*Submitted to the Annals of Statistics*

# ON POSTERIOR INFERENCE FOR THE NUMBER OF CLUSTERS IN DIRICHLET PROCESS MIXTURE MODELS

By Chiao-Yu Yang[†], Eric Xia[†], Nhat Ho[‡], and Michael I. Jordan[*,†]

*University of California, Berkeley*[†]
*University of Texas, Austin*[‡]

Dirichlet process mixture models (DPMMs) and Pitman-Yor process mixture models (PYPMMs) have been widely used in density estimation, mode detection, and other nonparametric estimation problems. Due to their discrete, combinatorial nature of the underlying random measures, these models have also been suggested for inference in clustering problems, but it has been shown that these methods are inconsistent in the number of components. We contribute to this line of work by giving quantitative statements about the clustering behavior. Specifically, we provide nonasymptotic and asymptotic lower bounds on the ratio between the posterior probabilities of consecutive number of clusters in these models under different choices of prior for the parameters.

**1. Introduction.** In the wake of the seminal work of Ferguson [2, 4], a line of research on Bayesian nonparametric inference emerged that was based on the use of the Dirichlet process and related combinatorial stochastic processes, such as the Pitman-Yor process [23], as prior distributions. Most commonly, these priors are used in the context of mixture models, where both the prior and the posterior place probability mass on an unbounded number of mixture components [1].

Statistical applications for these models include a variety of problems in density estimation [3, 6, 7, 8, 9, 10, 11, 26, 28] and parameter estimation [5, 22, 24, 27]. Further applications can be found in engineering and scientific fields such as computer vision, information retrieval, economics, astronomy, molecular biology, and genetics [3, 14, 15, 18, 21, 24, 27]. In these latter applications, the problem is often one of clustering, and Bayesian nonparametric mixture models are viewed as a substitute for classical methods such as $K$-means, which generally require the number of components $K$ to be known a priori. Particularly in the setting of dynamically growing data

[*] Supported by Army Research Office grant W911NF-17-1-0304

*MSC 2010 subject classifications:* Primary 00X00, 00X00; secondary 00X00

*Keywords and phrases:* Bayesian nonparametrics, posterior distribution, mixture models, number of clusters

sets, it is difficult and even unreasonable to fix the number of clusters a priori, and Bayesian nonparametric mixture models have been considered as a natural way to circumvent this issue since they allow the number of clusters to be random and subject to posterior inference.

However, there is little theoretical support for this methodology. Unlike the case of density estimation, results on the convergence of the number of components in Bayesian nonparametric mixtures are largely absent from the literature. Indeed, it has been demonstrated that under certain parametric assumptions, these models exhibit posterior inconsistencies in the number of clusters when the underlying data-generating distribution satisfies some mild conditions [19, 20]. Moreover, in practice, it has been observed that inference based on Dirichlet process or Pitman-Yor process mixture models can generate small clusters that do not reflect the underlying data-generating process, a serious concern when the real number of components is small.

We study the posterior distribution of the number of clusters of Bayesian clustering models based on the Dirichlet process and the Pitman-Yor process. In contrast to previous work on this topic, we do not assume the underlying data-generating process to be be a specific parametric family or even a finite mixture model, so our results apply to nonparametric data-generating processes as well. In this general setting, there is no analytic form for the posterior distribution; nonetheless, we have obtained an analytical characterization of the ratio between the probability of obtaining $k+1$ clusters and that of obtaining $k$ clusters, for any positive $k$. We provide novel lower bounds on these ratios, denoted $R(s|\{x_i\}_{i=1}^n)$, where $x_1, \dots, x_n$ are exchangeable observations. We study the setting where the prior on the parameters is either Gaussian or uniform. Our results are as follows:

- When the prior on the parameters is uniform over a bounded subset of $\mathbb{R}$, denoted $\Theta$, the lower bound on $R(s|\{x_i\}_{i=1}^n)$ is of order $\frac{\alpha}{s|\Theta|}$, under a Dirichlet process mixture model, where $\alpha$ is the scale parameter of the Dirichlet process prior. In the case of a Pitman-Yor process mixture model, the lower bound becomes $\frac{\alpha}{s|\Theta|} + \frac{\xi}{|\Theta|}$ up to a universal constant, where $\xi$ denotes the positive discount parameter. These bounds are nonasymptotic.
- When the prior on the parameters is Gaussian, $\mathcal{N}(0, \sigma^2)$, the asymptotic lower bound on $R(s|\{x_i\}_{i=1}^n)$ is of order $\frac{\alpha}{s^2(1+\sigma)}$ under a Dirichlet process mixture model and of order $\frac{\alpha}{s^2(1+\sigma)} + \frac{\xi}{s(1+\sigma)}$ under a Pitman-Yor process mixture model.

These lower bounds provide a fine-grained understanding of the posterior distribution induced by the Dirichlet process and the Pitman-Yor process on the number of clusters. We note, however, that in general there does not exist a finite upper bound. Such an upper bound, should it exist, would require additional assumptions on the data-generating process.

Besides the lower bounds we demonstrate in this paper, we note that one can use our techniques to obtain results for more specific cases where additional assumptions on the data-generating process are imposed.

The remainder of the paper is organized as follows. We begin in Section 2 with background on Dirichlet process and Pitman-Yor process mixture models. Section 3 is devoted to a study of the posterior distribution of the number of clusters from Dirichlet process mixture models under two choices of priors on the parameters: uniform and Gaussian priors. In Section 4, we extend these results to the case of Pitman-Yor process mixture models. Section 5 presents the results of experiments that explore some of the consequences of the theoretical results in previous sections. Finally, we conclude the paper with a discussion of potential future work in Section 6.

*Notation:* We use $\{x_i\}_{i=1}^n$ to denote the sample $\{x_1, \cdots, x_n\}$, $[n]$ to denote the set $\{1, \cdots, n\}$, and $A^{n,s} \in \rho_s(n)$ to denote a set $\{A_1^{n,s}, \cdots, A_s^{n,s}\}$ such that all of the elements $A_i$ form an $s$-partition of $[n]$, where $\rho_s(n)$ is the set of all $s$-partitions on $[n]$. For example, $\{\{1, 2\}, \{3, 4\}\}$ is in $\rho_2(4)$.

## 2. Preliminaries.

In this section, we provide necessary background on Dirichlet process mixture models and establish key notation for our subsequent analysis of the posterior distribution of the number of clusters in these models.

Dirichlet process mixture models (DPMMs) [1, 16] are specified as follows:

$$p(A^{n,s}) := \frac{\alpha^s}{\alpha^{(n)}} \prod_{i=1}^{s} (|A_i^{n,s}| - 1)! \tag{1}$$

$$p(\theta | A^{n,s}) := \prod_{i=1}^{s} \pi(\theta_i) \tag{2}$$

$$p\big(\{x_i\}_{i=1}^n \big| \{\theta_j\}_{j=1}^s, A^{n,s}\big) := \prod_{j=1}^{s} \prod_{x_i \in A_j^{n,s}} f_{\theta_j}(x_i), \tag{3}$$

where $\pi$ is a prior on the parameter $\theta$ and $\{f_\theta(\cdot)\}$ is a family of density functions.

In this paper, we focus on the application of DPMMs to clustering problems. The prior for the number of clusters with a data set of size $n$ is as follows:

$$\mathbb{P}(K_n = s) = \sum_{A^{n,s} \in \rho_s(n)} p(A^{n,s}).$$

Given this prior distribution, the posterior distribution for the number of clusters admits the following formulation:

$$
\begin{aligned}
&\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n) \\
&\quad = \frac{\mathbb{P}(\{x_i\}_{i=1}^n|K_n = s)\mathbb{P}(K_n = s)}{\mathbb{P}(\{x_i\}_{i=1}^n)} \\
&\quad \propto \sum_{A \in \rho_s(n)} p(A^{n,s}) \cdot \int_{\{\theta_j\}_{j=1}^s} p(\{x_i\}_{i=1}^n|\{\theta_j\}_{j=1}^s) p(\{\theta_j\}_{j=1}^s|A^{n,s}) d\{\theta_j\}_{j=1}^s \\
&\quad = \sum_{A \in \rho_s(n)} p(A^{n,s}) \cdot \int_{\{\theta_j\}_{j=1}^s \in \Theta^s} \Big( \prod_{j=1}^s \prod_{x_i \in A_j} f_{\theta_j}(x_i) \prod_{j=1}^s \pi(\theta_j) \Big) d\{\theta_j\}_{j=1}^s.
\end{aligned}
$$

To ease the ensuing discussion, we use $m(x_{A_j^{n,s}})$ to denote the probability density function of a cluster of samples, which is given by:

$$
m(x_{A_j^{n,s}}) = \int_{\theta_j} f_{\theta_j}(x_{j,1}) \cdots f_{\theta_j}(x_{j,a_j}) \pi(\theta_j) d\theta_j,
$$

where $x_{j,1}, \cdots, x_{j,a_j} \in A_j^{n,s}$ for all $1 \leq j \leq s$. Given this definition of $m(x_{A_j^{n,s}})$, we can rewrite $\mathbb{P}(K_n = s|x_n)$ as follows:

$$
\begin{aligned}
\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n) &\propto \sum_{A^{n,s} \in \rho_s(n)} \Big( p(A^{n,s}) \cdot \prod_{j=1}^s m(x_{A_j^{n,s}}) \Big) \\
&= \sum_{A^{n,s} \in \rho_s(n)} \Big( \frac{\alpha^s}{\alpha^{(n)}} \prod_{j=1}^s (|A_j^{n,s}| - 1)! \cdot \prod_{j=1}^s m(x_{A_j^{n,s}}) \Big). \qquad (4)
\end{aligned}
$$

Early theoretical results on these posterior probabilities can be found in [19, 20]. A general qualitative result from [20] is as follows.

