Journal of Machine Learning Research 24 (2023) 1-32 Submitted 4/22; Revised 5/23; Published 5/23

Consistent Model-based Clustering using the Quasi-Bernoulli Stick-breaking Process

Cheng Zeng CZENG1@UFL.EDU
Department of Statistics
University of Florida
Gainesville, FL 32611, USA

Jeffrey W Miller JWMILLER@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

Leo L Duan LI.DUAN@UFL.EDU
Department of Statistics
University of Florida
Gainesville, FL 32611, USA

Editor: Ryan Adams

Abstract

In mixture modeling and clustering applications, the number of components and clusters is often not known. A stick-breaking mixture model, such as the Dirichlet process mixture model, is an appealing construction that assumes infinitely many components, while shrinking the weights of most of the unused components to near zero. However, it is well-known that this shrinkage is inadequate: even when the component distribution is correctly specified, spurious weights appear and give an inconsistent estimate of the number of clusters. In this article, we propose a simple solution: when breaking each mixture weight stick into two pieces, the length of the second piece is multiplied by a quasi-Bernoulli random variable, taking value one or a small constant close to zero. This effectively creates a soft truncation and further shrinks the unused weights. Asymptotically, we show that as long as this small constant diminishes to zero at a rate faster than $o(1/n^2)$, with n the sample size and given data from a finite mixture model, the posterior distribution will converge to the true number of clusters. In comparison, we rigorously explore Dirichlet process mixture models using a concentration parameter that is either constant or rapidly diminishes to zero—both of which lead to inconsistency for the number of clusters. Our proposed model is easy to implement, requiring only a small modification of a standard Gibbs sampler for mixture models. In simulations and a data application of clustering brain networks, our proposed method recovers the ground-truth number of clusters, and leads to a small number of clusters.

Keywords: Consistent Clustering, Exchangeable Partition Probability Function, Sparse Simplex.

1. Introduction

Mixture models are frequently used to analyze data with unknown group/cluster structure. They give a generative view of the data $y = y_{1:n} = (y_1, \ldots, y_n)$, and provide uncertainty quantification on the cluster assignments. To review the main idea, suppose

$$y_i \overset{iid}{\sim} \sum_{k=1}^{K} w_k \mathcal{F}(\theta_k)$$

for $i = 1, \ldots, n$, where $\mathcal{F}(\theta_k)$ is a distribution parameterized by θ_k, and $w_1, \ldots, w_K \geq 0$ such that $\sum_{k=1}^{K} w_k = 1$. Using the Bayesian framework that posits a prior distribution for the weights $w = w_{1:K} = (w_1, \ldots, w_K)$ and the parameters $\theta = \theta_{1:K} = (\theta_1, \ldots, \theta_K)$, one can infer the posterior distribution of the weights w and the parameters θ, as well as the assignments of data points to mixture components, which yields a clustering of the data (Fraley and Raftery, 2002).

In practice, in addition to w_k and θ_k, we usually do not know the number of clusters either. Stick-breaking models provide an appealing solution in which the number of mixture components K is infinite, and the number of clusters in the data (that is, the number of components that the data are assigned to) can be any finite number. A general form of a stick-breaking model for the mixture weights w is

$$w_1 = v_1 \text{ and } w_k = v_k \prod_{l=1}^{k-1} (1 - v_l) \text{ for } k = 2, 3, \ldots, \tag{1}$$

where $v_1, v_2, \ldots$ are drawn from some prior distribution. The interpretation is that starting from a stick of length 1, for each k we break off a proportion $v_k \in [0, 1]$ from the remaining stick, and use it as the weight w_k.

Many priors have been proposed for the distribution of v_k. Perhaps the most widely used one is $v_k \sim \text{Beta}(1, \alpha)$, with which the probability distribution $\sum_{k=1}^{\infty} w_k \delta_{\theta_k}(\cdot)$ yields a realization of the Dirichlet process with concentration parameter α (Sethuraman, 1994), where $\delta_x(\cdot)$ represents the Dirac measure which satisfies $\delta_x(A) = \mathbb{1}(x \in A)$ for any measurable set A. More generally, when $v_k \sim \text{Beta}(1 - d, \alpha + kd)$ with $0 \leq d < 1$ and $\alpha > -d$, one obtains the Pitman–Yor process with discount parameter d and strength parameter α (Pitman and Yor, 1997; Ishwaran and James, 2001).

