# Posterior Distribution for the Number of Clusters in Dirichlet Process Mixture Models

**Chiao-Yu Yang**
UC Berkeley
chiaoyu@berkeley.edu

**Nhat Ho**
UC Berkeley
minhnhat@berkeley.edu

**Michael I. Jordan**
UC Berkeley
jordan@cs.berkeley.edu

## Abstract

Dirichlet process mixture models (DPMM) play a central role in Bayesian nonparametrics, with applications throughout statistics and machine learning. DPMMs are generally used in clustering problems where the number of clusters is not known in advance, and the posterior distribution is treated as providing inference for this number. Recently, however, it has been shown that the DPMM is inconsistent in inferring the true number of components in certain cases. This is an asymptotic result, and it would be desirable to understand whether it holds with finite samples, and to more fully understand the full posterior. In this work, we provide a rigorous study for the posterior distribution of the number of clusters in DPMM under different prior distributions on the parameters and constraints on the distributions of the data. We provide novel lower bounds on the ratios of probabilities between $s+1$ clusters and $s$ clusters when the prior distributions on parameters are chosen to be Gaussian or uniform distributions.

## 1 Introduction

Since the introduction of Dirichlet process in Ferguson's seminal paper [4, 2], models based on the Dirichlet process and related combinatorial stochastic processes, such as the Pitman-Yor process (see for example [22]), have been used for various statistical and machine learning problems. These models are generally mixture models, with the Dirichlet process or Pitman-Yor process used as a prior on mixture components [1]. Statistical applications have included a wide variety of problems in density estimation [3, 6, 7, 25, 11, 8, 9, 27, 10] and parameter estimation [26, 23, 5, 21]. Applied problems that have been studied include problems such as computer vision, and text processing, as well as in scientific fields such as economics, astronomy, molecular biology, and genetics [26, 23, 20, 3, 14, 13, 17]. The use of Dirichlet process mixture models (DPMMs) in these fields is generally motivated by the assertion that it can be used to determine the number of components in nonparametric mixture models.

Classical clustering algorithms such as $K$-means or Gaussian mixture models generally require us to set the number of clusters $k$ a priori. However, in practice the real number of clusters is rarely known, and almost never known for dynamically growing data sets. This has motivated the use of DPMMs to find the number of clusters. Unlike $k$-means or GMMs, the DPMM is based on the Dirichlet process which has infinite components and does not require one to specify the number of components at first. The goal is to use the posterior distributions of the number of clusters to find an optimal choice for clustering.

However, unlike the case of density estimation, a theoretical understanding of the convergence of the number of components in DPMM is still largely missing in the literature. On the negative side, it has been demonstrated that DPMM and PYPMM may exhibit posterior inconsistencies in the number of components when the true number of components is finite [18, 19]. Moreover, in practice, it has been observed that DPMM-based inference can generate small clusters that do not reflect the underlying

Preprint. Under review.

data-generating process, especially when the real number of components is small instead of infinite. Despite these observations, quantitatively we have little understanding about the behavior of the posterior distribution of the number of components when the number of samples goes to infinity, not to mention in the nonasymptotic regime.

To fill the gap, in this work, we study the posterior distribution of the number of clusters for DPMM-based clustering models. Our main results are lower bounds on the ratio of the probabilities of obtaining $s+1$ clusters and $s$ clusters under Gaussian or uniform priors for the parameters, with different assumptions and constraints on the data. This yields a fine-grained understanding of the posterior distribution induced by the DPMM on the number of clusters, and positions us for future work on topics such as the rate of growth for the number of clusters in the posterior when the number of clusters is not fixed but also growing with the sample size.

## 2 Model Description

We first introduce some key notation. We use $\{x_i\}_{i=1}^n$ to denote the $n$ samples $\{x_1, \cdots, x_n\}$, $[n]$ to denote the set $\{1, \cdots, n\}$, $A \in \rho_s(n)$ to denote the set $\{A_1, \cdots, A_s\}$ such that $A_i$'s form an s-partition of $[n]$, where $\rho_s(n)$ is the set of all $s$-partitions on $[n]$.