Theorem 2.1. *For a Dirichlet process mixture model, if the component distribution is Gaussian, log normal, Gamma, exponential, Weibull, or if it is discrete and has at least some point with nonzero measure for any choice of parameter, then, provided that the data are independently and identically distributed from a $K^*$-component mixture model, ($K^*$ is bounded), and with probability one we have:*

$$
\limsup_{n \to \infty} \mathbb{P}(K_n = K^*|\{x_i\}_{i=1}^n) < 1.
$$

Our goal is to supply a quantitative perspective on these posterior probabilities, in a general setting where we do not confine ourselves to the case of finite or parametric mixtures. In other words, under more general assumptions on the data-generating process, we want to characterize the behavior of the posterior distribution of the number of clusters.

In order to obtain this characterization, we consider the following ratio between its values at $K_n = s+1$ and $K_n = s$:

$$\begin{aligned} R(s|\{x_i\}_{i=1}^n) &:= \frac{\mathbb{P}(K_n = s+1 \,|\, \{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s \,|\, \{x_i\}_{i=1}^n)} \\ &= \frac{\sum_{A^{n,s+1} \in \rho_{s+1}(n)} \left( p(A^{n,s+1}) \cdot \prod_{j=1}^{s+1} m(x_{A_j^{n,s}}) \right)}{\sum_{A^{n,s} \in \rho_s(n)} \left( p(A^{n,s}) \cdot \prod_{j=1}^{s} m(x_{A_j^{n,s}}) \right)}. \end{aligned}$$

From the formulation of a DPMM, we further obtain that

$$\begin{aligned} &R(s|\{x_i\}_{i=1}^n) \\ &= \alpha \cdot \frac{\sum_{A^{n,s+1} \in \rho_{s+1}(n)} \left( (\prod_{i=1}^{s+1} (|A_i^{n,s+1}| - 1)!) \cdot \prod_{j=1}^{s+1} m(x_{A_j^{n,s+1}}) \right)}{\sum_{A^{n,s} \in \rho_s(n)} \left( (\prod_{i=1}^{s} (|A_i^{n,s}| - 1)!) \cdot \prod_{j=1}^{s} m(x_{A_j^{n,s}}) \right)}. \end{aligned} \tag{5}$$

Our focus will be to obtain lower bounds on this ratio.

## 3. A Quantitative Result for the Posterior of the DPMM.

In this section, we study the posterior distribution on the number of clusters induced by a DPMM. Our first result assumes a uniform prior on a finite interval, with a boundedness assumption on the data. The second result is an asymptotic result for the Gaussian prior.

### 3.1. *Uniform prior.*

We first consider the posterior distribution of the number of clusters of a DPMM when the data lie in a bounded set. Data sets of this form often arise in fields such as biology, genetics, and economics. In this case it is natural to let the parameter space $\Theta$ be a compact set [25].

To ease the complexity of the algebraic aspects of the proof, we specifically consider a simple uniform prior on the parameter space $\Theta$, where $\Theta$ is a bounded segment of $\mathbb{R}$ with size $|\Theta|$. With this choice, we have $\pi(\theta) = 1/|\Theta|$ for all $\theta \in \Theta$. We begin with the following result that establishes a lower bound on the ratio $R(s|\{x_i\}_{i=1}^n)$.

Theorem 3.1. *Given the DPMM defined in Eq. (3), with a uniform prior* $\mathrm{Unif}(\Theta)$ *on* $\theta$*, for sufficiently large* $n$*, if* $\min(\{x_i\}_{i=1}^n) > \min(\Theta)$ *and* $\max(\{x_i\}_{i=1}^n) < \max(\Theta)$*, then the ratio* $R(s|\{x_i\}_{i=1}^n)$ *between consecutive terms is lower bounded by*

$$R(s|\{x_i\}_{i=1}^n) \geq C \cdot \frac{\alpha}{s|\Theta|}, \tag{6}$$

*where $C > 0$ is a universal constant.*

Remark. Note that in many problems when one applies a uniform prior for the parameters, one expects the uniform prior to have support that is large enough to capture the means of all the components. It is also possible to extend to the case where some samples lie outside $\Theta$ by modifying Lemma 3.2, but the result would then have to be dependent on the specifics of the distribution and lose generality.

Proof. Given $A^{n,s} \in \rho_s(n)$, we define its $(s+1)$-*component extension set* $\eta_{ext}(A^{n,s})$, and its $(s-1)$-*component contraction set* $\eta_{con}(A^{n,s})$ as follows:

$$\begin{aligned}\eta_{ext}(A^{n,s}) &:= \{\tilde{A}^{n,s+1} \in \rho_{s+1}(n) : A^{n,s} \in \eta_{con}(\tilde{A}^{n,s+1})\}\\ \eta_{con}(A^{n,s}) &:= \Big\{\tilde{A}^{n,s-1} \in \rho_{s-1}(n) : !\exists\, i,j \in [s] :\\ &\qquad \{A_1^{n,s}, \cdots, A_s^{n,s}, A_i^{n,s} \cup A_j^{n,s}\}\backslash\{A_i^{n,s}, A_j^{n,s}\} = \{\tilde{A}_1^{n,s-1}, \cdots, \tilde{A}_{s-1}^{n,s-1}\}\Big\},\end{aligned}$$

where we require $s > 1$ in the definition of $\eta_{con}(A^{n,s})$.

In words, $\eta_{con}(A^{n,s})$ is the set of partitions in $\rho_{s-1}(n)$ that can be obtained from $A^{n,s}$ by combining two elements of $A^{n,s}$ into one element while keeping everything else the same. On the other hand, $\eta_{ext}(A^{n,s})$ is the set of partitions in $\rho_{s+1}(n)$ for which we can combine two of its elements into one to get $A^{n,s}$. For example, if $s = 2$, and we have $A^{n,3} = \{A_1^{n,3}, A_2^{n,3}, A_3^{n,3}\}$, then $\eta_{con}(A^{n,3})$ consists of following three partitions: $\{A_1^{n,3}, A_2^{n,3} \cup A_3^{n,3}\}, \{A_1^{n,3} \cup A_3^{n,3}, A_2^{n,3}\}, \{A_1^{n,3} \cup A_2^{n,3}, A_3^{n,3}\}$, and $\eta_{ext}(A^{n,3})$ is the set of all 4-partitions that can contract to $A^{n,3}$.

We further define the posterior probabilities of a partition to be:

$$p(A^{n,s}, x) := p(A^{n,s}) \cdot \prod_{i=1}^{s} m(x_{A_i^{n,s}}).$$

Given these definitions, we claim that we can rewrite the ratio $R(s|\{x_i\}_{i=1}^n)$ as follows:

$$\begin{aligned}R(s|\{x_i\}_{i=1}^n) &= \frac{\mathbb{P}(K_n = s+1|\{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)}\\ &= \frac{2}{(s+1)s} \cdot \frac{\sum_{A^{n,s}\in\rho_s(n)} \Big(\sum_{A^{n,s+1}\in\eta_{ext}(A^{n,s})} p(A^{n,s+1}, x)\Big)}{\sum_{A^{n,s}\in\rho_s(n)} p(A^{n,s}, x)},\end{aligned} \tag{7}$$

for $n \geq s+1$. The proof of this claim is deferred to the end of the proof and we proceed while assuming this result. Note that the last fractional term in the claim takes the generic form $\frac{\sum_a f(a)}{\sum_a g(a)}$, which motivates us to study $f(a)/g(a)$ in the following.