When the true data-generating distribution is a finite mixture from the assumed family, these models tend to shrink most of w_k's close to zero, leading to a small number of clusters in the posterior. However, Miller and Harrison (2013, 2014) showed a striking result: neither the Dirichlet process prior nor the Pitman–Yor process prior allows the posterior distribution on the number of components K to converge to the true number of clusters as n increases, even when the family of component distributions $\mathcal{F}$ is correctly specified.

To address this issue, we develop a prior on the w_k's that yields stronger shrinkage while remaining easy to use. Specifically, we modify the canonical stick-breaking construction as follows: at each break, we multiply the remaining proportion $(1 - v_k)$ by a discrete random variable that takes value either 1 or ϵ. When the latter happens at the L-th stick, the tail probability $\sum_{k=L+1}^{\infty} w_k$ is strictly bounded by ϵ. We show that if ϵ is chosen in a sample-size-dependent way such that $\epsilon(n) = o(1/n^{2+r})$ (with r a non-negative integer that depends

on α), then one can obtain posterior consistency for the number of clusters. In the special case of $\epsilon = 0$, this model reduces to a finite mixture model with a prior on the number of components (Miller and Harrison, 2018), which also yields posterior consistency for the number of clusters. Meanwhile, in the special case of $\epsilon = 1$ and $v_k \sim \text{Beta}(1, \alpha)$, this model reduces to a Dirichlet process mixture.

Therefore, using $\epsilon \in (0, 1)$ effectively interpolates between these two extremes. Compared to $\epsilon = 0$, using $\epsilon \in (0, 1)$ avoids having a singularity at $w_k = 0$. This relaxes the parameter space in a way that allows the Markov chain Monte Carlo (MCMC) sampler to more efficiently add and remove clusters, rather than employing an explicit discrete search over K. Further, it makes the technique compatible with more complex stick-breaking models, such as the ones involving kernels (Dunson and Park, 2008), geospatial processes (Rodríguez et al., 2010), and external predictors (Ren et al., 2011). On the other hand, compared to $\epsilon = 1$, it allows the model to behave effectively like a finite mixture with a prior on the number of components, leading to superior control on the number of clusters. We illustrate these advantages in simulations and a data application in clustering brain networks, using a mixture of low-rank probit models. A software implementation and the steps needed to replicate the results in this paper are provided on https://github.com/zeng-cheng/quasi-bernoulli-stick-breaking.

2. Quasi-Bernoulli Stick-breaking Process

In this section, we formally construct the quasi-Bernoulli stick-breaking process, and provide a theoretical analysis of this process and the corresponding mixture model.

2.1 Prior Construction

In the general form of the stick-breaking construction (Equation (1)), if the proportion v_L at step L is very close to 1, then w_L will take almost all the remaining sticks, resulting in $w_k \approx 0$ for $k \geq L + 1$. Based on this intuition, we introduce the following stick-breaking process:

$$\begin{aligned} &w_1 = v_1, \quad w_k = v_k \prod_{l=1}^{k-1} (1 - v_l), \text{ for } k \geq 2, \\ &v_k = 1 - b_k \beta_k, \\ &b_k \overset{iid}{\sim} \tilde{p}\delta_1(\cdot) + (1 - \tilde{p})\delta_\epsilon(\cdot), \\ &\beta_k \overset{iid}{\sim} \text{Beta}(\alpha, 1). \end{aligned} \tag{2}$$

Each b_k follows a discrete distribution such that $b_k = 1$ with probability $\tilde{p} \in (0, 1)$, and $b_k = \epsilon$ with probability $1 - \tilde{p}$, for some small $\epsilon \in (0, 1)$. We refer to b_k as a quasi-Bernoulli random variable, since it resembles the standard Bernoulli supported on $\{0, 1\}$. We refer to Equation (2) as the quasi-Bernoulli stick-breaking process (or simply the "quasi-Bernoulli process"). With this prior on weights $w_1, w_2, \ldots$, we obtain an infinite mixture model by letting $y_i \overset{iid}{\sim} \sum_{k=1}^{\infty} w_k \mathcal{F}(\theta_k)$, where $\theta_k \overset{iid}{\sim} \mathcal{G}$ from some base measure $\mathcal{G}$; we refer to this as a quasi-Bernoulli mixture model. Moreover, this marginal representation of the mixture model can be equivalently represented in a conditional form with the introduction of the

latent assignment variable c_i for each data point y_i. Specifically, with $c_i = k$ representing the event that y_i is drawn from the mixture component k,

$$\begin{aligned} \theta_k &\overset{iid}{\sim} \mathcal{G} \text{ for } k \geq 1, \\ c_i \mid w &\overset{iid}{\sim} \text{Categorical}(w), \\ y_i \mid c_i, \theta &\overset{indep}{\sim} \mathcal{F}(\theta_{c_i}) \text{ for } i = 1, \ldots, n. \end{aligned} \tag{3}$$

This conditional representation is useful if one is interested in using the model to perform model-based clustering.