The DPMM [1, 15] is specified as follows:

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

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

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

where $\pi$ stands for a given prior on the parameter $\theta$ while $\{f_\theta(\cdot)\}$ is a known family of density functions.

The DPMM has been widely used in machine learning and statistics for problems including density estimation and parameter estimation. In this paper, we specifically focus on the application of DPMM to clustering problems. For this application, the prior for the number of clusters with $n$ samples is given by

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

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

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

The central goal of this work is a rigorous study with the behavior of $\mathbb{P}(K_n = s|\{x_i\}_{i=1}^n)$ under two representative choices of prior $\pi$ and different assumptions on the data generating processes. In particular, to ease the ensuing discussion, we use $m(x_{A_j})$ to denote the cluster probability:

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

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

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

To understand the behavior of the posterior distribution of the number of clusters in (4), we consider the ratio between its values at $K_n = s+1$ and $K_n = s$, which can be computed directly as follows:

$$\begin{aligned} R(s|\{x_i\}_{i=1}^n) &:= \frac{\mathbb{P}(K_n = s+1 \,|\, \{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s \,|\, \{x_i\}_{i=1}^n)} \\ &= \frac{\sum_{A\in\rho_{s+1}(n)} \Big(p(A,s+1)\cdot\prod_{j=1}^{s+1} m(x_{A_j})\Big)}{\sum_{B\in\rho_s(n)} \Big(p(B,s)\cdot\prod_{j=1}^{s} m(x_{B_j})\Big)} \\ &= \alpha\cdot\frac{\sum_{A\in\rho_{s+1}(n)} \Big((\prod_{i=1}^{s+1}(|A_i|-1)!)\cdot\prod_{j=1}^{s+1} m(x_{A_j})\Big)}{\sum_{B\in\rho_s(n)} \Big((\prod_{i=1}^{s}(|B_i|-1)!)\cdot\prod_{j=1}^{s} m(x_{B_j})\Big)}. \end{aligned} \tag{5}$$

Throughout the paper, we consider the Dirichlet mixture of standard normals, i.e., $f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$ for all $x$.

## 3 Uniform Prior

We first consider a study with posterior distribution of the number of clusters of DPMM where the data lie on a bounded set [24]. Practically speaking, many data sets that naturally arise in fields such as biology, genetics, and economics are essentially bounded. Under this setting of data, the parameter space $\Theta$ is usually chosen as a compact set.

In this section, to ease the complexity of proof argument, 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 $\theta \in \Theta$. Now, we start with the following result regarding the lower bound of $R(s|\{x_i\}_{i=1}^n)$ under certain assumptions with the data:

**Theorem 1.** *Given DPMM defined in* (3) *with a uniform prior* $\mathrm{Unif}(\Theta)$ *on* $\theta$*. Then, 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*

$$R(s|\{x_i\}_{i=1}^n) \gtrsim \frac{\alpha}{s|\Theta|}. \tag{6}$$

**Remark.** *The condition of Theorem 1 regarding the data can be relaxed to requiring only most of the data to be within* $\Theta$*; however, that weaker condition will require a slightly more complicated proof. Additionally, that condition is mild in many problems since for a clustering problem, when one applies a uniform prior for the parameters (the means of normal distributions), one expects the uniform prior to be big enough to capture the means of all the components.*

*Proof.* Now, for $A \in \rho_{s+1}(n)$ and $B \in \rho_s(n)$, we define two key terms $\eta_{s+1}(A)$ and $\tilde{\eta}_s(B)$ as follows:

$$\begin{aligned}\eta_{s+1}(A) &:= \{\tilde{A} \in \rho_s(n) : !\exists\, i \in [s] : \forall j \neq i, A_j = \tilde{A}_j, A_i \cup A_{s+1} = \tilde{A}_i\}, \\ \tilde{\eta}_s(B) &:= \{\tilde{B} \in \rho_{s+1}(n) : B \in \eta_{s+1}(\tilde{B})\}.\end{aligned}$$