For each partition in $\eta_{ext}(A^{n,s})$, there exist exactly two sets that can be combined to get one set in $A^{n,s}$, and all the others are the same as the other clusters of $A^{n,s}$. We now let $\eta_{ext}^{i,j}(A^{n,s})$ denote the subcollection of $\eta_{ext}(A^{n,s})$ such that for any partition in this subcollection, two sets can be combined into $A_i^{n,s}$, and that one set has size $j$. Since the order of sets in each partition does not matter, we will assume for simplicity that the two sets are the $i$-th and $(s+1)$-th clusters. To distinguish the $s$ partition and its induced $s+1$ partitions, we use $B^{n,s}$ to denote the $s$-partition, and $A^{n,s+1} \in \rho_{ext}(B^{n,s})$ to denote its induced partitions. To simplify the notation, we use $b_1, \ldots, b_s$ to denote the cardinalities of $B_1^{n,s}, \cdots, B_s^{n,s}$ and $a_1, \ldots, a_{s+1}$ to denote those of $A_1^{n,s+1}, \ldots, A_{s+1}^{n,s+1}$. Furthermore, in the following derivation, the value $n$ is fixed, so will omit it in the notation unless there is the necessity to discuss different $n$. We obtain the following:

$$\begin{aligned}
&\frac{\sum_{A^{s+1}\in\eta_{ext}(B^s)} p(A^{s+1}|x)}{p(B^s|x)} \\
&\quad= \sum_{A^{s+1}\in\eta_{ext}(B^s)} \alpha \cdot \frac{\prod_{i=1}^{s+1}(|A_i^{s+1}|-1)!\, m(x_{A_i^{s+1}})}{\prod_{i=1}^{s}(|B_i^s|-1)!\, m(x_{B_i^s})} \\
&\quad= \alpha \cdot \sum_{i=1}^{s}\sum_{j=1}^{b_i-1} \sum_{A^{s+1}\in\eta_{ext}^{i,j}(B^s)} \frac{(j-1)!(b_i-j-1)!\, m(x_{A_i^{s+1}}) m(x_{A_{s+1}^{s+1}})}{(b_i-1)!\, m(x_{B_i^s})}
\end{aligned}$$

$$(8) \qquad = \alpha \cdot \sum_{i=1}^{s}\sum_{j=1}^{b_i-1} \Big\{ \frac{(j-1)!(b_i-j-1)!}{(b_i-1)!} \Big( \sum_{A\in\eta_{ext}^{i,j}} \frac{m(X_{A_i^{s+1}}) m(X_{A_{s+1}^{s+1}})}{m(X_{B_i^s})} \Big) \Big\}.$$

Recall that, by definition, $A_i^{s+1} \cup A_{s+1}^{s+1} = B_i^s$.

Lemma 3.2. *For a set $x_S$ with $n_s$ elements and $x_T$ with $n_t$ elements, suppose $\Theta \subseteq [\theta_{\min}, \theta_{\max}]$ is an interval in $\mathbb{R}$ such that $\min x_S, \min x_T > \theta_{\min} + c$ and $\max x_S, \max x_T < \theta_{\max} - c$ for some $c > 0$, we have:*

$$\frac{m(x_S)m(x_T)}{m(x_S \cup x_T)} \geq \frac{c_1\sqrt{2\pi}}{|\Theta|} \cdot \frac{\sqrt{n_s + n_t}}{\sqrt{n_s\, n_t}},$$

*for some positive constant $c_1$ that does not depend on $x_S, x_T$ but only on $c$.*

Given Lemma 3.2, we can derive the following bounds for the term in (8):

$$(9) \qquad \frac{\sum_{A^{s+1}\in\eta_{ext}(B^s)} p(A^{s+1}|x)}{p(B^s x)} \geq \frac{\alpha\, c_1\sqrt{2\pi}}{|\Theta|} \cdot \sum_{i=1}^{s}\sum_{j=1}^{b_i-1} \Big(\frac{b_i}{j(b_i-j)}\Big)^{3/2}.$$

Note that:

$$\sum_{j=1}^{b_i-1}\left(\frac{b_i}{j(b_i-j)}\right)^{3/2} \geq \int_{x=1}^{b_i-1}\left(\frac{b_i}{x(b_i-x)}\right)^{3/2} dx \geq \frac{4(b_i-2)}{\sqrt{(b_i-1)b_i}}$$

The inequality implies that the leftmost term is no less than 2 for $b_i \geq 4$. When $b_i = 2, 3$, simple algebra indicates that $\sum_{j=1}^{b_i-1}\left(\frac{b_i}{j(b_i-j)}\right)^{3/2} \geq 2$. Invoking these results, we have the following lower bound:

$$\frac{\sum_{A^{s+1}\in\eta_{ext}(B^s)} p(A^{s+1}|x)}{p(B^s|x)} \geq \frac{\alpha\, c_1^2\sqrt{2\pi}}{|\Theta|} \cdot \sum_{i=1}^{s} 2 \cdot I_{b_i \geq 2} \gtrsim \frac{\alpha s}{|\Theta|}.$$

Combining this lower bound with Eq. (7), we obtain the following evaluation of the ratio between consecutive terms $R(s)$:

$$R(s|\{x_i\}_{i=1}^n) = \frac{\mathbb{P}(K_n = s+1|\{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)} \gtrsim \frac{\alpha}{s|\Theta|}.$$

As a consequence, we reach the conclusion of the theorem.

*Proof of claim* (7)*:* Using equation (4), we can rewrite the ratio between the posterior probability of $s+1$ components and that of $s$ components as follows:

$$\frac{\mathbb{P}(K_n = s+1|\{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)} = \frac{\sum_{A^{s+1}\in\rho_{s+1}(n)} p(A^{s+1}, x)}{\sum_{B^s\in\rho_s(n)} p(B^s, x)}.$$

Note that for each $A^{s+1} \in \rho_{s+1}(n)$, we can merge any two of its $s+1$ parts to obtain some $B^s \in \rho_s(n)$. The number of distinct ways to do so is exactly $\binom{s+1}{2}$. Also, for each $B^s \in \rho_s(n)$, the set $\eta_{ext}(B^s)$ contains all $A^{s+1} \in \rho_{s+1}(n)$ such that they can merge some parts to get $B^s$. Thus, the index of the numerator on the right-hand side counts each $A^{s+1} \in \rho_{s+1}(n)$ exactly $\binom{s+1}{2}$ times, from which the equation follows. □

*Proof of Lemma 3.2:* For a set $x_S$ with $n_s$ elements, we use the following shorthand:

$$\overline{x_S} = \frac{\sum_{X\in x_S} X}{n_s}; \quad S_{x_S}^2 = \sum_{X\in x_S} X^2.$$

After some algebra, we can verify that

$$\begin{aligned}
&\frac{m(x_S)m(x_T)}{m(x_S \cup x_T)} \\
&= \frac{\int_{\theta\in\Theta} \exp\left( - \sum_{X \in x_S} \frac{(X-\theta)^2}{2} \right) d\theta \cdot \int_{\theta\in\Theta} \exp\left( - \sum_{X \in x_T} \frac{(X-\theta)^2}{2} \right) d\theta}{|\Theta| \int_{\theta\in\Theta} \exp\left( - \sum_{X \in x_S \cup x_T} \frac{(X-\theta)^2}{2} \right) d\theta} \\
&= \frac{\frac{2\pi}{\sqrt{n_s n_t}} \exp\left( - \frac{(S^2_{x_S} + n_1 \overline{x_S}^2)}{2} - \frac{(S^2_{x_T} + n_t \overline{x_T}^2)}{2} \right) P_{x_S}(\Theta) P_{x_S \cup x_T}(\Theta)}{|\Theta| \sqrt{\frac{2\pi}{n_s + n_t}} \exp\left( - \frac{(S^2_{x_S \cup x_T} + (n_s + n_t)\overline{x_S \cup x_T}^2)}{2} \right)} \\
&= \frac{\sqrt{2\pi}}{|\Theta|} \cdot \frac{\sqrt{n_s + n_t}}{\sqrt{n_s\, n_t}} \cdot \exp\left( \frac{n_s \overline{x_S}^2 + n_t \overline{x_T}^2 - (n_s + n_t)\overline{x_S \cup x_T}^2}{2} \right) \\
&\qquad\qquad \times \frac{P_{x_S}(\Theta)\, P_{x_T}(\Theta)}{P_{x_S \cup x_T}(\Theta)},
\end{aligned}$$

where $P_x(\Theta) := P(\theta \in \Theta | \theta \sim N(\overline{x}, \frac{1}{|x|}))$, which can be shown to be at least $\mathrm{erf}(c|x|/\sqrt{2})$, where erf is the error function. Hence:

$$(\mathrm{erf}(c/\sqrt{2}))^2 < \frac{P_{x_1}(\Theta)\, P_{x_2}(\Theta)}{P_{x_1 \cup x_2}(\Theta)} < \frac{1}{\mathrm{erf}(c/\sqrt{2})}.$$

Note that this range in general is very small (around 1) for $c$ that is not too small. For example, the range above is $(0.997^2, 1/0.997)$ for $c = 3$. Now, we write $c_1 = (\mathrm{erf}(c/\sqrt{2}))^2$.