Note that if $\epsilon = 1$, then we would have $v_k = 1 - \beta_k \sim \text{Beta}(1, \alpha)$, yielding the stick-breaking representation of the Dirichlet process $\text{DP}(\alpha, \mathcal{G})$; whereas if $\epsilon = 0$, then we would have a random truncation on $(w_1, w_2, \ldots)$. Using $\epsilon \in (0, 1)$ yields a soft truncation that provides advantages from both of these extremes.

Before we move into more technical discussions, we first illustrate an intuition for why the Dirichlet process with a fixed α (as a data generating mechanism) tends to produce small clusters, and how our proposed prior mitigates this. Consider the scenario where we have n data points already assigned to K clusters, and there are m new data points to be assigned. Assume that there are only K clusters in the population, but we are modeling the data as drawn from a Dirichlet process mixture model. Ideally we want to assign all of the m new data into the existing K clusters. The prior probability of adding one or more new clusters can be calculated using the predictive rule (as in the "restaurant process") $p(c_i \mid c_1, \ldots, c_{i-1})$ for $i \in \{n+1, \ldots, n+m\}$ recursively for m times. For the Dirichlet process with concentration parameter α, this probability has a closed form:

$$p\left(\sum_{l=1}^{m} \mathbb{1}(c_{n+l} > K) > 0 \mid c_1, \ldots, c_n\right) = 1 - \prod_{l=1}^{m} \left(\frac{n+l-1}{n+l-1+\alpha}\right),$$

which follows directly from the predictive distribution under the Dirichlet process (Blackwell and MacQueen, 1973, Equation (2)). Although the probability is small for $m = 1$, it increases rapidly and becomes non-trivial as m grows. Note that the above event includes not only assigning all m data points into a single new cluster, but also having them scattered into several new clusters; hence this is the union of all outcomes of having small clusters.

For the quasi-Bernoulli process, although we do not have a simple closed-form expression for the above probability, we can numerically calculate it based on the partition probability function Equation (4) introduced in the next section. In Figure 1, we examine the case when given two existing clusters with $n_1 = n_2 = 50$, and assigning additional m data points. It can be seen that the prior probability of creating new cluster(s) quickly increases in the Dirichlet process, whereas the quasi-Bernoulli processes (with $\alpha = 1$ and two values for $\tilde{p}$) substantially slow down the growth of this probability. We also provide results with another rate of $\epsilon(n)$ in the appendix.

We will carefully examine the posterior behaviors of the above models, including comparing with the Dirichlet process with $\alpha = \alpha(n)$ tending to zero as $n \to \infty$.

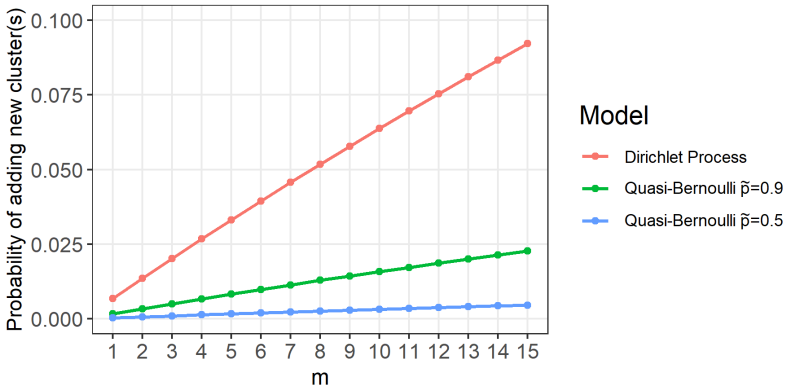


Figure 1: Under the prior, the Dirichlet process exhibits rapid growth in the probability of adding one or more new clusters for m future data points ($n = 100$), favoring the creation of additional clusters *a priori*. Meanwhile, the quasi-Bernoulli process exhibits much slower growth of this probability. For the quasi-Bernoulli process, we use $\epsilon = \epsilon(n, m) = 1/(n+m)^{2.1}$ and $\alpha = 1$ as suggested in our theory Theorem 5. For the Dirichlet process, we use $\alpha = 0.69$, which has the prior expected number of clusters close to the one under the quasi-Bernoulli process with $\tilde{p} = 0.9$ (see Table 1).