To avoid notation cluttering, in the following we fix $s$ in our discussion and will write $\eta := \eta_{s+1}$ and $\tilde{\eta} := \tilde{\eta}_s$, unless otherwise specified. To interpret our results above, we note that $\eta(A)$ is the set of partitions in $\rho_s(n)$ that can be obtained from $A$ by combining two elements of $A$ into one element

and keeping everyone else the same. Conversely, $\tilde{\eta}_s(B)$ is the set of partitions in $\rho_{s+1}(n)$ which can combine two of its elements into one to get $B$. Toward showing the result, we further define the posterior probabilities of a partition for $A \in \rho_{s+1}(n), B \in \rho_s(n)$ to be:

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

Based on the above definitions, we can rewrite the ratio $R(s)$ as follows:

$$R(s) = \frac{\mathbb{P}(K_n = s+1 | \{x_i\}_{i=1}^n)}{\mathbb{P}(K_n = s | \{x_i\}_{i=1}^n)} = \frac{2}{(s+1)s} \cdot \frac{\sum_{B \in \rho_s(n)} \Big( \sum_{A \in \tilde{\eta}_s(B)} p(A,x) \Big)}{\sum_{B \in \rho_s(n)} p(B,x)} \tag{7}$$

for $n \geq s+1$.

Assume that the above claim is given at the moment (the proof of this claim is deferred to the end of the proof of Theorem 1). We proceed to finish the proof of the theorem. Let $B \in \rho_s(n)$ and denote $\tilde{\eta}^i(B)$ the set of $A \in \rho_{s+1}(n)$ such that $A_i \cup A_{s+1} = B_i$ and $B_j = A_j$ for $j \neq i$. Additionally, let $\tilde{\eta}^{i,j}(B)$ be the subset of $\tilde{\eta}^i(B)$ such that the corresponding $A$ has $|A_i| = j$. Here, we note that the order in the partition does not matter, so we choose $i$ and $s+1$ for notational convenience. Furthermore, to ease the ensuing presentation, we denote $|A_i| = a_i$ and $|B_i| = b_i$. Then, we obtain the following equations

$$\begin{aligned}
\frac{\sum_{A \in \tilde{\eta}_s(B)} p(A|x)}{p(B|x)} &= \sum_{A \in \tilde{\eta}_s(B)} \alpha \cdot \frac{\prod_{i=1}^{s+1} (|A_i|-1)!\, m(x_{A_i})}{\prod_{i=1}^{s} (|B_i|-1)!\, m(x_{B_i})} \\
&= \alpha \cdot \sum_{i=1}^{s} \sum_{j=1}^{b_i - 1} \sum_{A \in \tilde{\eta}_{i,j}(B)} \frac{(j-1)!(b_i - j - 1)!\, m(x_{A_i}) m(x_{A_{s+1}})}{(b_i - 1)!\, m(x_{B_i})} \\
&= \alpha \cdot \sum_{i=1}^{s} \sum_{j=1}^{b_i - 1} \frac{(j-1)!(b_i - j - 1)!}{(b_i - 1)!} \Big( \sum_{A \in \tilde{\eta}_{i,j}(B)} \frac{m(X_{A_i}) m(X_{A_{s+1}})}{m(X_{B_i})} \Big).
\end{aligned} \tag{8}$$

Given the above results, we define the following shorthands:

$$\overline{X}_{A_i} = \frac{1}{|A_i|} \cdot \sum_{x \in x_{A_i}} x; \quad S^2_{A_i} = \sum_{x \in x_{A_i}} x^2.$$