On the other hand,

$$\frac{n_s \overline{x_S}^2 + n_t \overline{x_T}^2 - (n_s + n_t)\overline{x_S \cup x_T}^2}{2} = \frac{n_s\, n_t(\overline{x_S} - \overline{x_T})^2}{2(n_s + n_t)} \geq 0.$$

Combining these results, we obtain the desired inequality. □

The bound in the result of Theorem 3.1 does not require the original distribution to be a mixture distribution. Instead, it holds provided that the true underlying distribution has finite and nonzero variance.

In particular, note that by the law of large numbers, if the observations in $x_S, x_T$ are i.i.d., we may consider the empirical average of the exponential term in Lemma 3.2 and Eq. (8):

$$\exp\left( n_s n_t (\overline{x_S} - \overline{x_T})^2 / 2(n_s + n_t) \right).$$

The term inside the exponential is approximately $\frac{\sigma\sqrt{n_s n_t}}{2\sqrt{n_s+n_t}}Z$-distributed as $n_s$ and $n_t$ go to infinity, where $Z$ is the standard normal distribution and $\sigma^2$ is the variance of the original distribution. Using the moment-generating function of the $\chi_2$ distribution, we can see that this term grows unbounded as $n_s$ and $n_t$ grow, in contrast to the constant value one we derived in the proof of the Theorem 3.1. Therefore, given the result of Theorem 3.1, we obtain the following corollary:

Corollary 3.3. *If the conditions in Theorem 3.1 hold as $n \to \infty$, then for any $s$ fixed, we have:*

$$\lim_{n\to\infty} R(s|\{x_i\}_{i=1}^n) \to \infty.$$

We note that the rate at which $R(s|\{x_i\}_{i=1}^n)$ grows is unknown and may be very slow when $\sigma^2$ is small, due to the complicated combinatorial behavior. Combining the results from Theorem 3.1 and Corollary 3.3, we can see that for any distribution with nonzero variance, the posterior probability of obtaining $s+1$ clusters will eventually exceed that of obtaining $s$ clusters, and their ratio grows in an unbounded way. An important implication of this result is that, with a larger sample size we will tend to fit a larger number of clusters, and this model ultimately leads to an infinite number of clusters almost surely. However, in the finite sample-size regime, the model's behavior depends more on the variance of the original distribution. In practice, one should rely on the finite bound in Theorem 3.1 unless additional information about the distribution is available.

3.2. *Gaussian Prior.* We now consider the more commonly used Gaussian prior on the parameter $\theta$ [3, 17]. In particular, we choose the prior density to be that of a univariate Gaussian distribution, $\mathcal{N}(0,\sigma^2)$, with fixed variance $\sigma > 0$. Given this choice of prior, we have the following asymptotic lower bound on $R(s|\{x_i\}_{i=1}^n)$:

Theorem 3.4. *For the DPMM defined in Eq. (3), with a Gaussian prior $\mathcal{N}(0,\sigma^2)$ on $\theta$, as $n$ goes to infinity, the ratio $R(s|\{x_i\}_{i=1}^n)$ satisfies the following asymptotic lower bound:*

$$\lim_{n\to\infty} R(s|\{x_i\}_{i=1}^n) \geq C \cdot \frac{\alpha}{s^2(1+\sigma)}, \tag{10}$$

*where $C > 0$ is a universal constant.*

Remark. The result of Theorem 3.4 holds asymptotically. Its performance in finite samples is unknown and is complex, depending heavily on the original distribution. It would be of interest to characterize the finite-sample behavior of distributions satisfying certain conditions on variance and the true number of components or the true rate of growth in the number of components for an infinite-component distribution induced by processes such as the Dirichlet process.

Proof. We use the notation from the proof of Theorem 3.1 in this proof. As in Eq. 8, we study the following term: $(m(x_{A_i^{s+1}})m(x_{A_{s+1}^{s+1}}))/m(x_{B_i^s})$, where as in Theorem 3.1 we assume $B^s$ to be some $s$-cluster formed by merging two clusters in $A^{s+1}$.

Lemma 3.5. *For sets $x_S$ with $n_s$ elements and $x_T$ with $n_t$ elements, suppose $\Theta = \mathbb{R}$ and $\pi(\theta) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{\theta^2}{2})$. We have:*

$$\frac{m(x_S)m(x_T)}{m(x_S \cup x_T)} \geq \sqrt{\frac{1}{2} \cdot \frac{\tau}{1+\tau} \cdot \frac{n_s + n_t}{n_s n_t}},$$

*with probability approaching one as $n_s + n_t$ goes to infinity.*

Returning to the computation of the ratio $p(A|x)/p(B|x)$, for fixed $s$ and a partition $B \in \rho_s(n)$, we define

$$U(B) := \{i \in [s] : b_i \geq \frac{n}{s^2}\}.$$

For any $i \in U(B)$, since $b_i$ increases as $n$ increases, we may assume that the aforementioned condition that $F(\tau; x_{B_1}, x_{B_2}) \geq 0$ holds asymptotically for any fixed proportion (less than one) for all the partitions of $B_i$. Then, for sufficiently large $n$ we have:

$$\begin{aligned}
&\sum_{A^{s+1} \in \eta_{ext}(B^s)} \frac{p(A^{s+1}|x)}{p(B^s|x)} \\
&\quad = \alpha \cdot \sum_{i=1}^{s} \sum_{j=1}^{b_i - 1} \frac{(j-1)!(b_i - j - 1)!}{(b_i - 1)!} \Big( \sum_{A^{s+1} \in \eta_{ext}^{i,j}(B^s)} \frac{m(X_{A^{s+1}{s_i}})m(X_{{}^{s+1}A_{s+1}})}{m(X_{B_i^s})} \Big) \\
&\quad \overset{w.h.p.}{\geq} \frac{\alpha}{2} \frac{\sqrt{\tau}}{\sqrt{1+\tau}} \cdot \sum_{i \in U(B)} \sum_{j=1}^{b_i - 1} \frac{(j-1)!(b_i - j - 1)!}{(b_i - 1)!} \Big( \sum_{A \in \eta_{ext}^{i,j}(B^s)} \sqrt{\frac{b_i}{j(b_i - j)}} \Big) \\
&\quad \gtrsim \frac{\alpha\sqrt{\tau}}{\sqrt{1+\tau}}.
\end{aligned}$$

Here, the second step follows with high probability by our previous argument, and the last step follows by a similar argument as in the case of uniform prior, except that we have only a constant term instead of an $s$ in the case of uniform prior. This is true because it is possible to have $|U(A)| \ll s$, so the result can only be bounded by a constant multiple of $\alpha \cdot \frac{\sqrt{\tau}}{\sqrt{1+\tau}}$ without the $s$ factor in the uniform case.

Finally, note that as $n$ goes to infinity, the result holds with probability one. Therefore, we obtain that

$$\lim_{n\to\infty} R(s|\{x_i\}_{i=1}^n) = \lim_{n\to\infty} \frac{\mathbb{P}(K_n = s+1|\{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)} \gtrsim \frac{\alpha}{s^2} \cdot \frac{\sqrt{\tau}}{\sqrt{1+\tau}}.$$

As a consequence, we reach the conclusion of the theorem. □

*Proof of Lemma 3.5:* We have:

$$\begin{aligned}
&\frac{m(x_S)m(x_T)}{m(x_S \cup x_T)} \\
&= \frac{1}{\sqrt{\sigma^2}} \sqrt{\frac{(n_s+n_t)+\frac{1}{\sigma^2}}{(n_s+\frac{1}{\sigma^2})(n_t+\frac{1}{\sigma^2})}} \\
&\qquad\qquad \times \exp\left(\frac{1}{2}\left(\frac{n_s^2\overline{x_S}^2}{n_s+\frac{1}{\sigma^2}} + \frac{n_t^2\overline{x_T}^2}{n_t+\frac{1}{\sigma^2}} - \frac{(n_s+n_t)^2\overline{x_S \cup x_T}^2}{(n_s+n_t)+\frac{1}{\sigma^2}}\right)\right).
\end{aligned}$$

To simplify the notation, we let $\tau := \frac{1}{\sigma^2}$ be the precision, and rewrite this expression as:

$$\sqrt{\tau} \cdot \sqrt{\frac{n_s+n_t+\tau}{(n_s+\tau)(n_t+\tau)}} \exp\Big(\frac{1}{2} \underbrace{\Big(\frac{n_s^2\overline{x_S}^2}{n_s+\tau} + \frac{n_t^2\overline{x_T}^2}{n_t+\tau} - \frac{(n_s+n_t)^2\,\overline{x_S \cup x_T}^2}{(n_s+n_t)+\tau}\Big)}_{F(\tau;x_S,x_T)}\Big).$$