2.2 Exchangeable Random Partitioning

A large class of stick-breaking models enjoys the property of partition exchangeability—that is, the probability distribution of the induced partition only depends on the cluster sizes, and is invariant to any permutation of the cluster index (see Pitman, 1995, proposition 5). Letting $i \in \{1, \ldots, n\}$ be the data index, if there are t unique values in the cluster assignments $c = (c_1, \ldots, c_n)$, then we can form a corresponding partition $\mathcal{A} = \{A_1, \ldots, A_t\}$ in the following way: (1) initialize $A_1 = \{1\}$ and $t = 1$; (2) sequentially for $i = 2, \ldots, n$, if $c_i = c_j$ for any $j \leq i - 1$ and $j \in A_k$, then add i to the same set A_k containing j; otherwise create a new set $A_{t+1} = \{i\}$ and increment t to $t + 1$.

Theorem 1 (Exchangeable Partition Probability Function) *The probability mass function of the random partition $\mathcal{A} = \{A_1, \ldots, A_t\}$ induced by c in the quasi-Bernoulli stick-breaking process is*

$$p_{\epsilon,n}(\mathcal{A}) = \frac{\alpha^t \Gamma(\alpha)}{\Gamma(n+\alpha)} \left(\prod_{j=1}^{t} \Gamma(n_j + 1) \right) \sum_{\sigma \in S_t} \prod_{j=1}^{t} \frac{\tilde{p} + (1-\tilde{p}) I_\epsilon(g_{j+1}(\sigma) + \alpha, n_{\sigma_j} + 1)/\epsilon^\alpha}{g_j(\sigma) + \alpha(1-\tilde{p})(1-\epsilon^{g_j(\sigma)})} \tag{4}$$

where $\sigma = (\sigma_1, \ldots, \sigma_t)$ is a permutation of $\{1, \ldots, t\}$, S_t is the set of all permutations of $\{1, \ldots, t\}$, $n_j = |A_j|$, $g_j(\sigma) = \sum_{l=j}^{t} n_{\sigma_l}$, $\Gamma(x) = \int_0^\infty z^{x-1} e^{-z}\, dz$ and $I_\epsilon(q_1, q_2)$ is the cumulative distribution function of Beta(q_1, q_2) evaluated at ϵ.

For conciseness, we defer all the proofs to the appendix. Using $p_{\epsilon,n}(\mathcal{A})$, we can substantially simplify the model in Equations (2) and (3) into an equivalent generative process:

$$\begin{aligned} \mathcal{A} &\sim p_{\epsilon,n}(\mathcal{A}), \\ \theta_k \mid \mathcal{A} &\overset{iid}{\sim} \mathcal{G} \text{ for } k = 1, \ldots, t, \\ y_i \mid \mathcal{A}, \theta &\overset{indep}{\sim} \mathcal{F}(\theta_j) \text{ for } i \in A_j,\ A_j \in \mathcal{A}. \end{aligned} \tag{5}$$

We now use the above representation to study the asymptotics of the clustering when $n \to \infty$.

2.3 Consistent Estimation of the Number of Clusters

By definition, the number of clusters is $t = |\mathcal{A}|$. We use T to denote the associated random variable in the model. Let $\mathcal{H}_t(n)$ denote the set of all partitions of $\{1, \ldots, n\}$ into t disjoint sets. Using Equation (5), the marginal posterior of T is

$$p_\epsilon(T = t \mid y_{1:n}) = \frac{\sum_{\mathcal{A} \in \mathcal{H}_t(n)} p(y_{1:n} \mid \mathcal{A}) p_{\epsilon,n}(\mathcal{A})}{\sum_{\mathcal{A} \in \cup_{t'=1}^{n} \mathcal{H}_{t'}(n)} p(y_{1:n} \mid \mathcal{A}) p_{\epsilon,n}(\mathcal{A})}, \tag{6}$$

where $p(y_{1:n} \mid \mathcal{A}) = \prod_{A \in \mathcal{A}} m(y_A)$, $y_A = (y_i : i \in A)$, and $m(y_A) = \int_\Theta \big(\prod_{i \in A} f_\theta(y_i)\big)\, d\mathcal{G}(\theta)$. Here, f_θ denotes the density of the component distribution $\mathcal{F}(\theta)$.

Suppose the data are generated from k_0 mixture components, with k_0 a fixed and finite number. We establish general conditions under which $p_\epsilon(T = k_0 \mid y_{1:n}) \to 1$ as $n \to \infty$. Our proof involves two main parts: (i) we establish that this consistency property holds for the finite-dimensional model obtained by setting $\epsilon = 0$; and (ii) we bound the total variation distance between the prior distributions $p_{\epsilon,n}(\mathcal{A})$ and $p_{0,n}(\mathcal{A})$. We then show that this implies posterior consistency.