With simple algebra, we can verify that

$$\begin{aligned}
\frac{m(x_{A_i}) m(x_{A_{s+1}})}{m(x_A)} &= \frac{\displaystyle\int_{\theta \in \Theta} \exp\Big( - \sum_{x \in x_{A_i}} \frac{(x-\theta)^2}{2} \Big) d\theta \cdot \int_{\theta \in \Theta} \exp\Big( - \sum_{x \in x_{A_{s+1}}} \frac{(x-\theta)^2}{2} \Big) d\theta}{|\Theta| \displaystyle\int_{\theta \in \Theta} \exp\Big( - \sum_{x \in A} \frac{(x-\theta)^2}{2} \Big) d\theta} \\
&= \frac{\sqrt{\frac{2\pi}{a_i}} \exp\Big( - \frac{(S_i^2 + a_i \overline{X}_i^2)}{2} \Big) P_i(\Theta) \cdot \sqrt{\frac{2\pi}{a_{s+1}}} \exp\Big( - \frac{(S_{s+1}^2 + a_{s+1} \overline{X}_{s+1}^2)}{2} \Big) P_{s+1}(\Theta)}{|\Theta| \sqrt{\frac{2\pi}{a_i + a_{s+1}}} \exp\Big( - \frac{(S^2 + (a_i + a_{s+1}) \overline{X}^2)}{2} \Big) P_{i \cup s+1}(\Theta)} \\
&= \frac{\sqrt{2\pi}}{|\Theta|} \cdot \frac{\sqrt{a_i + a_{s+1}}}{\sqrt{a_i\, a_{s+1}}} \cdot \exp\Big( \frac{a_i \overline{X}_i^2 + a_{s+1} \overline{X}_{s+1}^2 - (a_i + a_{s+1}) \overline{X}^2}{2} \Big) \cdot \frac{P_i(\Theta)\, P_{s+1}(\Theta)}{P_{i \cup s+1}(\Theta)},
\end{aligned}$$

where $P_i(\Theta) := P(\theta \in \Theta | \theta \sim N(\overline{X}_i, \frac{1}{n_i}))$. If the samples $\{x_i\}_{i=1}^n$ satisfies that $\min_i x_i - \min(\Theta) > c$ and $\max(\Theta) - \max_i x_i > c$ for some $c > 0$, where for simplicity in presentattion we may choose $c = 3$ but note that the result holds for any $c > 0$ with some constant depending on $c$, then we have:

$$(0.997)^2 < \frac{P_i(\Theta)\, P_{s+1}(\Theta)}{P(\Theta)} < \frac{1}{0.997}.$$

On the other hand, for any $k$ sets $X_1, \cdots, X_k$ with sizes $n_1, \cdots, n_k$ and means $\overline{X}_1, \cdots, \overline{X}_k$, whose union is $X$ with size $n$ and mean $\overline{X}$, we have

$$\begin{aligned}\sum_{i=1}^{n} n_i \overline{X}_i^2 - n\overline{X}^2 &= \sum_{i=1}^{n} n_i \overline{X}_i^2 - \frac{\left(\sum_{i=1}^{n} n_i \overline{X}_i\right)^2}{n} \\ &= \sum_{i=1}^{n} \frac{n_i(n-n_i)}{n} \overline{X}_i^2 - \sum_{i \neq j} \frac{n_i n_j}{n} \overline{X}_i \overline{X}_j \\ &= \frac{1}{n} \sum_{i<j} n_i n_j (\overline{X}_i - \overline{X}_j)^2.\end{aligned}$$

The above result leads to the following equation

$$\frac{a_i \overline{X}_{A_i}^2 + a_{s+1} \overline{X}_{A_{s+1}}^2 - (a_i + a_{s+1}) \overline{X}^2}{2} = \frac{a_i \, a_{s+1} (\overline{X}_{A_i} - \overline{X}_{A_{s+1}})^2}{2(a_i + a_{s+1})}.$$

Combining all the results above, we obtain the following inequality

$$\frac{m(x_{A_i}) m(x_{A_{s+1}})}{m(x_{A_i \cup A_{s+1}})} \geq \frac{(0.997)^2 \sqrt{2\pi}}{|\Theta|} \cdot \frac{\sqrt{a_i + a_{s+1}}}{\sqrt{a_i \, a_{s+1}}} \cdot \exp\Big(\frac{a_i \, a_{s+1} (\overline{X}_{A_i} - \overline{X}_{A_{s+1}})^2}{2(a_i + a_{s+1})}\Big).$$

