$, and the base measure H on $\theta = (\mu, \lambda)$ is $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, $\lambda \sim \text{Gamma}(a, b)$ independently (where $\text{Gamma}(\lambda|a, b) \propto \lambda^{a-1}e^{-b\lambda}$). Further, a hyperprior is placed on b, by taking $b \sim \text{Gamma}(a_0, b_0)$. The remaining parameters are set to $\mu_0 = (\max\{x_i\} + \min\{x_i\})/2$, $\sigma_0 = \max\{x_i\} - \min\{x_i\}$, $a = 2$, $a_0 = 0.2$, and $b_0 = 10/\sigma_0^2$. Note that the parameters μ_0, σ_0, and b_0 are functions of the observed data $x_1, \ldots, x_n$. See Richardson and Green (1997) for the rationale behind these parameter choices. (Note: This choice of σ_0 may be a bit too large, affecting the posteriors on the number of clusters and components, however, we stick with it to enable comparisons to Richardson and Green (1997).) For the MFM, following Richardson and Green (1997), we take $K \sim \text{Uniform}\{1, \ldots, 30\}$ and $\gamma = 1$. For the DPM, we take $\alpha \sim \text{Exponential}(1)$.

Inference. As before, we use the non-conjugate split-merge sampler of Jain and Neal (2007) coupled with Algorithm 8 of Neal (2000), and Gibbs updates to the DPM concentration parameter α are made using Metropolis–Hastings. We use Gibbs sampling to handle the hyperprior on b (i.e., append b to the state of the Markov chain, run the sampler given b as usual, and periodically sample b given everything else). More general hyperprior structures can be handled similarly. In all other respects, the same inference algorithm as in Section 7.1 was used.

We do not restrict the parameter space in any way (e.g., forcing the component means to be ordered to obtain identifiability, as was done by Richardson and Green (1997)). All of the quantities we consider are invariant to the labeling of the clusters. See Jasra et al. (2005) for discussion on this point.

The sampler was run for 5,000 burn-in iterations, and 45,000 sample iterations. This appeared to be more than sufficient for mixing. Cluster sizes were recorded after each iteration, and the full state of the chain was recorded every 50 iterations. Each iteration took approximately $8 \times 10^{-6}\, n$ seconds, with $n = 82$.

Results. Figure 5 shows the estimated densities and the posteriors on the number of clusters and components. Comparing this with Figure 2(c) of Richardson and Green (1997), we see that our MFM density estimate is visually indistinguishable from theirs (as it should be, since we are using the same model with the same parameters).

Table 1 compares our estimate of the MFM posterior on the number of components k with the results of Richardson and Green (1997). Again, the results are very close, as expected.

7.3. **Discriminating cancer types using gene expression data.** In cancer research, gene expression profiling—that is, measuring the degree to which each gene is expressed by a given tissue sample under given conditions—enables the identification of distinct subtypes of cancer, leading to greater understanding of the mechanisms underlying cancers as well as

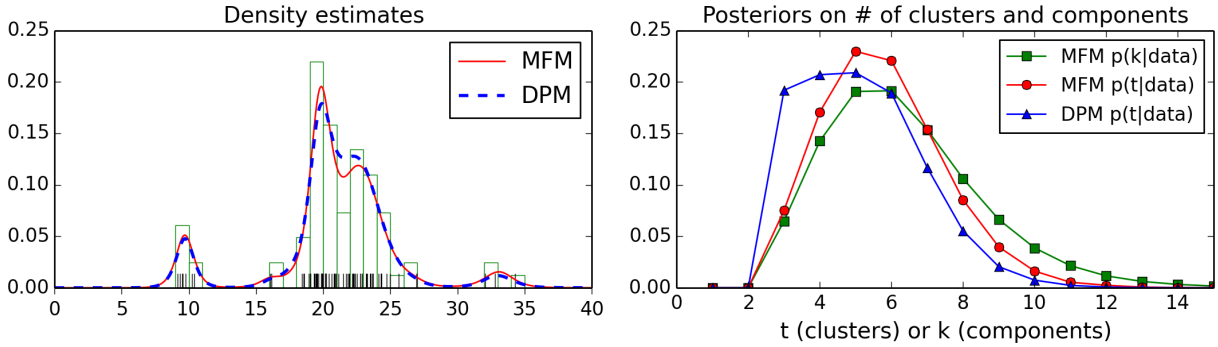

FIGURE 5. Results on the galaxy dataset. Left: Histogram of the data (green bars), rug plot of the data (black ticks), and estimated densities using the MFM (red solid line) and DPM (blue dashed line). Right: MFM and DPM posteriors on the number of clusters (t), along with the MFM posterior on the number of components (k).

TABLE 1. Estimate of the MFM posterior on k for the galaxy dataset.

k	1	2	3	4	5	6	7
Here	0.000	0.000	0.065	0.143	0.191	0.191	0.153
R&G	0.000	0.000	0.061	0.128	0.182	0.199	0.160

8	9	10	11	12	13	14	15
0.106	0.066	0.039	0.021	0.012	0.006	0.003	0.002
0.109	0.071	0.040	0.023	0.013	0.006	0.003	0.002

potentially providing patient-specific diagnostic tools. In gene expression datasets, there are typically a small number of very high-dimensional data points, each consisting of the gene expression levels in a given tissue sample under given conditions.