Consider the case when $\epsilon = 0$, that is, when b_k in Equation (2) is a Bernoulli random variable with success rate $\tilde{p}$. In this case, we refer to Equation (2) with $\epsilon = 0$ as the Bernoulli stick-breaking process. Let $K = \min\{k : b_k = 0\}$. When this occurs, we have $w_{K+1} = w_{K+2} = \cdots = 0$, and therefore, w is effectively K-dimensional. Observe that K follows a geometric distribution: $\pi_K(k) = \tilde{p}^{k-1}(1-\tilde{p})$ for $k \in \{1, 2, \ldots\}$. Thus, the Bernoulli stick-breaking process has the following equivalent representation:

$$\begin{aligned} K &\sim \text{Geometric}(1 - \tilde{p}), \\ v_k &\overset{iid}{\sim} \text{Beta}(1, \alpha), \quad \text{for } k = 1, \ldots, K-1; \quad v_K = 1, \\ w_1 = v_1, &\quad w_k = v_k \prod_{l=1}^{k-1} (1 - v_l), \quad \text{for } k \geq 2. \end{aligned} \tag{7}$$

The following result establishes the exchangeable partition probability function for the Bernoulli stick-breaking process.

Lemma 2 *The probability mass function of the random partition $\mathcal{A} = \{A_1, \ldots, A_t\}$ in the Bernoulli stick-breaking process is*

$$p_{\epsilon=0,n}(\mathcal{A}) = \frac{\alpha^t \Gamma(\alpha)}{\Gamma(n+\alpha)} \left(\prod_{j=1}^{t} \Gamma(n_j + 1) \right) \sum_{\sigma \in S_t} \prod_{j=1}^{t} \frac{\tilde{p} + \mathbb{1}(j = t)(1 - \tilde{p})/(\alpha \text{B}(\alpha, n_{\sigma_t} + 1))}{g_j(\sigma) + \alpha(1 - \tilde{p})}$$

where $n_j = |A_j|$, $g_j(\sigma) = \sum_{l=j}^{t} n_{\sigma_l}$, *and* $\mathrm{B}(q_1, q_2) = \int_0^1 z^{q_1-1}(1-z)^{q_2-1}\,dz$, *the Beta function.*

We now establish that the Bernoulli stick-breaking process mixture model (quasi-Bernoulli mixture model with $\epsilon = 0$) exhibits posterior consistency for K (the number of non-zero w_k's) and T (the number of clusters). To clarify, even for a large n, T could be less than the number of components K, if some component does not have any data point assigned to it. Let Ω denote the set of parameter tuples $\phi := (k, w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$ such that $k \in \{1, 2, \ldots\}$, $w_1, \ldots, w_k > 0$, $\sum_{l=1}^{k} w_l = 1$, and $\theta_1, \ldots, \theta_k \in \Theta$, the parameter space. Let Ω' denote the subset of Ω such that $\theta_i \neq \theta_j$ for all $i \neq j$. Further, let P_ϕ denote the mixture distribution $P_\phi := \sum_{l=1}^{k} w_l \mathcal{F}(\theta_l)$. When $\phi \in \Omega'$ is identifiable from P_ϕ up to permutation of the mixture components, we can define a transformation $\eta : \Omega \to \Omega'$ such that the parameter $\phi' = \eta(\phi)$ is fully identifiable from $P_{\phi'}$. See Nobile (1994, Section 3.2) for the details. Any prior on Ω defines an induced prior on Ω' through η.

Theorem 3 *Assume $\phi \in \Omega'$ is identifiable up to permutation of the mixture components. Let Π_0 be the prior on Ω under the model defined by Equations* (3) *and* (7)*, and assume $\Pi_0(\{\phi : \exists\ i \neq j$ such that $\theta_i = \theta_j\}) = 0$. Let Π_0' be the corresponding prior on Ω' induced by η. Then there is a subset $\Omega_0' \subset \Omega'$ with $\Pi_0'(\Omega_0') = 1$ such that for any $\phi_0 = (k_0, w_1^0, \ldots, w_{k_0}^0, \theta_1^0, \ldots, \theta_{k_0}^0) \in \Omega_0'$, if $y_1, y_2, \ldots \mid \phi_0 \overset{iid}{\sim} P_{\phi_0}$ and the component density f_θ is continuous (with respect to θ) at each θ_k^0, then as $n \to \infty$, we have*

$$p_{\epsilon=0}(K = k_0 \mid y_{1:n}) \to 1 \quad \text{a.s.}[P_{\phi_0}],$$
$$p_{\epsilon=0}(T = k_0 \mid y_{1:n}) \to 1 \quad \text{a.s.}[P_{\phi_0}],$$

where a.s.$[P_{\phi_0}]$ *denotes almost surely convergence under the probability distribution P_{ϕ_0}.*