Given the above inequality, we can derive the following bounds for the term in (8):

$$\begin{aligned}\sum_{A \in \tilde{\eta}_s(B)} \frac{p(A|x)}{p(B|x)} &\geq \frac{\alpha (0.997)^2 \sqrt{2\pi}}{|\Theta|} \cdot \sum_{i=1}^{s} \sum_{j=1}^{b_i - 1} \Big(\frac{b_i}{j(b_i - j)}\Big)^{3/2} \\ &\quad \times \Bigg(\frac{1}{\binom{b_i}{j}} \sum_{A \in \tilde{\eta}_{i,j}(B)} \exp\Big(\frac{j(b_i - j)(\overline{X}_{A_i} - \overline{X}_{A_{s+1}})^2}{2b_i}\Big)\Bigg) \\ &\geq \frac{\alpha (0.997)^2 \sqrt{2\pi}}{|\Theta|} \cdot \sum_{i=1}^{s} \sum_{j=1}^{b_i - 1} \Big(\frac{b_i}{j(b_i - j)}\Big)^{3/2}. \end{aligned} \tag{9}$$

Direct computations lead to

$$\begin{aligned}\int_{x=1}^{b_i - 1} \Big(\frac{b_i}{x(b_i - x)}\Big)^{3/2} dx &\leq \sum_{j=1}^{b_i - 1} \Big(\frac{b_i}{j(b_i - j)}\Big)^{3/2} \\ &\leq \int_{x=1}^{b_i - 1} \Big(\frac{b_i}{x(b_i - x)}\Big)^{3/2} dx + 2\Big(\frac{b_i}{1(b_i - 1)}\Big)^{3/2}.\end{aligned}$$

The above result yields that

$$\frac{4(b_i - 2)}{\sqrt{(b_i - 1) b_i}} \leq \sum_{j=1}^{b_i - 1} \Big(\frac{b_i}{j(b_i - j)}\Big)^{3/2} \leq \frac{4(b_i - 2)}{\sqrt{(b_i - 1) b_i}} + 2^{5/2}.$$

When $b_i = 2, 3$, simple algebra indicates that $\sum_{j=1}^{b_i - 1} \Big(\frac{b_i}{j(b_i - j)}\Big)^{3/2} \geq 2$. Additionally, the left hand side in the inequalities above is always no less than 2 for $b_i \geq 4$. Invoking these results, we have the following lower bound

$$\sum_{A \in \tilde{\eta}_s(B)} \frac{p(A|x)}{p(B|x)} \geq \frac{\alpha (0.997)^2 \sqrt{2\pi}}{|\Theta|} \cdot \sum_{i=1}^{s} 2 \cdot I_{b_i \geq 2} \succsim \frac{\alpha s}{|\Theta|}.$$

Combining the above lower bound with equation (7), we eventually obtain the following evaluation with the ratio between consecutive terms $R(s)$

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

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:

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

Note that for each $A \in \rho_{s+1}(n)$, we can merge any two of its $s+1$ parts to get some $B \in \rho_s(n)$. The number of distinct ways to do so is exactly $\binom{s+1}{2}$. Also, for each $B \in \rho_s(n)$, the set $\eta(B)$ finds all $A \in \rho_{s+1}(n)$ such that they can merge some parts to get $B$. Thus, the index of the numerator in the second equation's right hand side counts each $A \in \rho_{s+1}(n)$ exactly $\binom{s+1}{2}$, from which the equation follows. Note that $n \gg s$ is required to prevent the case we have degenerate components. Although we only need $n \geq s$, but for simplicity and consistence in the proof argument, we choose to have $n \gg s$. Therefore, we achieve the conclusion of claim (7). □

The bound in the result of Theorem 1 does not require the data-generating distribution to be a mixture distribution. In particular, noting that empirical average of the exponential term in equation (9) goes to infinity as $n \to \infty$ provided that the true underlying distribution has finite and nonzero variance. This result is implied by the moment generating function of the Chi-squared distribution. Therefore, given the result of Theorem 1, we obtain the following corollary:

**Corollary 2.** *If the conditions in Theorem 1 hold as $n \to \infty$, then we obtain that*

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

Combining the results from Theorem 1 and Corollary 2, we can see that for any true distributions with finite but nonzero variance, the posterior probability of obtaining $s+1$ clusters will eventually exceed that of obtaining $s$ clusters, and their ratio will grow in an unbounded way. Provided the original distribution has a finite number of components, with more samples the result may even worsen since the model ultimately will fit an infinite number of clusters almost surely. However, in finite samples, their behavior depends more on the distribution's properties.