One approach to analyzing gene expression data is to use Gaussian mixture models to identify clusters which may represent distinct cancer subtypes (Yeung et al., 2001; McLachlan et al., 2002; Medvedovic and Sivaganesan, 2002; Medvedovic et al., 2004; de Souto et al., 2008; Rasmussen et al., 2009; McNicholas and Murphy, 2010). In fact, in a comparative study of seven clustering methods on 35 cancer gene expression datasets with known ground truth, de Souto et al. (2008) found that finite mixtures of Gaussians provided the best results—when the number of components k was set to the true value. However, in practice, choosing an appropriate value of k can be difficult. Using the methods developed in this paper, the MFM provides a principled approach to inferring the clusters even when k is unknown, as well as doing inference for k, provided that the components are well-modeled by Gaussians. (However, see Section 8 for some potential pitfalls.)

The purpose of this example is to demonstrate that our approach can work well even in very high-dimensional settings, and may provide a useful tool for this application. It should be emphasized that we are not cancer scientists, so the results reported here should not be interpreted as scientifically relevant, but simply as a proof-of-concept.

Data. We apply the MFM to gene expression data collected by Armstrong et al. (2001) in a study of leukemia subtypes. Armstrong et al. (2001) measured gene expression levels in samples from 72 patients who were known to have one of two leukemia types, acute lymphoblastic leukemia (ALL) or acute myelogenous leukemia (AML), and they found that a previously undistinguished subtype of ALL, which they termed mixed-lineage leukemia (MLL), could be distinguished from conventional ALL and AML based on the gene expression profiles.

We use the preprocessed data provided by de Souto et al. (2008), which they filtered to include only genes with expression levels differing by at least 3-fold in at least 30 samples, relative to their mean expression level across all samples. The resulting dataset consists of 72 samples and 1081 genes per sample, i.e., $n = 72$ and $d = 1081$. Following standard practice, we take the base-2 logarithm of the data before analysis, and normalize each dimension to have zero mean and unit variance.

Model. For simplicity, we use multivariate Gaussian component densities with diagonal covariance matrices, i.e., the dimensions are independent univariate Gaussians, and we place independent conjugate priors on each dimension. Thus, for each component, for $i = 1, \dots, d$, dimension i is $\mathcal{N}(\mu_i, \lambda_i^{-1})$, with $\lambda_i \sim \text{Gamma}(a, b)$ and $\mu_i|\lambda_i \sim \mathcal{N}(0, (c\lambda_i)^{-1})$. We choose $a = 1$, $b = 1$, and $c = 1$. (Recall that the data is zero mean, unit variance in each dimension.) For the MFM, $K \sim \text{Geometric}(0.1)$ and $\gamma = 1$, and for the DPM, $\alpha \sim \text{Exponential}(1)$. These are all simply default settings and have not been tailored to the problem; a careful scientific investigation would involve thorough prior elicitation, sensitivity analysis, and model checking.

Inference. Given the partition $\mathcal{C}$ of the data into clusters, the parameters can be integrated out analytically since the prior is conjugate. Thus, for both the MFM and DPM, we use the split-merge sampler of Jain and Neal (2004) for conjugate priors, coupled with Algorithm 3 of Neal (2000). Following Jain and Neal (2004), we use the (5,1,1) scheme: 5 intermediate scans to reach the split launch state, 1 split-merge move per iteration, and 1 incremental Gibbs scan per iteration.

Due to the high-dimensionality of the parameters, this has far better mixing time than sampling the parameters, as is done in reversible jump MCMC.

The sampler was run for 1,000 burn-in iterations, and 19,000 sample iterations. This appears to be many more iterations than required for burn-in and mixing in this particular example—in fact, only 5 to 10 iterations are required to separate the clusters, and the results are indistinguishable when using only 10 burn-in and 190 sample iterations. The full state of the chain was recorded every 20 iterations. Each iteration took approximately $1.3 \times 10^{-3}\, n$ seconds, with $n = 72$.

Results. The posteriors on the number of clusters t are concentrated at 3 (see Figure 6), in agreement with the division into ALL, MLL, and AML determined by Armstrong et al. (2001). The MFM posterior on k is shifted slightly to the right because there are a small number of observations; this accounts for uncertainty regarding the possibility of additional components that were not observed in the given data.