Note that the term $F(\tau; x_S, x_T)$ is nonnegative at zero, since

$$F(0; x_S, x_T) = \frac{n_s n_t}{n_s+n_t}(\overline{x_S} - \overline{x_T})^2 \geq 0.$$

Solving a quadratic function shows that the equation $F(\tau; x_1, x_2) = 0$ has its positive root at:

$$\begin{cases}
\frac{1}{4}\Big[\sqrt{8n_s n_t \frac{(\overline{x_S}-\overline{x_T})^2}{\overline{x_S}\,\overline{x_T}} + Q^2} + Q\Big], & \text{if } \overline{x_S}\,\overline{x_T} > 0 \\
\frac{1}{4}\Big[\sqrt{8n_s n_t \frac{(\overline{x_S}-\overline{x_T})^2}{\overline{x_S}\,\overline{x_T}} + Q^2} + Q\Big] \text{ or } \emptyset, & \text{if } \overline{x_S}\,\overline{x_T} < 0 \\
\emptyset & \text{if } \overline{x_S} = 0, \overline{x_T} \neq 0 \text{ or } \overline{x_S} \neq 0, \overline{x_T} = 0 \\
\mathbb{R}^+ & \text{if } \overline{x_S} = \overline{x_T} = 0,
\end{cases}$$

where $Q := n_t\dfrac{n_t\overline{x_T}}{n_s\overline{x_S}} + n_s\dfrac{n_s\overline{x_S}}{n_t\overline{x_T}} - 2(n_s+n_t)$.

If there is no positive root or every positive number is a root, then $F(\tau; x_S, x_T) \geq 0$ for any $\tau > 0$. Otherwise, there exists a unique positive root $r_+(x_S, x_T) := \frac{1}{4}\left[\sqrt{8n_sn_t\frac{(\overline{x}_S-\overline{x}_T)^2}{\overline{x}_S\overline{x}_T}+Q^2}+Q\right]$, a random variable depending on $n_s, n_t$, whose probability density function concentrates on increasingly larger values as long as one of $n_s, n_t$ goes to infinity. That is, the root grows larger stochastically as $n_s$ and $n_t$ increase. For any fixed $\tau$, as $n_s + n_t$ goes to infinity, it follows that for any fixed $\tau > 0$, the probability that $[0, r_+(x_S, x_T)]$ contains $\tau$ approaches one asymptotically, so we have that $F(\tau; x_S, x_T) \geq 0$.

Note that for any positive integers $n_s, n_t$ and nonnegative number $\tau$, we have

$$\frac{n_s+n_t+\tau}{(n_s+\tau)(n_t+\tau)} > \frac{1}{2}\cdot\frac{1}{1+\tau}\cdot\frac{n_s+n_t}{n_sn_t}.$$

The result follows. □

## 4. Extension to Pitman-Yor process mixture models.

In this section, we extend our results to the Pitman-Yor process mixture model, a model that is similar to the Dirichlet process mixture model but allows a faster rate of creation of clusters. Accordingly, we will see that the ratio between the posterior probabilities of consecutive number of clusters is larger than that of the DPMM.

We retain most of the notation used in the previous section unless otherwise specified. We also use the standard Chinese restaurant process nomenclature to describe the dynamics associated with the partitions formed by the Pitman-Yor process, where the items to be partitioned are referred to as "customers," and partions are referred to as "tables" [23]. Letting $(\alpha, \xi)$ denote the parameters of the Pitman-Yor process prior, we obtain:

$$\mathbb{P}(\text{customer } n+1 \text{ joins table } c \,|\, A^{n,s}) = \begin{cases} \frac{|c|-\xi}{\alpha+n} & \text{if } c \leq s, \\ \frac{\alpha+\xi s}{\alpha+n} & \text{otherwise.} \end{cases} \tag{11}$$

The idea is that joining an existing table $c$ has a probability of $\frac{|c|-\xi}{\alpha+n}$, a probability proportional to $|c| - \xi$, the number of customers already at the table, discounted by the value $\xi$, and the probability of joining new table is proportional to $\alpha + \xi K_n$. Then, the probability of obtaining $A^{n,s}$ is:

$$\mathbb{P}(A^{n,s}) = \frac{\alpha(\alpha+\xi)\cdots(\alpha+\xi(s-1))}{\alpha^{(n)}}\prod_{i=1}^{s}(1-\xi)(2-\xi)\cdots(|A_i^{n,s}|-1-\xi),$$

where $\alpha^{(n)} = \alpha(\alpha+1)\cdots(\alpha+n-1)$. Note that if $|A_i^{n,s}| = 1$, the corresponding term in the product is 1. We denote

$$|c \ominus \xi|! = (1-\xi)(2-\xi)\cdots(c-\xi),$$

and define $|0 \ominus \xi|! = 1$.

Overall, we write:

$$(A^{n,s}, s) \sim \mathrm{PYP}(\alpha, \xi),$$

and define the *Pitman-Yor Process Mixture Model* (PYPMM) as follows:

$$(A^{n,s}, s) \sim \mathrm{PYP}(\alpha, \xi) \tag{12}$$

$$p(\theta \,|\, A^{n,s}, s) = \prod_{i=1}^{s} \pi(\theta_i) \tag{13}$$

$$p(\{x_i\}_{i=1}^n \,|\, \{\theta_j\}_{j=1}^s, A^{n,s}, s) = \prod_{j=1}^{s} \prod_{x \in A_j^{n,s}} f_{\theta_j}(x_i), \tag{14}$$

where $\pi$ is a prior on $\theta$ and $f_{\theta_j}$ belongs to a family of (known) densities. Moreover, analogously with the case of the DPMM, we obtain the following form for the posterior:

$$\begin{aligned}
&\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n) \\
&= \frac{\mathbb{P}(\{x_i\}_{i=1}^n \,|\, K_n = s)\mathbb{P}(K_n = s)}{\mathbb{P}(\{x_i\}_{i=1}^n)} \\
&\propto \sum_{A^{n,s} \in \rho_s(n)} \mathbb{P}(A^{n,s}) \cdot \int_{\{\theta_j\}_{j=1}^n} p(\{x_i\}_{i=1}^n \,|\{\theta_j\}_{j=1}^s) p(\{\theta_j\} \,|A^{n,s})\, d\{\theta_j\}_{j=1}^s \\
&= \sum_{A^{n,s} \in \rho_s(n)} \mathbb{P}(A^{n,s}) \int_{\{\theta_j\}_{j=1}^s} \left( \prod_{j=1}^{s} \prod_{x_i \in A_j^{n,s}} f_{\theta_j}(x_i) \right) \prod_{j=1}^{s} \pi(\theta_j)\, d\{\theta_j\}_{j=1}^s \\
&= \sum_{A^{n,s} \in \rho_s(n)} \mathbb{P}(A^{n,s}) \prod_{j=1}^{s} \int_{\theta_j} \left( \prod_{x_i \in A_j^{n,s}} f_{\theta_j}(x_i) \right) \pi(\theta_j)\, d\theta_j \\
&= \sum_{A^{n,s} \in \rho_s(n)} \mathbb{P}(A^{n,s}) \prod_{j=1}^{s} m(x_{A_j^{n,s}}).
\end{aligned}$$

Thus, we have

$$\begin{aligned}&\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)\mathbb{P}(\{x_i\}_{i=1}^n)\\ &= \sum_{A^{n,s}\in\rho_s(n)} \frac{\alpha(\alpha+\xi)\cdots(\alpha+\xi(s-1))}{\alpha^{(n)}}\prod_{i=1}^{s} |(|A_i^{n,s}|-1)\ominus\xi|!\cdot m(x_{A_i^{n,s}}).\end{aligned}$$