## 4 Gaussian Prior

Moving beyond the uniform prior, we consider the Gaussian prior on the parameter $\theta$, which has been widely employed with DPMM [3, 16]. In particular, we choose the prior density coming from the univariate Gaussian distribution $\mathcal{N}(0,\sigma^2)$ with fixed variance $\sigma > 0$, namely, $\pi(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{\theta^2}{2\sigma^2})$ for all $\theta \in \mathbb{R}$. Given this prior on $\theta$, we have the following asymptotic result regarding the lower bound of $R(s|\{x_i\}_{i=1}^n)$:

**Theorem 3.** *For the DPMM defined in* (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:*

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

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

**Remark.** *This result also holds in high probability in finite samples, provided sufficiently large number of samples. However, the number of samples required to attain this bound with fixed probability is highly dependent on the real data distribution.*

*Proof.* To ease the ensuing presentation, we reuse the notation from the proof of Theorem 1 in this proof. Direct computations yield the following result:

$$\frac{m(x_{A_1})m(x_{A_2})}{m(x_A)} = \frac{1}{\sqrt{\sigma^2}}\sqrt{\frac{(a_1+a_2)+\frac{1}{\sigma^2}}{(a_1+\frac{1}{\sigma^2})(a_2+\frac{1}{\sigma^2})}}\exp\left(\frac{1}{2}\left(\frac{a_1^2\overline{X}_1^2}{a_1+\frac{1}{\sigma^2}}+\frac{a_2^2\overline{X}_2^2}{a_2+\frac{1}{\sigma^2}}-\frac{a^2\overline{X}^2}{a+\frac{1}{\sigma^2}}\right)\right).$$

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

$$\sqrt{\tau}\cdot\sqrt{\frac{a_1+a_2+\tau}{(a_1+\tau)(a_2+\tau)}}\exp\Big(\frac{1}{2}\underbrace{\Big(\frac{a_1^2\overline{X}_1^2}{a_1+\tau}+\frac{a_2^2\overline{X}_2^2}{a_2+\tau}-\frac{a^2\overline{X}^2}{a+\tau}\Big)}_{F(\tau;X_1,X_2)}\Big)$$

.

Note that the term $F(\tau; x_{A_1}, x_{A_2})$ is nonnegative at zero since

$$F(0; x_{A_1}, x_{A_2}) = \frac{a_1 a_2}{a_1 + a_2}(\overline{X}_1 - \overline{X}_2)^2 \geq 0.$$

Solving a quadratic function gives that $F(\tau; x_{A_1}, x_{A_2}) = 0$ has its positive root on:

$$\begin{cases} \frac{1}{4}\Big[\sqrt{8a_1a_2\frac{(\overline{X}_1-\overline{X}_2)^2}{\overline{X}_1\overline{X}_2}+Q^2}+Q\Big], & \text{if } \overline{X}_1\overline{X}_2 > 0 \\ \frac{1}{4}\Big[\sqrt{8a_1a_2\frac{(\overline{X}_1-\overline{X}_2)^2}{\overline{X}_1\overline{X}_2}+Q^2}+Q\Big] \text{ or } \emptyset, & \text{if } \overline{X}_1\overline{X}_2 < 0 \\ \emptyset & \text{if } \overline{X}_1 = 0, \overline{X}_2 \neq 0 \text{ or } \overline{X}_1 \neq 0, \overline{X}_2 = 0 \\ \mathbb{R}^+ & \text{if } \overline{X}_1 = \overline{X}_2 = 0, \end{cases}$$

where $Q := a_2\frac{a_2\overline{X}_2}{a_1\overline{X}_1} + a_1\frac{a_1\overline{X}_1}{a_2\overline{X}_2} - 2(a_1+a_2)$.