Figure 6 also shows the MFM pairwise probability matrix, that is, the matrix in which entry (i, j) is the posterior probability that data points i and j belong to the same cluster; in the figure, white is probability 0, black is probability 1. (The DPM matrix, not shown, is indistinguishable from the MFM matrix.) The rows and columns of the matrix are ordered

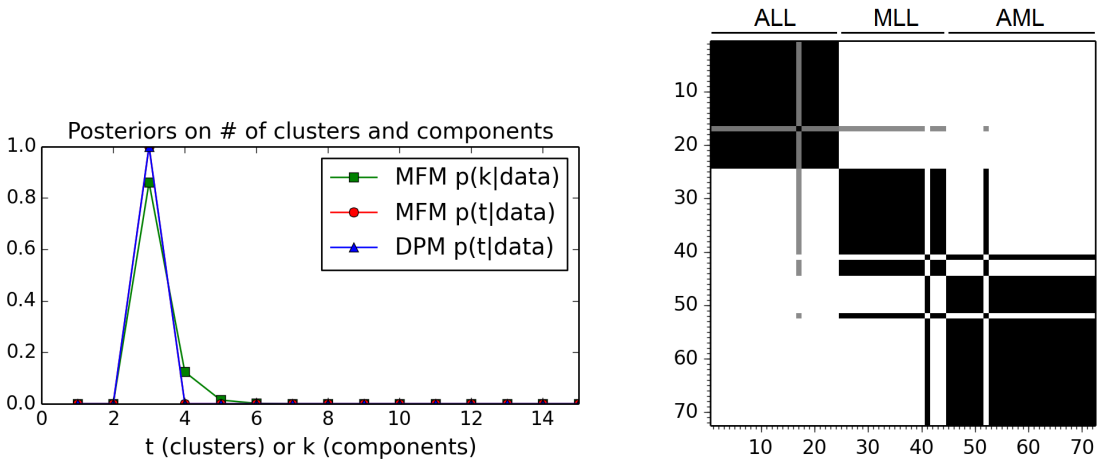

FIGURE 6. Results on the leukemia gene expression dataset. Left: Posteriors on the number of clusters and components. Right: MFM pairwise probability matrix (the DPM matrix is the same). See the text for discussion.

according to ground truth, such that 1–24 are ALL, 25–44 are MLL, and 45–72 are AML. The model has clearly separated the subjects into these three groups, with a small number of exceptions: subject 41 is clustered with the AML subjects instead of MLL, subject 52 with the MLL subjects instead of AML, and subject 17 is about 50% ALL and 50% MLL.

8. Discussion

Due to the fact that inference for the number of components is a topic of high interest in many research communities, it seems prudent to make some cautionary remarks in this regard. Many approaches have been proposed for estimating the number of components (Henna, 1985; Keribin, 2000; Leroux, 1992; Ishwaran et al., 2001; James et al., 2001; Henna, 2005; Woo and Sriram, 2006, 2007). In theory, the MFM model provides a Bayesian approach to consistently estimating the number of components, making it a potentially attractive method of assessing the heterogeneity of the data. However, there are several possible pitfalls to consider, some of which are more obvious than others.

An obvious potential issue is that in many applications, the clusters which one wishes to distinguish are purely notional (for example, perhaps, clusters of images or documents), and a mixture model is used for practical purposes, rather than because the data is actually thought to arise from a mixture. Clearly, in such cases, inference for the "true" number of components is meaningless. On the other hand, in some applications, the data definitely comes from a mixture (for example, extracellular recordings of multiple neurons)—so there is in reality a true number of components—however, usually the form of the mixture components is far from clear.

More subtle issues are that the posteriors on k and t can be

(1) strongly affected by the base measure H, and
(2) sensitive to misspecification of the family of component distributions $\{f_\theta\}$.

Issue (1) can be seen, for instance, in the case of normal mixtures: it might seem desirable to choose the prior on the component means to have large variance in order to be less

informative, however, this causes the posteriors on k and t to favor smaller values (Richardson and Green, 1997; Stephens, 2000; Jasra et al., 2005). The basic mechanism at play here is the same as in the Bartlett–Lindley paradox, and shows up in many Bayesian model selection problems. With some care, this issue can be dealt with by varying the base measure H and observing the effect on the posterior—that is, by performing a sensitivity analysis—for instance, see Richardson and Green (1997).

Issue (2) is more serious—in practice, we typically cannot expect our choice of $\{f_\theta : \theta \in \Theta\}$ to contain the true component densities (assuming the data is even from a mixture). When the model is misspecified in this way, the posteriors of k and t can be severely affected and depend strongly on n. For instance, if the model uses mixtures of Gaussians, and the true data distribution is not a finite mixture of Gaussians, then these posteriors can be expected to diverge to infinity as n increases. Consequently, the effects of misspecification need to be carefully considered if these posteriors are to be used as measures of heterogeneity. Steps toward addressing the issue of robustness have been taken by Woo and Sriram (2006, 2007) and Rodríguez and Walker (2014), however, this is an important problem demanding further study.

Despite these issues, sample clusterings and estimates of the number of components or clusters can provide a useful tool for exploring complex datasets, particularly in the case of high-dimensional data that cannot easily be visualized. It should always be borne in mind, though, that the results can be interpreted as being correct only to the extent that the model assumptions are correct.

Acknowledgments

We are very grateful to Steve MacEachern for many helpful suggestions. This work was supported in part by the National Science Foundation (NSF) grants DMS-1007593 and DMS-1309004, by the National Institute of Mental Health (NIMH) grant R01MH102840, and by the Defense Advanced Research Projects Agency (DARPA) contract FA8650-11-1-715.