The first part of the result is a corollary of Nobile (1994, Proposition 3.5), the proof of which is an application of the Doob's theorem. The intuition for the second part is that since $w_1^0, \ldots, w_{k_0}^0$ are positive and do not change with n, we can expect that at least some data will be assigned to each component $k = 1, \ldots, k_0$, and thus that K and T will match in the posterior.

Now, consider the case of $\epsilon > 0$. Intuitively, when ϵ is small, we would expect the posterior to behave similarly to the case of $\epsilon = 0$. Formally, we show that this is indeed the case when $\epsilon = \epsilon(n) \to 0$ at an appropriate rate as $n \to \infty$. To show this, we employ the following bound on the distance between the partition distributions for $\epsilon > 0$ and $\epsilon = 0$ under the prior.

Theorem 4 (Prior equivalence as $\epsilon(n) \to 0$) *Assume $\epsilon \leq 1/n$. Under the quasi-Bernoulli priors, the total variation distance between the partition distributions for $\epsilon > 0$ and $\epsilon = 0$ satisfies the bound*

$$\sup_{A \in \mathscr{A}} |p_{\epsilon,n}(A) - p_{0,n}(A)| \leq \sqrt{\frac{\alpha n \epsilon}{2(\alpha + 1 - \alpha\epsilon n)}}$$

where $\mathscr{A}$ denotes the set of all subsets of $\cup_{t=1}^{n} \mathcal{H}_t(n)$, and $p_{\epsilon,n}(A) = \sum_{\mathcal{A} \in A} p_{\epsilon,n}(\mathcal{A})$. In particular, if $\epsilon(n) = o(1/n)$, then

$$\sup_{A \in \mathscr{A}} |p_{\epsilon(n),n}(A) - p_{0,n}(A)| \xrightarrow[n \to \infty]{} 0.$$

The interpretation of this result is that if we control ϵ to be slightly smaller than $1/n$, then we have a stick-breaking model supported in the infinite-dimensional space that asymptotically approximates the finite-dimensional model with posterior consistency guarantees.

Now, moving to the posterior, we have the consistency of the quasi-Bernoulli model for the number of clusters.

Theorem 5 (Posterior consistency) *Under the same assumptions and notations of Theorem 3, if $\epsilon(n) = o(1/n^{2+r})$, where r is the integer such that $\max(\alpha - 1, 0) \leq r < \alpha$, then*

$$p_{\epsilon(n)}(T = k_0 \mid y_{1:n}) \xrightarrow[n\to\infty]{} 1 \quad \text{a.s.}[P_{\phi_0}].$$

Note that here we assume $\epsilon(n) = o(1/n^{2+r})$ rather than $o(1/n)$ as in Theorem 4, however, we expect that the involved inequalities could be tightened further.

It should also be noted that letting $\epsilon(n)$ depend on n makes the resulting sequence of models no longer projective. That is, the model for n data points does not coincide with the distribution obtained by taking the model for $n+1$ data points and integrating over data point $n+1$. However, to achieve certain optimal asymptotic behaviors such as consistency, it is common to calibrate the prior based on the sample size (for example, see Castillo et al., 2015 on the choice of prior for variable selection). Alternatively, one could always use $\epsilon = 0$, which achieves both consistency and projectivity, but this is less flexible and less efficient in terms of computation.

2.4 Comparison with the Asymptotic Behavior of the Dirichlet Process

The results of Miller and Harrison (2013, 2014) on the inconsistency of the Dirichlet process mixture model show that for a fixed value of the concentration parameter α, the model asymptotically over-estimates the number of clusters on data from a finite mixture.

Our theoretic finding in the quasi-Bernoulli raises a tempting question: can we achieve consistency with a Dirichlet process mixture if we let $\alpha = \alpha(n)$ go to 0 at an appropriate rate as n increases? To our knowledge, this remains an open question. However, here we provide a partial answer (in the negative), by showing that if $\alpha(n) \to 0$ too fast, then the Dirichlet process remains inconsistent for the number of clusters.