If there is no positive root or every positive number is a root, then $F(\tau; x_{A_1}, x_{A_2}) \geq 0$ for any $\tau > 0$. Otherwise, as shown in the above, there exists a unique positive root $r_+(X_1, X_2) := \frac{1}{4}\Big[\sqrt{8a_1a_2\frac{(\overline{X}_1-\overline{X}_2)^2}{\overline{X}_1\overline{X}_2}+Q^2}+Q\Big]$, a random variable depending on $a_1, a_2$, whose probability density function favors larger and larger values as long as one of $a_1, a_2$ goes to infinity. That is, the root grows larger in probability as $a_1, a_2$ increases, where rate it scales up depends on the real data distribution. For any fixed $\tau$, as $a_1 + a_2$ goes to infinity, it follows that for most partitions of $x_A$ into $x_{A_1}$ and $x_{A_2}$, $\tau$ falls into $[0, r_+(X_1, X_2)]$, so we have that $F(\tau; x_{A_1}, x_{A_2}) \geq 0$.

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 $F(\tau; x_{B_1}, x_{B_2}) \geq 0$ asymptotically holds for any fixed proportion (less than 1) for all the partitions of $B_i$. Note that for any positive integers $a_1, a_2$ and nonnegative number $\tau$, we have

$$\frac{a_1 + a_2 + \tau}{(a_1+\tau)(a_2+\tau)} > \frac{1}{2} \cdot \frac{1}{1+\tau} \cdot \frac{a_1+a_2}{a_1 a_2}.$$

Then, for sufficiently large $n$ we have:

$$\begin{aligned} \sum_{A\in\tilde{\eta}_s(B)} \frac{p(A|x)}{p(B|x)} &= \alpha \cdot \sum_{i=1}^{s}\sum_{j=1}^{b_i-1} \frac{(j-1)!(b_i-j-1)!}{(b_i-1)!}\Big(\sum_{A\in\tilde{\eta}_{i,j}(B)} \frac{m(X_{A_i})m(X_{A_{s+1}})}{m(X_{B_i})}\Big) \\ &\overset{w.h.p.}{\geq} C_0\alpha\sqrt{\tau} \cdot \sum_{i\in U(B)}\sum_{j=1}^{b_i-1} \frac{(j-1)!(b_i-j-1)!}{(b_i-1)!} \\ &\qquad \times\Big(\sum_{A\in\tilde{\eta}_{i,j}(B)} \sqrt{\frac{b_i+\tau}{(j+\tau)(b_i-j+\tau)}}\Big) \\ &\geq C_0\alpha\sqrt{\tau} \cdot \sum_{i\in U(B)}\sum_{j=1}^{b_i-1} \frac{(j-1)!(b_i-j-1)!}{(b_i-1)!} \\ &\qquad \times\Big(\sum_{A\in\tilde{\eta}_{i,j}(B)} \frac{1}{\sqrt{2(1+\tau)}}\sqrt{\frac{b_i}{j(b_i-j)}}\Big) \\ &\geq \frac{C\sqrt{\tau}}{1+\sqrt{\tau}} \cdot \alpha, \end{aligned} \tag{11}$$

with high probability where $C$ is a universal constant. Here, the second step follows with high probability by our previous argument where $C_0$ is a positive universal constant between zero and one, and the last step follows by a similar argument as in the case of uniform prior with $C$ being some constant independent of $\alpha, n, s$, except that here it is possible to have $|U(A)| \ll s$, so the result can only be bounded by a constant multiple of $\alpha \cdot \frac{\sqrt{\tau}}{1+\sqrt{\tau}}$ without the $s$ factor in the uniform case.

Finally, note that as $n$ goes to infinity, the above result holds in probability 1. Therefore, we obtain that

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

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

The result of Theorem 3 holds in the asymptotic regime. Its performance in finite samples is more complicated to study, and always heavily depends on the original distribution. It would be of interest to characterize the finite-sample behaviors for 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.