Appendix A. Formulas for some posterior quantities

Below are some details regarding computation of the posterior on k and of the density estimates.

Posterior on the number of components k. The posterior on $t = |\mathcal{C}|$ is easily estimated from posterior samples of $\mathcal{C}$. To compute the MFM posterior on k, note that

$$p(k|x_{1:n}) = \sum_{t=1}^{\infty} p(k|t, x_{1:n}) p(t|x_{1:n}) = \sum_{t=1}^{n} p(k|t) p(t|x_{1:n}),$$

by Equation 3.8 and the fact that t cannot exceed n. Using this and the formula for $p(k|t)$ given by Equation 3.6, it is simple to transform our estimate of the posterior on t into an estimate of the posterior on k. For the DPM, the posterior on the number of components k is always trivially a point mass at infinity.

Density estimates. Using the restaurant process (Theorem 4.1), it is straightforward to show that if $\mathcal{C}$ is a partition of $[n]$ and $\phi = (\phi_c : c \in \mathcal{C})$ then

$$p(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}) \propto \frac{V_{n+1}(t+1)}{V_{n+1}(t)} \gamma\, m(x_{n+1}) + \sum_{c \in \mathcal{C}} (|c| + \gamma) f_{\phi_c}(x_{n+1}) \tag{A.1}$$

where $t = |\mathcal{C}|$, and, using the recursion for $V_n(t)$ (Equation 3.9), this is normalized when multiplied by $V_{n+1}(t)/V_n(t)$. Further,

$$p(x_{n+1} \mid \mathcal{C}, x_{1:n}) \propto \frac{V_{n+1}(t+1)}{V_{n+1}(t)} \gamma\, m(x_{n+1}) + \sum_{c \in \mathcal{C}} (|c| + \gamma) \frac{m(x_{c \cup \{n+1\}})}{m(x_c)}, \tag{A.2}$$

with the same normalization constant. Therefore, when the single-cluster marginals $m(x_c)$ can be easily computed, Equation A.2 can be used to estimate the posterior predictive density $p(x_{n+1}|x_{1:n})$ based on samples from $\mathcal{C} \mid x_{1:n}$. When $m(x_c)$ cannot be easily computed, Equation A.1 can be used to estimate $p(x_{n+1}|x_{1:n})$ based on samples from $\mathcal{C}, \phi \mid x_{1:n}$, along with samples $\theta_1, \ldots, \theta_N \overset{\text{iid}}{\sim} H$ to approximate $m(x_{n+1}) \approx \frac{1}{N} \sum_{i=1}^{N} f_{\theta_i}(x_{n+1})$. (Thanks to Steve MacEachern for pointing out how to handle $m(x_{n+1})$ here.)

The posterior predictive density is, perhaps, the most natural estimate of the density. However, following Green and Richardson (2001), a simpler way to obtain a natural estimate is by assuming that element $n+1$ is added to an existing cluster; this will be very similar to the posterior predictive density when n is sufficiently large. To this end, we define $p_*(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}) = p(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}, |\mathcal{C}_{n+1}| = |\mathcal{C}|)$, where $\mathcal{C}_{n+1}$ is the partition of $[n+1]$, and observe that

$$p_*(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}) = \sum_{c \in \mathcal{C}} \frac{|c| + \gamma}{n + \gamma t} f_{\phi_c}(x_{n+1})$$

where $t = |\mathcal{C}|$ (Green and Richardson, 2001). Using this, we can estimate the density by

$$\frac{1}{N} \sum_{i=1}^{N} p_*(x_{n+1} \mid \mathcal{C}^{(i)}, \phi^{(i)}, x_{1:n}), \tag{A.3}$$

where $(\mathcal{C}^{(1)}, \phi^{(1)}), \ldots, (\mathcal{C}^{(N)}, \phi^{(N)})$ are samples from $\mathcal{C}, \phi \mid x_{1:n}$. The corresponding expressions for the DPM are all very similar, using its restaurant process instead. The density estimates shown in this paper are obtained using this approach.

These formulas are conditional on additional parameters such as γ for the MFM, and α for the DPM. If priors are placed on such parameters and they are sampled along with $\mathcal{C}$ and ϕ given $x_{1:n}$, then the posterior predictive density can be estimated using the same formulas as above, but also using the posterior samples of these additional parameters.

Appendix B. Proofs

Proof of Theorem 3.1. Letting $E_i = \{j : z_j = i\}$, and writing $\mathcal{C}(z)$ for the partition induced by $z = (z_1, \ldots, z_n)$, by Dirichlet-multinomial conjugacy we have

$$p(z|k) = \int p(z|\pi) p(\pi|k) d\pi = \frac{\Gamma(k\gamma)}{\Gamma(\gamma)^k} \frac{\prod_{i=1}^{k} \Gamma(|E_i| + \gamma)}{\Gamma(n + k\gamma)} = \frac{1}{(k\gamma)^{(n)}} \prod_{c \in \mathcal{C}(z)} \gamma^{(|c|)},$$

for $z \in [k]^n$, provided that $p_K(k) > 0$. It follows that for any partition $\mathcal{C}$ of $[n]$,