Lemma 6 *Suppose the data are $y_1, \ldots, y_n \overset{iid}{\sim} 0.5\,\mathrm{N}(0,1) + 0.5\,\mathrm{N}(\kappa, 1)$ for any fixed $\kappa \in \mathbb{R}$, where $\mathrm{N}(\mu, \sigma^2)$ is the Gaussian distribution with density $f_{\mu,\sigma^2}(x) \propto \exp\{-(x-\mu)^2/(2\sigma^2)\}$. Consider a Dirichlet process mixture model with the mixture components $\mathcal{F}(\theta) = \mathrm{N}(\theta, 1)$, the base measure $\mathcal{G} = \mathrm{N}(0,1)$ (the prior on the parameter θ), and concentration parameter $\alpha = \alpha(n)$ such that $\alpha(n) = o(\exp(-a_0\, n))$ for some constant $a_0 > 1/2 + \kappa^2/4$. Then $p(T = 2 \mid y_{1:n}) \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.*

To clarify, the purpose of the above result is not to provide practical guidance on choosing the rate of $\alpha(n) \to 0$. Indeed, the problem of the Dirichlet process mixture is usually over-estimation rather than underestimation of the number of clusters in the limit (Yang et al., 2020). Rather, this result shows that if consistency could be achieved for the Dirichlet process mixture model under a sample-size-dependent $\alpha(n)$, the rate of this hyper-parameter needs to satisfy both an upper and a lower bounds, which may turn out to be practically

challenging for the users—there is a somewhat delicate sensitivity issue. In comparison, a strength of the quasi-Bernoulli mixture is that for obtaining consistency ϵ only needs to satisfy one upper bound $o(1/n^{2+\alpha})$, which means the user can simply use a small ϵ such as $1/n^{2+\alpha}$ (or smaller), without worrying about impacting the consistency. We provide empirical comparison via simulations with different cases of $\alpha(n)$ and $\epsilon(n)$ in the Appendix B.2.

The above result does not exclude the possibility that the Dirichlet process with $\alpha \to 0$ at a slower rate could achieve consistency. For example, recently Ohn and Lin (2023) show that setting $\alpha \approx n^{-a_1}$ for some positive number a_1 will guarantee $\mathrm{pr}(T > Ck_0 \mid y_{1:n}) \to 0$ for some constant $C > 1$, hence preventing severe over-estimation in the number of clusters—although whether one could exactly recover k_0 is still unknown. Alternatively, another possibility is to put a hyper-prior on α. Ascolani et al. (2023) show that this method can achieve consistency when the component distribution $\mathcal{F}$ has bounded support. To our best knowledge, the consistency with general $\mathcal{F}$ remains an open question.

3. Posterior Sampling Algorithm

Since the quasi-Bernoulli mixture model involves a small modification to classic stick-breaking construction, we can use an efficient slice sampling algorithm [Kalli et al. (2011), as the improved version of Walker (2007)] for posterior inference. We use a sequence of decreasing positive constants $\xi_1, \xi_2, \ldots$ that converges to zero. In this article, we choose $\xi_i = 0.5^i$ for $i \geq 1$ as suggested in Kalli et al. (2011). Given c_i (the component assignment), consider a latent uniform $u_i \sim \mathrm{Uniform}(0, \xi_{c_i})$, then we have a joint likelihood proportional to $\prod_{i=1}^n \mathbb{1}(u_i < \xi_{c_i}) w_{c_i}/\xi_{c_i} f_{\theta_{c_i}}(y_i)$. We define the state of the Markov chain to be (c, θ, w, u) and the target distribution is the posterior $p(c, \theta, w, u \mid y)$, where $y = y_{1:n}$, $c = c_{1:n}$, $\theta = \theta_{1:\infty}$, and $w = w_{1:\infty}$.

The slice sampler iterates the following steps:

1. *Sample c from its full conditional.* For $i = 1, \ldots, n$: sample $c_i \sim \mathrm{Categorical}(\tilde{w})$ where
$$\tilde{w}_k = \frac{w_k/\xi_k f_{\theta_k}(y_i)}{\sum_{\{l:\xi_l > u_i\}} w_l/\xi_l f_{\theta_l}(y_i)}, \quad \text{for } k \in \{l : \xi_l > u_i\}.$$
Since the sequence $\xi_1, \xi_2, \ldots$ converges to zero, the index set $\{l : \xi_l > u_i\}$ is finite. Compute $n_k := \sum_i \mathbb{1}(c_i = k)$, and $m_k := \sum_i \mathbb{1}(c_i > k)$.

2. *Sample u from its full conditional.* For $i = 1, \ldots, n$: sample u_i from the uniform distribution over the interval $(0, \xi_{c_i})$.