Recall the definitions

$$\begin{aligned}&\eta_{ext}(A^{n,s}) := \{\tilde{A}^{n,s+1}\in\rho_{s+1}(n) : A^{n,s}\in\eta_{con}(\tilde{A}^{n,s+1})\},\\ &\eta_{con}(A^{n,s}) := \Big\{\tilde{A}^{n,s-1}\in\rho_{s-1}(n) : !\exists\, i,j\in[s] :\\ &\qquad\{A_1^{n,s},\cdots,A_s^{n,s},A_i^{n,s}\cup A_j^{n,s}\}\backslash\{A_i^{n,s},A_j^{n,s}\} = \{\tilde{A}_1^{n,s-1},\cdots,\tilde{A}_{s-1}^{n,s-1}\}\Big\},\end{aligned}$$

and note that $\eta_{s+1}(A)$ is the set of partitions formed by combining $A_i$ and $A_{s+1}$, and $\tilde{\eta}_s(B)$ is all sets in $\rho_{s+1}(n)$ that can be formed by splitting one of $B$ into two clusters (and then making one of them the “last cluster”). This yields:

$$\begin{aligned}R(s|\{x_i\}_{i=1}^n) &= \frac{\mathbb{P}(K_n = s+1\,|\,\{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s\,|\,\{x_i\}_{i=1}^n)}\\ &= \frac{2}{(s+1)s}\cdot\frac{\sum_{B^{n,s}\in\rho_s(n)}\left(\sum_{A^{n,s-1}\in\eta_{ext}(B^{n,s})}p(A^{n,s-1},x)\right)}{\sum_{B^{n,s}\in\rho_s(n)}p(B^{n,s},x)}.\end{aligned}\tag{15}$$

Thus, continuing with the proof, we have:

$$\begin{aligned}&\frac{\sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})}p(A^{n,s+1}|x)}{p(B^{n,s}|x)}\\ &= \sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})}(\alpha+\xi s)\frac{\prod_{i=1}^{s+1}|(|A_i^{n,s+1}|-1)\ominus\xi|!\cdot\prod_{i=1}^{s+1}m(x_{A_i^{n,s+1}})}{\prod_{i=1}^{s}|(|B_i^{n,s}|-1)\ominus\xi|!\cdot\prod_{i=1}^{s}m(x_{B_i^{n,s}})}\\ &= (\alpha+\xi s)\sum_{i=1}^{s}\sum_{j=1}^{b_i-1}\sum_{A^{n,s+1}\in\eta_{ext}^{i,j}(B^{n,s})}\frac{|(j-1)\ominus\xi|!\cdot|(b_i-j-1)\ominus\xi|!}{|(b_i-1)\ominus\xi|!}\\ &\qquad\qquad\times\frac{m(x_{A_i^{n,s+1}})m(x_{A_{s+1}^{n,s+1}})}{m(x_{B_i^{n,s}})}.\end{aligned}\tag{16}$$

Note that we let $b_i = |B_i|$, and let $\eta_{ext}^{i,j}(B^{n,s})$ be defined to be all $A^{n,s+1}\in\eta_{ext}(B^{n,s})$ obtained by splitting $B_i^{n,s}$ into two clusters of size $j$ and $b_i-j-1$. We now separate into two cases, where the prior is uniform and Gaussian respectively.

4.1. *Uniform prior.* Similar to our analysis with the DPMM, we first consider the behavior of the posterior distribution for the number of clusters when the prior is a uniform distribution.

THEOREM 4.1. *Given the PYPMM defined in Eq.* (14) *with a uniform prior* $\mathrm{Unif}(\Theta)$ *on* $\theta$*, when* $n$ *is sufficiently large, if* $\min(\{x_i\}_{i=1}^n) > \min(\Theta)+c$ *and* $\max(\{x_i\}_{i=1}^n) < \max(\Theta)-c$ *for some* $c>0$*, then the ratio* $R(s|\{x_i\}_{i=1}^n)$ *between consecutive terms is lower bounded by*

$$R(s|\{x_i\}_{i=1}^n) \succsim \frac{(\alpha+\xi s)}{s|\Theta|}. \tag{17}$$

PROOF. By Lemma (3.2) and under the same condition, we have that when $B_i^{n,s} = A_i^{n,s+1} \cup A_{s+1}^{n,s+1}$ and letting $a_k = |A_k^{n,s+1}|$ for $k \in [s+1]$,

$$\frac{m(x_{A_i^{n,s+1}})m(x_{A_{s+1}^{n,s+1}})}{m(x_{B_i^{n,s}})} \geq \frac{c_1\sqrt{2\pi}}{|\Theta|}\frac{\sqrt{a_i+a_{s+1}}}{\sqrt{a_i a_{s+1}}}.$$

Now, we obtain that

$$\begin{aligned}
&\frac{\sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} p(A^{n,s+1}|x)}{p(B^{n,s}|x)} \\
&\geq (\alpha+\xi s)\sum_{i=1}^{s}\sum_{j=1}^{b_i-1}\frac{|(j-1)\ominus\xi|!|(b_i-j-1)\ominus\xi|!}{|(b_i-1)\ominus\xi|!} \\
&\qquad\qquad\times \sum_{A^{s+1}\in\eta_{ext}^{i,j}(B^s)}\frac{c_1\sqrt{2\pi}}{|\Theta|}\cdot\frac{\sqrt{b_i}}{\sqrt{j(b_i-j)}} \\
&= (\alpha+\xi s)\frac{c_1\sqrt{2\pi}}{|\Theta|}\sum_{i=1}^{s}\sum_{j=1}^{b_i-1}\frac{|(j-1)\ominus\xi|!|(b_i-j-1)\ominus\xi|!}{|(b_i-1)\ominus\xi|!}\cdot\binom{b_i}{j}\sqrt{\frac{b_i}{j(b_i-j)}}.
\end{aligned}$$

Note for $j=1$ and $j=b_i-1$ we have that for $b_i\geq 2$

$$\frac{|(j-1)\ominus\xi|!|(b_i-j-1)\ominus\xi|!}{|(b_i-1)\ominus\xi|!}\cdot\binom{b_i}{j}\sqrt{\frac{b_i}{j(b_i-j)}} = \frac{b_i}{b_i-1-\xi}\sqrt{\frac{b_i}{b_i-1}} \geq 1.$$

Plugging this back into the expression above

$$\frac{\sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} p(A^{n,s+1}|x)}{p(B^{n,s}|x)} \geq (\alpha+\xi s)\frac{(0.997)^2\sqrt{2\pi}}{|\Theta|}\sum_{i=1}^{s}2\cdot I_{b_i\geq 2} \gtrsim \frac{(\alpha+\xi s)s}{|\Theta|}.$$

Then, using Eq. 16 we get that

$$\begin{aligned} R(s|\{x_i\}_{i=1}^n) &= \frac{2}{(s+1)s} \cdot \frac{\sum_{B^{n,s}\in\rho_s(n)} \left(\sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} p(A^{n,s+1}|x)\right)}{\sum_{B^{n,s}\in\rho_s(n)} p(B^{n,s}|x)} \\ &\geq \frac{2}{(s+1)s} \min_{B^{n,s}\in\rho_s(n)} \sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} \frac{p(A^{n,s+1}|x)}{p(B^{n,s}|x)} \\ &\gtrsim \frac{2}{(s+1)s} \frac{(\alpha+\xi s)s}{|\Theta|} \gtrsim \frac{(\alpha+\xi s)}{s|\Theta|} = \frac{\alpha}{s|\Theta|} + \frac{\xi}{|\Theta|}. \end{aligned}$$

As a consequence, we obtain the conclusion of the theorem. □

4.2. *Gaussian Prior.* We now turn to the case where the prior is a Gaussian distribution. We make similar assumptions on the prior as we made in Section 3.2 with the DPMM.

THEOREM 4.2. *For the PYPMM defined in Eq. (14), with a Gaussian prior $\mathcal{N}(0,\sigma^2)$ on $\theta$, as $n$ goes to infinity, the ratio $R(s|\{x_i\}_{i=1}^n)$ satisfies the following asymptotic lower bound:*

$$\lim_{n\to\infty} R(s|\{x_i\}_{i=1}^n) \geq \frac{C(\alpha+\xi s)}{s^2} \cdot \frac{1}{1+\sqrt{\sigma^2}}, \tag{18}$$

*where $C > 0$ is a universal constant.*

PROOF. Most steps are the same as in the derivation for the DPMM. We denote

$$D_{j,b_i} = \frac{|(j-1)\ominus\xi|!|(b_i-j-1)\ominus\xi|!}{|(b_i-1)\ominus\xi|!}.$$