$$\begin{aligned} p(\mathcal{C}|k) &= \sum_{z \in [k]^n \,:\, \mathcal{C}(z) = \mathcal{C}} p(z|k) \\ &= \#\Big\{ z \in [k]^n : \mathcal{C}(z) = \mathcal{C} \Big\} \frac{1}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \\ &= \frac{k_{(t)}}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)}, \end{aligned} \tag{B.1}$$

where $t = |\mathcal{C}|$, since $\#\big\{ z \in [k]^n : \mathcal{C}(z) = \mathcal{C} \big\} = \binom{k}{t} t! = k_{(t)}$. Finally,

$$p(\mathcal{C}) = \sum_{k=1}^{\infty} p(\mathcal{C}|k) p_K(k) = \Big(\prod_{c \in \mathcal{C}} \gamma^{(|c|)} \Big) \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} \, p_K(k) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)},$$

with $V_n(t)$ as in Equation 3.2. □

Proof of Equation 3.3. Theorem 3.1 shows that the distribution of $\mathcal{C}$ is as shown. Next, note that instead of sampling only $\theta_1, \ldots, \theta_k \overset{\text{iid}}{\sim} H$ given $K = k$, we could simply sample $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} H$ independently of K, and the distribution of $X_{1:n}$ would be the same. Now, $Z_{1:n}$ determines which subset of the i.i.d. variables $\theta_1, \theta_2, \ldots$ will actually be used, and the indices of this subset are independent of $\theta_1, \theta_2, \ldots$; hence, denoting these random indices $I_1 < \cdots < I_T$, we have that $\theta_{I_1}, \ldots, \theta_{I_T} | Z_{1:n}$ are i.i.d. from H. For $c \in \mathcal{C}$, let $\phi_c = \theta_{I_i}$ where i is such that $c = \{j : z_j = I_i\}$. This completes the proof. □

Proof of the properties in Section 3.1. Abbreviate $x = x_{1:n}$, $z = z_{1:n}$, and $\theta = \theta_{1:k}$, and assume $p(z, k) > 0$. Letting $E_i = \{j : z_j = i\}$, we have $p(x|\theta, z, k) = \prod_{i=1}^{k} \prod_{j \in E_i} f_{\theta_i}(x_j)$ and

$$\begin{aligned} p(x|z,k) &= \int_{\Theta^k} p(x|\theta, z, k) p(d\theta|k) = \prod_{i=1}^{k} \int_{\Theta} \Big[\prod_{j \in E_i} f_{\theta_i}(x_j) \Big] H(d\theta_i) \\ &= \prod_{i=1}^{k} m(x_{E_i}) = \prod_{c \in \mathcal{C}(z)} m(x_c). \end{aligned}$$

Since this last expression depends only on z, k through $\mathcal{C} = \mathcal{C}(z)$, we have $p(x|\mathcal{C}) = \prod_{c \in \mathcal{C}} m(x_c)$, establishing Equation 3.4. Next, recall that $p(\mathcal{C}|k) = \frac{k_{(t)}}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)}$ (where $t = |\mathcal{C}|$) from Equation B.1, and thus

$$p(t|k) = \sum_{\mathcal{C} : |\mathcal{C}| = t} p(\mathcal{C}|k) = \frac{k_{(t)}}{(\gamma k)^{(n)}} \sum_{\mathcal{C} : |\mathcal{C}| = t} \prod_{c \in \mathcal{C}} \gamma^{(|c|)},$$

(where the sum is over partitions $\mathcal{C}$ of $[n]$ such that $|\mathcal{C}| = t$) establishing Equation 3.5. Equation 3.6 follows, since

$$p(k|t) \propto p(t|k) p(k) \propto \frac{k_{(t)}}{(\gamma k)^{(n)}} \, p_K(k),$$

(provided $p(t) > 0$) and the normalizing constant is precisely $V_n(t)$. To see that $\mathcal{C} \perp K \mid T$ (Equation 3.7), note that if $t = |\mathcal{C}|$ then

$$p(\mathcal{C}|t,k) = \frac{p(\mathcal{C},t|k)}{p(t|k)} = \frac{p(\mathcal{C}|k)}{p(t|k)},$$

(provided $p(t,k) > 0$) and due to the form of $p(\mathcal{C}|k)$ and $p(t|k)$ just above, this quantity does not depend on k; hence, $p(\mathcal{C}|t,k) = p(\mathcal{C}|t)$. To see that $X \perp K \mid T$ (Equation 3.8), note that $X \perp K \mid \mathcal{C}$; using this in addition to $\mathcal{C} \perp K \mid T$, we have

$$p(x|t,k) = \sum_{\mathcal{C}:|\mathcal{C}|=t} p(x|\mathcal{C},t,k)p(\mathcal{C}|t,k) = \sum_{\mathcal{C}:|\mathcal{C}|=t} p(x|\mathcal{C},t)p(\mathcal{C}|t) = p(x|t).$$