## 5 Discussion

In this paper, we establish lower bounds on the ratio of posterior probabilities $R(s|\{x_i\}_{i=1}^n)$ for the number of clusters under several settings of prior distributions on the parameter space. The aim of our study is to increase our understanding of the posterior distribution of the number of clusters in both the non-asymptotic and asymptotic regimes. As our results suggest, comparing to the popular application of DPMM to density estimation, DPMM is not as succesful in fitting the number of components due to the combinatorial structure in the prior, which has an effect (that may not be favorable) on the posterior distribution of the number of clusters even when the sample size $n$ goes to infinity.

The current work lays useful foundations for several future research directions that we now discuss. One interesting open problem is that when the original distribution contains infinite components. Does DPMM guarantee an infinite number of clusters in this case? This is important in the case where the data distribution is dynamic and continues to generate new components. In this case, even if DPMM guarantees an infinite number of components asymptotically, it is not clear whether it does so in a rate that matches that of the original distribution. In particular, an interesting case is when the real distribution is indeed a Dirichlet process of normals or just a general Dirichlet process. Then, does DPMM generate a posterior number of clusters in the same rate as implied by the Dirichlet process?

Another important problem is to investigate a natural way to resolve inconsistency when the distribution is finite-component or to resolve mismatch in the rate of growth when the distribution is infinite-component. In the literature, truncation of the number of clusters is a popular way to deal with the growing number of clusters in DPMM. Recently, this method has been shown to yield consistency with the number of clusters when the true data generated distribution is in fact a finite mixtures [12]. However, the truncation method generally requires a tuning with the separation among parameters or the lower bound for the ratios of the clusters, which are not available in practice. The question, therefore, is whether there exists a natural way to correct the problem instead of truncating the number of clusters?

## References


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

[2] D. Blackwell and J. MacQueen. Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

[3] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[4] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[5] E. Fox, E. Sudderth, M. I. Jordan, and A. Willsky. A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5:1020–1056, 2011.

[6] S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. *Bayesian nonparametrics*, 28:35, 2010.

[7] S. Ghosal, J. K. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[8] S. Ghosal and A. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29:1233–1263, 2001.

[9] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 2007.

[10] S. Ghosal and A. van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35:697–723, 2007.

[11] J. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics. Springer, 2003.

[12] A. Guha, N. Ho, and L. Nguyen. On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Arxiv preprint Arxiv: 1901.05078*, 2019.

[13] J. P. Huelsenbeck and P. Andolfatto. Inference of population structure under a Dirichlet process model. *Genetics*, 175(4):1787–1802, 2007.

[14] N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–1109, 2004.

[15] A. Lo. On a class of Bayesian nonparametric estimates I: Density estimates. *Annals of Statistics*, 12(1):351–357, 1984.

[16] S. MacEachern and P. Mueller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.

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

[18] J. W. Miller and M. T. Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in neural information processing systems*, pages 199–206, 2013.

[19] J. W. Miller and M. T. Harrison. Inconsistency of Pitman-Yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15(1):3333–3370, 2014.

[20] E. Otranto and G. M. Gallo. A nonparametric Bayesian approach to detect the number of regimes in markov switching models. *Econometric Reviews*, 21(4):477–496, 2002.

[21] J. Paisley, C. Wang, D. Blei, and M. I. Jordan. Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:256–270, 2015.

[22] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[23] A. Rodriguez, D. Dunson, and A. Gelfand. The nested Dirichlet process. *J. Amer. Statist. Assoc.*, 103(483):1131–1154, 2008.

[24] J. Rousseau. Rates of convergence for the posterior distributions of mixtures of Betas and adaptive nonparametric estimation of the density. *Annals of Statistics*, 38:146–180, 2010.

[25] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Annals of Statistics*, 29:687–714, 2001.

[26] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:1566–1581, 2006.

[27] S. Walker, A. Lijoi, and I. Prunster. On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.*, 35(2):738–746, 2007.