Then, by equation (16) we have that

$$\begin{aligned} &\sum_{A^{n,s+1}\in\eta_{ext}(B^{n,s})} \frac{p(A^{n,s+1}|x)}{p(B^{n,s}|x)} \\ &= (\alpha+\xi s) \sum_{i=1}^{s} \sum_{j=1}^{b_i-1} D_{j,b_i} \sum_{A^{n,s+1}\in\eta_{ext}^{i,j}(B^{n,s})} \frac{m(x_{A_i^{n,s+1}})m(x_{A_{s+1}^{n,s+1}})}{m(x_{B_i^{n,s}})} \\ &\overset{w.h.p.}{\geq} C_0(\alpha+\xi s)\sqrt{\tau} \sum_{i\in U(B^{n,s})} \sum_{j=1}^{b_i-1} D_{j,b_i} \times \left( \sum_{A^{n,s+1}\in\eta_{ext}^{i,j}(B^{n,s})} \sqrt{\frac{b_i+\tau}{(j+\tau)(b_i-j+\tau)}} \right) \end{aligned}$$

$$\geq C_0(\alpha+\xi s)\sqrt{\tau}\sum_{i\in U(B^{n,s})}\sum_{j=1}^{b_i-1} D_{j,b_i}\times\left(\sum_{A\in\eta_{ext}^{i,s}(B^{n,s})}\frac{1}{\sqrt{2(1+\tau)}}\sqrt{\frac{b_i}{j(b_i-j)}}\right)$$

$$\geq \frac{C\sqrt{\tau}}{1+\sqrt{\tau}}(\alpha+\xi s),$$

where most of the steps follow similarly as in the proof of the DPMM in Theorem 3.4. Thus we conclude that

$$\lim_{n\to\infty} R(s\,|\,\{x_i\}_{i=1}^n) \gtrsim \frac{(\alpha+\xi s)}{s^2}\cdot\frac{\sqrt{\tau}}{1+\sqrt{\tau}} = \frac{\sqrt{\tau}}{1+\sqrt{\tau}}\left(\frac{\alpha}{s^2}+\frac{\xi}{s}\right).$$

As a consequence, we reach the conclusion of the theorem. □

Remark. Compared to the DPMM, the PYPMM has a faster rate of table generation, and accordingly a longer tail for the number of clusters. As we can see in Theorems 4.1 and 4.2, this leads to a larger lower bound in the ratio between consecutive posterior probabilities. Essentially, the lower bound for the uniform case is raised to a constant, and that for the Gaussian case is raised to the order of $1/s$, which corresponds to our intuition that PYPMM is more likely to fit a bigger number of clusters in the posterior.

## 5. Experiments.

In the current section, we discuss the results of numerical experiments that aim to empirically strengthen our understanding of the clustering behavior of the DPMM and PYPMM. Throughout this section, for simplicity of implementation, we consider the Dirichlet mixture of standard normals; i.e., we assume $f_\theta(x)=\frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2/2\sigma^2}$, where $f_\theta(x)$ is the likelihood function in Eq. (3). As for the prior density on the parameters, we choose $\pi(\theta)$ to be Gaussian.

To compute the posterior density given the data points, we note that if data points $\{x_i\}_{i=1}^k$ fall within the cluster, then it follows that

$$\begin{aligned} p(\theta|\{x_i\}_{i=1}^n) &\propto p(\theta)p(\{x_i\}_{i=1}^n|\theta) \\ &= \frac{1}{\sqrt{2\pi}}e^{-\theta^2/2\sigma^2}\prod_{i=1}^n\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_i-\theta)^2} \\ &\sim \mathcal{N}\left(\frac{\sigma^2\sum_{i=1}^n x_i}{n\sigma^2+1},\frac{\sigma^2}{n\sigma^2+1}\right). \end{aligned}$$

Moreover, to compute the marginal density of a specific $x$, we compute as follows:

$$\begin{aligned} f(x) &= \int f_\theta(x)\pi(\theta)\, d\theta \\ &= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\theta^2}{2\sigma^2}}\, d\theta \\ &\sim \mathcal{N}(0, \sigma^2+1). \end{aligned}$$

We pick $\alpha = 1$ and $\sigma = 1$ for our simulations. Specifically, we use the "dirichletprocess" R package [12] for these simulations, where the software package runs a Gibbs sampler on the DPMM model. For all the experiments, we run the Gibbs sampler with a burn-in period with $2 \times 10^5$ burn-in steps, and then take 10000 succeeding samples with step size 100.

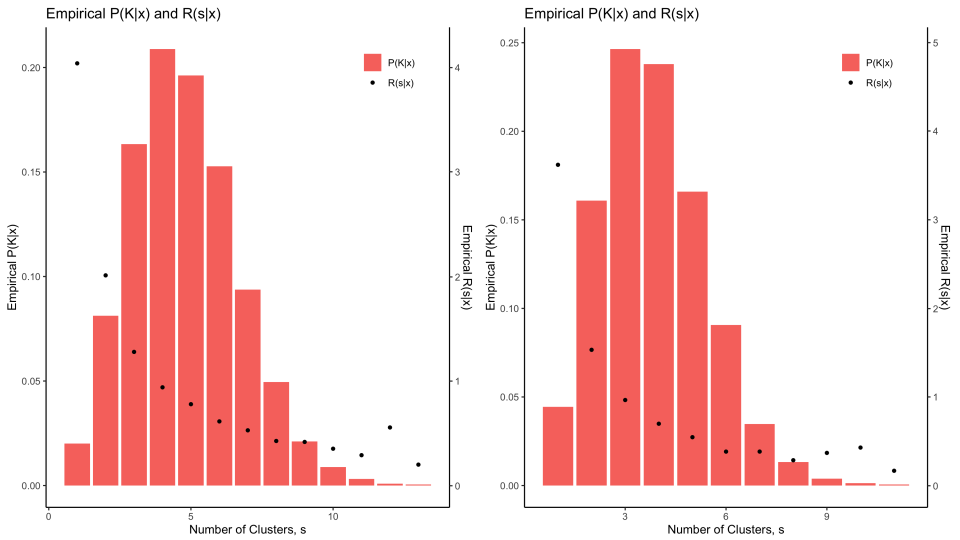
5.1. *Finite-cluster models.* In this subsection we consider two finite-cluster distributions. In the first experiment, we generate 300 i.i.d. data points following a standard normal distribution. In the second experiment, we generate another 300 data points with the following two-cluster distribution:

$$f(x) = \frac{1}{2}, \quad \forall x \in [0,1] \cup [2,3].$$

After running the Gibbs sampler, we compute the posterior frequency of the number of clusters $F_s$ for each $s$ with some nonzero count of posterior frequency. This gives an empirical distribution of $\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)$, from which we may approximate $R(s|\{x_i\}_{i=1}^n)$. Note, however, that this approximation may be of low quality for large $s$ given the corresponding low posterior frequency.

Figure 1 presents a plot of posterior frequencies and their ratios. The red bars correspond to the posterior frequency of the number of components, and the black dots are the ratios between the posterior frequencies, $F_{s+1}/F_s$.

The theoretical bound asserts that the ratio between consecutive posterior frequencies is at least of the order of $1/s^2$ under a Gaussian prior. To evaluate this, we plot $R(s) \cdot s^2$ versus $s$, which, roughly speaking, should look flat or have an upwards trend. In Figure 2a, we plot $\widehat{R}(s|x) \cdot s^2$ for both the one-cluster Gaussian and two-cluster uniform cases, where we observe the expected upward trend. In Figure 2b, we plot $\widehat{R}(s|x) \cdot s$ to assess whether the relationship is roughly linear. If linearity holds true then we would expect to see the dots for both settings to be roughly flat. Finally we note that it is reasonable to have erratic behavior of $\widehat{R}(s|x)$ when $s$ is large since the posterior frequencies for large $s$ are all very close to zero and the estimator $\widehat{R}$ is unstable in that case.



(a) Gaussian, one cluster

(b) Uniform, two clusters

Fig 1: Plots for empirical $\mathbb{P}(K_n|x)$ and $R(s|x)$

5.2. *Dirichlet process mixture models.* It is also of interest to investigate the behavior when the data follow an infinite-component distribution.

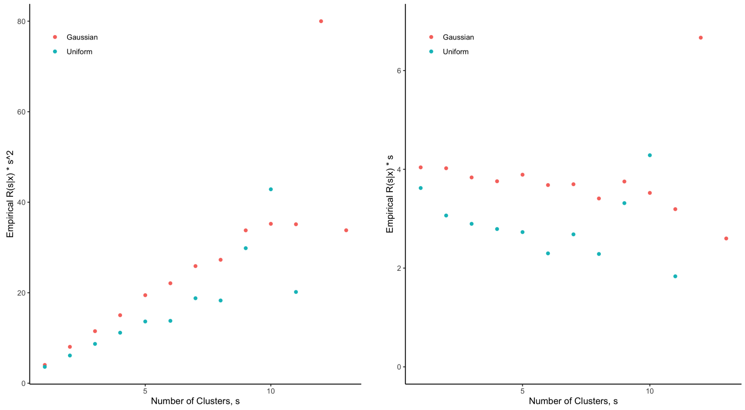
To study this case, we investigate a Dirichlet process data-generating model, with the $\alpha = 3$, $\theta_i \overset{i.i.d.}{\sim} N(0, 25)$ for each $i$, and

$$x_i|\theta_j, i \in A_j^{n,s} \sim N(\theta_j, 1).$$

Again, we generate 300 samples for three independent experiments, which result in 11, 14, and 17 clusters respectively.