□

Proof of Theorem 4.1. Let $\mathcal{C}_\infty$ be the random partition of $\mathbb{Z}_{>0}$ as in Section 3.3, and for $n \in \{1,2,\dots\}$, let $\mathcal{C}_n$ be the partition of $[n]$ induced by $\mathcal{C}_\infty$. Then

$$p(\mathcal{C}_n|\mathcal{C}_{n-1},\dots,\mathcal{C}_1) = p(\mathcal{C}_n|\mathcal{C}_{n-1}) \propto q_n(\mathcal{C}_n)\, I(\mathcal{C}_n \setminus n = \mathcal{C}_{n-1}),$$

where $\mathcal{C} \setminus n$ denotes $\mathcal{C}$ with element n removed, and $I(\cdot)$ is the indicator function ($I(E) = 1$ if E is true, and $I(E) = 0$ otherwise). Recalling that $q_n(\mathcal{C}_n) = V_n(|\mathcal{C}_n|)\prod_{c\in\mathcal{C}_n}\gamma^{(|c|)}$ (Equation 3.1), we have, letting $t = |\mathcal{C}_{n-1}|$,

$$p(\mathcal{C}_n|\mathcal{C}_{n-1}) \propto \begin{cases} V_n(t+1)\gamma & \text{if } n \text{ is a singleton in } \mathcal{C}_n, \text{ i.e., } \{n\} \in \mathcal{C}_n \\ V_n(t)(\gamma + |c|) & \text{if } c \in \mathcal{C}_{n-1} \text{ and } c \cup \{n\} \in \mathcal{C}_n, \end{cases}$$

for $\mathcal{C}_n$ such that $\mathcal{C}_n \setminus n = \mathcal{C}_{n-1}$ (and $p(\mathcal{C}_n|\mathcal{C}_{n-1}) = 0$ otherwise). With probability 1, $q_{n-1}(\mathcal{C}_{n-1}) > 0$, thus $V_{n-1}(t) > 0$ and hence also $V_n(t) > 0$, so we can divide through by $V_n(t)$ to get the result. □

Proof of Theorem 4.2. Let $G \sim \mathcal{M}(p_K, \gamma, H)$ and let $\beta_1,\dots,\beta_n \overset{\text{iid}}{\sim} G$, given G. Then the joint distribution of $(\beta_1,\dots,\beta_n)$ (with G marginalized out) is the same as $(\theta_{Z_1},\dots,\theta_{Z_n})$ in the original model (Equation 2.1). Let $\mathcal{C}_n$ denote the partition induced by $Z_1,\dots,Z_n$ as usual, and for $c \in \mathcal{C}_n$, define $\phi_c = \theta_I$ where I is such that $c = \{j : Z_j = I\}$; then, as in the proof of Equation 3.3, $(\phi_c : c \in \mathcal{C}_n)$ are i.i.d. from H, given $\mathcal{C}_n$.

Therefore, we have the following equivalent construction for $(\beta_1,\dots,\beta_n)$:

$$\begin{aligned} &\mathcal{C}_n \sim q_n, \text{ with } q_n \text{ as in Section 3.3} \\ &\phi_c \overset{\text{iid}}{\sim} H \text{ for } c \in \mathcal{C}_n, \text{ given } \mathcal{C}_n \\ &\beta_j = \phi_c \text{ for } j \in c,\ c \in \mathcal{C}_n, \text{ given } \mathcal{C}_n, \phi. \end{aligned}$$

Due to the self-consistency property of $q_1, q_2, \dots$ (Proposition 3.3), we can sample $\mathcal{C}_n, (\phi_c : c \in \mathcal{C}_n), \beta_{1:n}$ sequentially for $n = 1, 2, \dots$ by sampling from the restaurant process for $\mathcal{C}_n|\mathcal{C}_{n-1}$, sampling $\phi_{\{n\}}$ from H if n is placed in a cluster by itself (or setting $\phi_{c\cup\{n\}} = \phi_c$ if n is added to $c \in \mathcal{C}_{n-1}$), and setting β_n accordingly.

In particular, if the base measure H is continuous, then the ϕ's are distinct with probability 1, so conditioning on $\beta_{1:n-1}$ is the same as conditioning on $\mathcal{C}_{n-1}, (\phi_c : c \in \mathcal{C}_{n-1}), \beta_{1:n-1}$, and hence we can sample $\beta_n|\beta_{1:n-1}$ in the same way as was just described. In view of the form of the restaurant process (Theorem 4.1), the result follows. □

We use the following elementary result in the proof of Theorem 5.1; it is a special case of the dominated convergence theorem.

Proposition B.1. *For $j = 1, 2, \ldots$, let $a_{1j} \geq a_{2j} \geq \cdots \geq 0$ such that $a_{ij} \to 0$ as $i \to \infty$. If $\sum_{j=1}^{\infty} a_{1j} < \infty$ then $\sum_{j=1}^{\infty} a_{ij} \to 0$ as $i \to \infty$.*

Proof of Theorem 5.1. For any $x > 0$, writing $x^{(n)}/n! = \Gamma(x+n)/(n!\,\Gamma(x))$ and using Stirling's approximation, we have

$$\frac{x^{(n)}}{n!} \sim \frac{n^{x-1}}{\Gamma(x)}$$

as $n \to \infty$. Therefore, the $k = t$ term of $V_n(t)$ is

$$\frac{t_{(t)}}{(\gamma t)^{(n)}}\, p_K(t) \sim \frac{t!}{n!}\, \frac{\Gamma(\gamma t)}{n^{\gamma t - 1}}\, p_K(t).$$

The first $t-1$ terms of $V_n(t)$ are 0, so to prove the result, we need to show that the rest of the series, divided by the $k = t$ term, goes to 0. (Recall that we have assumed $p_K(t) > 0$.) To this end, let

$$b_{nk} = (\gamma t)^{(n)}\, \frac{k_{(t)}}{(\gamma k)^{(n)}}\, p_K(k).$$

We must show that $\sum_{k=t+1}^{\infty} b_{nk} \to 0$ as $n \to \infty$. We apply Proposition B.1 with $a_{ij} = b_{t+i,t+j}$. For any $k > t$, $b_{1k} \geq b_{2k} \geq \cdots \geq 0$. Further, for any $k > t$,

$$\frac{(\gamma t)^{(n)}}{(\gamma k)^{(n)}} \sim \frac{n^{\gamma t - 1}}{\Gamma(\gamma t)}\, \frac{\Gamma(\gamma k)}{n^{\gamma k - 1}} \longrightarrow 0$$

as $n \to \infty$, hence, $b_{nk} \to 0$ as $n \to \infty$ (for any $k > t$). Finally, observe that $\sum_{k=t+1}^{\infty} b_{nk} \leq (\gamma t)^{(n)} V_n(t) < \infty$ for any $n \geq t$. Therefore, by Proposition B.1, $\sum_{k=t+1}^{\infty} b_{nk} \to 0$ as $n \to \infty$. This proves the result. □

Proof of Theorem 5.2. For any $t \in \{1, \ldots, k\}$,

$$p_n(K = t \mid T = t) = \frac{1}{V_n(t)}\, \frac{t_{(t)}}{(\gamma t)^{(n)}}\, p_K(t) \longrightarrow 1 \tag{B.2}$$

as $n \to \infty$ (where p_n denotes the MFM distribution with n samples), by Equation 3.6 and Theorem 5.1. For any $n \geq k$,

$$p(K = k \mid x_{1:n}) = \sum_{t=1}^{k} p(K = k \mid T = t, x_{1:n})\, p(T = t \mid x_{1:n}),$$

and note that by Equations 3.8 and B.2, $p(K = k \mid T = t, x_{1:n}) = p_n(K = k \mid T = t) \longrightarrow I(k = t)$ for $t \leq k$. The result follows. □

References

D. J. Aldous. *Exchangeability and related topics*. Springer, 1985.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47, 2001.

D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

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

T. Broderick, M. I. Jordan, and J. Pitman. Beta processes, stick-breaking and power laws. *Bayesian Analysis*, 7(2):439–476, 2012.

C. A. Bush and S. N. MacEachern. A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.

A. Cerquetti. Generalized Chinese restaurant construction of exchangeable Gibbs partitions and related results. *arXiv:0805.3853*, 2008.

A. Cerquetti. Conditional α-diversity for exchangeable Gibbs partitions driven by the stable subordinator. *arXiv:1105.0892*, 2011.

Y. Chung and D. B. Dunson. Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104(488), 2009.

D. B. Dahl. An improved merge-split sampler for conjugate Dirichlet process mixture models. *Technical Report, Department of Statistics, University of Wisconsin – Madison*, 2003.

D. B. Dahl. Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11, 2005.

D. B. Dahl. Modal clustering in a class of product partition models. *Bayesian Analysis*, 4 (2):243–264, 2009.

M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9(1):497, 2008.

D. B. Dunson and J.-H. Park. Kernel stick-breaking processes. *Biometrika*, 95(2):307–323, 2008.

R. Durrett. *Probability: Theory and Examples*, volume 2. Cambridge University Press, 1996.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

S. Favaro, A. Lijoi, and I. Pruenster. On the stick-breaking representation of normalized inverse Gaussian priors. *Biometrika*, 99(3):663–674, 2012.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

S. Ghosal and A. Van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007.

A. Gnedin. A species sampling model with finitely many types. *Elect. Comm. Probab.*, 15: 79–88, 2010.

A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685, 2006.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

P. J. Green and S. Richardson. Modeling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375, June 2001.

J. E. Griffin and M. J. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.

B. Hansen and J. Pitman. Prediction rules for exchangeable sequences related to species sampling. *Statistics & Probability Letters*, 46(3):251–256, 2000.

J. A. Hartigan. Partition models. *Communications in Statistics – Theory and Methods*, 19 (8):2745–2756, 1990.

J. Henna. On estimating of the number of constituents of a finite mixture of continuous distributions. *Annals of the Institute of Statistical Mathematics*, 37(1):235–240, 1985.

J. Henna. Estimation of the number of components of finite mixtures of multivariate distributions. *Annals of the Institute of Statistical Mathematics*, 57(4):655–664, 2005.