We again plot the empirical values of $\mathbb{P}(K_n|x)$ and $R(s|x)$ in Figure 3, and investigate the nature of the growth rate of the posterior number of clusters in Figure 4. As we see in the latter figure, the ratios between consecutive posterior frequencies of the number of clusters are at least of order $1/s^2$.

**6. Discussion.** We have established lower bounds on the ratio of posterior probabilities $R(s|\{x_i\}_{i=1}^n)$ for the number of clusters for both the Dirichlet process mixture model and Pitman-Yor process mixture model, under several choices of prior distributions on the parameter space. While the complicated combinatorial structure of the DPMM and PYPMM preclude general characterizations of the posterior distribution, we have shown that it is possible to obtain analytical results for the posterior distribution



(a) Empirical $\widehat{R}(s|x) * s^2$

(b) Empirical $\widehat{R}(s|x) * s$

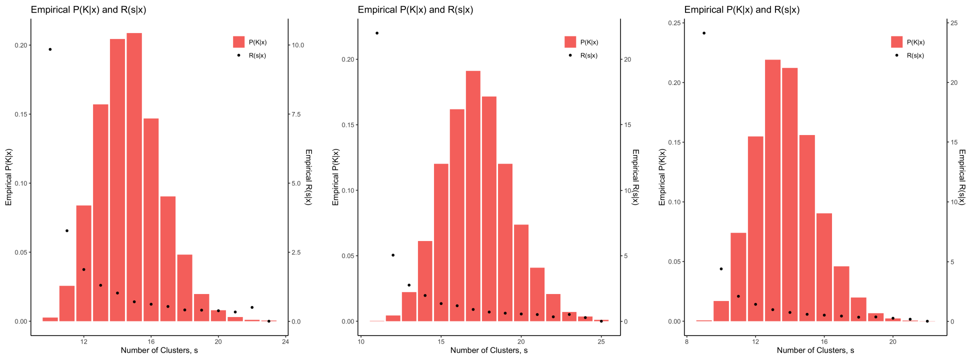
Fig 2: Plots of $\widehat{R}(s|x)$ with respect to different order of powers

of the number of clusters. We obtained both asymptotic and nonasymptotic results for the distribution of this important quantity.

An interesting open problem is to consider whether our rates are optimal or can be further improved. In particular, in the case of a Gaussian prior, our simulation suggests that instead of the $1/s^2$ rate predicted by our theory, a tighter $1/s$ rate may be possible. This seems plausible given that the $1/s$ rate arises in the case of uniform prior.

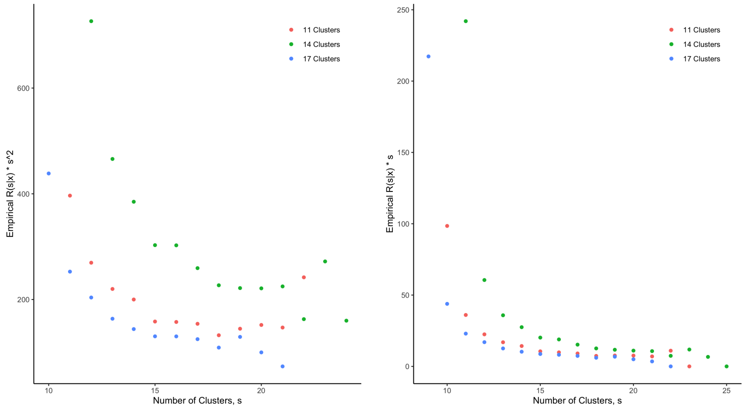
Our results provide a strong negative response to the question of whether basic nonparametric Bayesian models such as the DPMM or the PYPMM are able to infer the true number of clusters when the true number of clusters is finite. Indeed, our results provide a quantitative refutation of this naive hope. Additional mechanisms, such as truncation or some form of regularization, will be necessary to obtain consistent inference of the number of clusters in the finite setting. Indeed, recent work has shown that truncation can yield consistency with the number of clusters when the true data-generation mechanism is a finite mixtures [13]. However, the truncation method brings additional tuning parameters into the picture, including notions of separation of clusters, which may not be easily determined in practice. The problem of consistent estimation of the number of clusters remains open.

There is, however, another interesting open problem, which arises when the true distribution contains an infinite number of components. Do the DPMM or the PYPMM guarantee an infinite number of clusters in this



(a) 11 clusters (b) 14 clusters (c) 17 clusters

Fig 3: Plots for empirical $\mathbb{P}(K_n|x)$ and $R(s|x)$ in three simulations from Dirichlet process.



(a) Empirical $R(s|x) * s^2$ (b) Empirical $R(s|x) * s$

Fig 4: Plots of $\widehat{R}(s|x)$ with respect to different order of powers in three simulations from the Dirichlet process.

case? Does the rate of growth of clusters in the posterior match that of the data-generating distribution, at least asymptotically? What about the case in which the truth is Dirichlet process mixture of normals or simply a general Dirichlet process? Does DPMM generate a posterior number of clusters at the same rate as implied by the Dirichlet process? In answering these questions, it will be necessary to derive some sort of upper bound on the $R(s|X)$, and focus on the asymptotic regime.

**7. Acknowledgements.** We wish to acknowledge support from the by Army Research Office grant W911NF-17-1-0304 and the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

## References.

[1] Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2** 1152-1174. (Cited on pages 1 and 3.)

[2] Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Polya urn schemes. *Annals of Statistics* **1** 353–355. (Cited on page 1.)

[3] Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90** 577–588. (Cited on pages 1 and 10.)

[4] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1** 209–230. (Cited on page 1.)

[5] Fox, E., Sudderth, E., Jordan, M. I. and Willsky, A. (2011). A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics* **5** 1020-1056. (Cited on page 1.)

[6] Ghosal, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics*. Cambridge University Press Walker, Cambridge. (Cited on page 1.)

[7] Ghosal, S., Ghosh, J. K. and van der Vaart, A. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28** 500–531. (Cited on page 1.)

[8] Ghosal, S. and van der Vaart, A. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statististics* **29** 1233–1263. (Cited on page 1.)

[9] Ghosal, S. and van der Vaart, A. (2007a). Convergence rates of posterior distributions for noniid observations. *Annals of Statistics* **35** 192–223. (Cited on page 1.)

[10] Ghosal, S. and van der Vaart, A. (2007b). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Annals of Statistics* **35** 697–723. (Cited on page 1.)

[11] Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics. Springer Series in Statistics*. Springer. (Cited on page 1.)

[12] Gordon, J. R. and Markwick, D. (2019). dirichletprocess: Build Dirichlet process objects for Bayesian modelling R package version 0.3.0. (Cited on page 19.)

[13] Guha, A., Ho, N. and Nguyen, L. (2019). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Arxiv preprint Arxiv: 1901.05078.* (Cited on page 21.)

[14] Huelsenbeck, J. P. and Andolfatto, P. (2007). Inference of population structure under a Dirichlet process model. *Genetics* **175** 1787–1802. (Cited on page 1.)

[15] Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21** 1095–1109. (Cited on page 1.)

[16] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *Annals of Statistics* **12** 351–357. (Cited on page 3.)

[17] MacEachern, S. and Mueller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7** 223-238. (Cited on page 10.)

[18] Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18** 1194–1206. (Cited on page 1.)

[19] Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems* 199–206. (Cited on pages 2 and 4.)

[20] Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research* **15** 3333–3370. (Cited on pages 2 and 4.)

[21] Otranto, E. and Gallo, G. M. (2002). A nonparametric Bayesian approach to detect the number of regimes in Markov switching models. *Econometric Reviews* **21** 477–496. (Cited on page 1.)

[22] Paisley, J., Wang, C., Blei, D. and Jordan, M. I. (2015). Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37** 256-270. (Cited on page 1.)

[23] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability* 855–900. (Cited on pages 1 and 13.)

[24] Rodriguez, A., Dunson, D. and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association* **103** 1131–1154. (Cited on page 1.)

[25] Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of Betas and adaptive nonparametric estimation of the density. *Annals of Statistics* **38** 146-180. (Cited on page 5.)

[26] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics* **29** 687–714. (Cited on page 1.)

[27] Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101** 1566–1581. (Cited on page 1.)

[28] Walker, S., Lijoi, A. and Prunster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Annals of Statistics* **35** 738–746. (Cited on page 1.)

Chiao-Yu Yang
Department of Statistics
E-mail: chiaoyu@berkeley.edu

Eric Xia
Department of Statistics
E-mail: ericzxia@berkeley.edu

Nhat Ho
Department of Statistics and Data Sciences
E-mail: minhnhat@utexas.edu

Michael I. Jordan
Department of Statistics
Department of Electrical Engineering
and Computer Sciences
E-mail: jordan@cs.berkeley.edu