N. L. Hjort. Bayesian analysis for a generalised Dirichlet process prior. *Technical Report, University of Oslo*, 2000.

M.-W. Ho, L. F. James, and J. W. Lau. Gibbs partitions (EPPF's) derived from a stable subordinator are Fox H and Meijer G transforms. *arXiv:0708.0619*, 2007.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453), 2001.

H. Ishwaran and L. F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13(4):1211–1236, 2003.

H. Ishwaran and M. Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

H. Ishwaran, L. F. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96(456), 2001.

S. Jain and R. M. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1), 2004.

S. Jain and R. M. Neal. Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472, 2007.

L. F. James, C. E. Priebe, and D. J. Marchette. Consistent estimation of mixture complexity. *The Annals of Statistics*, pages 1281–1296, 2001.

A. Jasra, C. Holmes, and D. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhya Ser. A*, 62(1): 49–66, 2000.

W. Kruijer, J. Rousseau, and A. Van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

B. G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20 (3):1350–1360, 1992.

A. Lijoi and I. Prünster. Models beyond the Dirichlet process. *Bayesian Nonparametrics*, 28:80, 2010.

A. Lijoi, R. H. Mena, and I. Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291, 2005.

A. Lijoi, R. H. Mena, and I. Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007.

A. Lijoi, I. Prünster, and S. G. Walker. Bayesian nonparametric estimators derived from conditional Gibbs structures. *The Annals of Applied Probability*, 18(4):1519–1547, 2008.

J. S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics – Simulation and Computation*, 23(3):727–741, 1994.

S. N. MacEachern. Computational methods for mixture of Dirichlet process models. In *Practical nonparametric and semiparametric Bayesian statistics*, pages 23–43. Springer, 1998.

S. N. MacEachern. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55, 1999.

S. N. MacEachern. Dependent Dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, 2000.

S. N. MacEachern and P. Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

G. J. McLachlan, R. W. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422, 2002.

P. D. McNicholas and T. B. Murphy. Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*, 26(21):2705–2712, 2010.

M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.

M. Medvedovic, K. Y. Yeung, and R. E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

J. W. Miller. *Nonparametric and Variable-Dimension Bayesian Mixture Models: Analysis, Comparison, and New Methods*. PhD thesis, Division of Applied Mathematics, Brown University, 2014.

J. W. Miller and M. T. Harrison. Inconsistency of Pitman–Yor process mixtures for the number of components. *Journal of Machine Learning Research*, 15:3333–3370, 2014.

P. Müller and F. Quintana. Random partition models with regression on covariates. *Journal of Statistical Planning and Inference*, 140(10):2801–2808, 2010.

P. Müller, F. Quintana, and G. L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1), 2011.

R. M. Neal. Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*, pages 197–211. Springer, 1992.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

X. L. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

A. Nobile. *Bayesian Analysis of Finite Mixture Distributions.* PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 1994.

A. Nobile. Bayesian finite mixtures: a note on prior specification and posterior computation. *Technical Report, Department of Statistics, University of Glasgow*, 2005.

A. Nobile and A. T. Fearnside. Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162, 2007.

M. Pagel and A. Meade. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology*, 53(4):571–581, 2004.

J. W. Paisley, A. K. Zaas, C. W. Woods, G. S. Ginsburg, and L. Carin. A stick-breaking construction of the beta process. In *Proceedings of the 27th International Conference on Machine Learning*, pages 847–854, 2010.

O. Papaspiliopoulos and G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

J.-H. Park and D. B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, 20:1203–1226, 2010.

D. B. Phillips and A. F. M. Smith. Bayesian model comparison via jump diffusions. In *Markov chain Monte Carlo in Practice*, pages 215–239. Springer, 1996.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.

J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996.

J. Pitman. *Combinatorial Stochastic Processes*. Springer–Verlag, Berlin, 2006.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

F. A. Quintana and P. L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574, 2003.

C. E. Rasmussen, B. J. de la Cruz, Z. Ghahramani, and D. L. Wild. Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):615–628, 2009.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, 2000.

S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997.

A. Rodriguez and D. B. Dunson. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1), 2011.

C. E. Rodríguez and S. G. Walker. Univariate Bayesian nonparametric mixture modeling with unimodal kernels. *Statistics and Computing*, 24(1):35–49, 2014.

K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

J. Sethuraman and R. C. Tiwari. Convergence of Dirichlet measures and the interpretation of their parameter. *Technical Report, Department of Statistics, Florida State University*, 1981.

C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2. IEEE, 1999.

M. Stephens. Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74, 2000.

Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 556–563, 2007.

R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 564–571, 2007.

S. G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics – Simulation and Computation*, 36(1):45–54, 2007.

M. West. Hyperparameter estimation in Dirichlet process mixture models. *ISDS Discussion Paper #92-A03, Duke University*, 1992.

M. West, P. Müller, and M. D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. In P. Freeman and A. F. Smith, editors, *Aspects of Uncertainty: A Tribute to D.V. Lindley*, pages 363–386. Wiley, 1994.

M.-J. Woo and T. N. Sriram. Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101(476), 2006.

M.-J. Woo and T. N. Sriram. Robust estimation of mixture complexity for count data. *Computational Statistics and Data Analysis*, 51(9):4379–4392, 2007.

K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.