3. *Sample w from its full conditional.* For $k \in \cup_{i=1}^n \{l : \xi_l > u_i\}$:
   - Sample $b_k \sim q\delta_1(\cdot) + (1-q)\delta_\epsilon(\cdot)$ where
   $$q = \frac{\tilde{p}}{\tilde{p} + (1-\tilde{p})\epsilon^{-\alpha} I_\epsilon(m_k + \alpha, n_k + 1)}.$$
   - Sample β_k by drawing $X \sim \mathrm{Beta}_{(0,b_k)}(m_k + \alpha, n_k + 1)$ and setting $\beta_k = X/b_k$, where $\mathrm{Beta}_{(0,\epsilon)}$ denotes a Beta distribution truncated to the interval $(0, \epsilon)$.
   - Compute w_k from $b_{1:k}$ and $\beta_{1:k}$ using Equation (2).

4. *Sample θ from its full conditional.* For $k \in \cup_{i=1}^{n}\{l : \xi_l > u_i\}$: sample θ_k from the distribution proportional to $g(\theta_k)\prod_{i:c_i=k} f_{\theta_k}(y_i)$, where g is the density of the base measure $\mathcal{G}$.

Here, f_θ denotes the density of the component distribution $\mathcal{F}(\theta)$. Note that in the w update, we first sample b_k marginalized over β_k, then sample $\beta_k \mid b_k$, and compute the resulting value of w_k.

4. Simulations

In this section, we assess the empirical performance of the quasi-Bernoulli (QB) mixture model in simulation studies. We compare our method with three popular alternatives: Dirichlet process (DP) mixture, Pitman–Yor process (PY) mixture and finite mixture with a prior on the number of components (MFM).

We demonstrate the consistency of the quasi-Bernoulli mixture model for the number of clusters T when the family of component distributions is correctly specified. We set $\tilde{p} = 0.9$ for the quasi-Bernoulli probability in Equation (2), yielding a prior mean of no more than $1/(1-\tilde{p}) = 10$ components with mixture weights larger than ϵ. The reason is that the random variable K for the first time $b_k = \epsilon$ follows Geometric$(1-\tilde{p})$, and the weights after K are less than ϵ. Alternatively, one could place a Beta hyper-prior on $\tilde{p}$ to make the prior even more weakly informative. We set $\alpha = 1$ and $\epsilon = 1/n^{2.1}$, which satisfies the theoretical condition that $\epsilon = o(1/n^2)$ in Theorem 5. To have a fair comparison, we set the Dirichlet process concentration parameter α so that the expected number of clusters under the Dirichlet process prior is as close as possible to the one under the quasi-Bernoulli process prior for each n, as shown in Table 1 in the appendix. Similarly, for the Pitman–Yor process prior PY(α, d) where each v_k follows Beta$(1-d, \alpha+kd)$, we choose α and d to match both the expectation and the variance of the number of clusters with the quasi-Bernoulli process prior. For the MFM model, we follow Miller and Harrison (2018) and set $\Pi_K(k) = p^{k-1}(1-p)$ where $p = 0.9$ for the prior on number of components, and $(w_1, \ldots, w_K) \sim \text{Dir}_K(\alpha, \ldots, \alpha)$ where $\alpha = 1$ for the prior on mixture weights.

For each experiment, we run the Markov chain for 50,000 iterations, discard the first 20,000 as burn-ins, and use thinning by keeping only every 50th iteration.

4.1 Simulations with Gaussian Mixtures

To compare performance in terms of consistency and MCMC mixing, we first consider simulations using Gaussian distributions $\text{N}(\mu, \Sigma)$ for the mixture components.

We first generate data with sample sizes $n \in \{50, 100, 250, 1000, 2500\}$ from a three-component univariate Gaussian mixture distribution: $0.3\,\text{N}(-4, 1^2)+0.3\,\text{N}(0, 1^2)+0.4\,\text{N}(5, 1^2)$. Following Richardson and Green (1997), we use a data-dependent prior (that is, base measure $\mathcal{G}$) on the component parameters (μ, Σ): $\mu \sim \text{N}(m_\mu, s_\mu^2)$ and $\Sigma \sim \text{Gamma}^{-1}(2, \gamma)$ where $\text{Gamma}^{-1}(a, b)$ has density $f(x) \propto x^{-a-1}\exp(-b/x)$, with a hyper-prior $\text{Gamma}(g, h)$ on γ, where $m_\mu = (\max\{y_{1:n}\} + \min\{y_{1:n}\})/2$, $s_\mu = \max\{y_{1:n}\} - \min\{y_{1:n}\}$, $g = 0.2$, and $h = 10/s_\mu^2$.

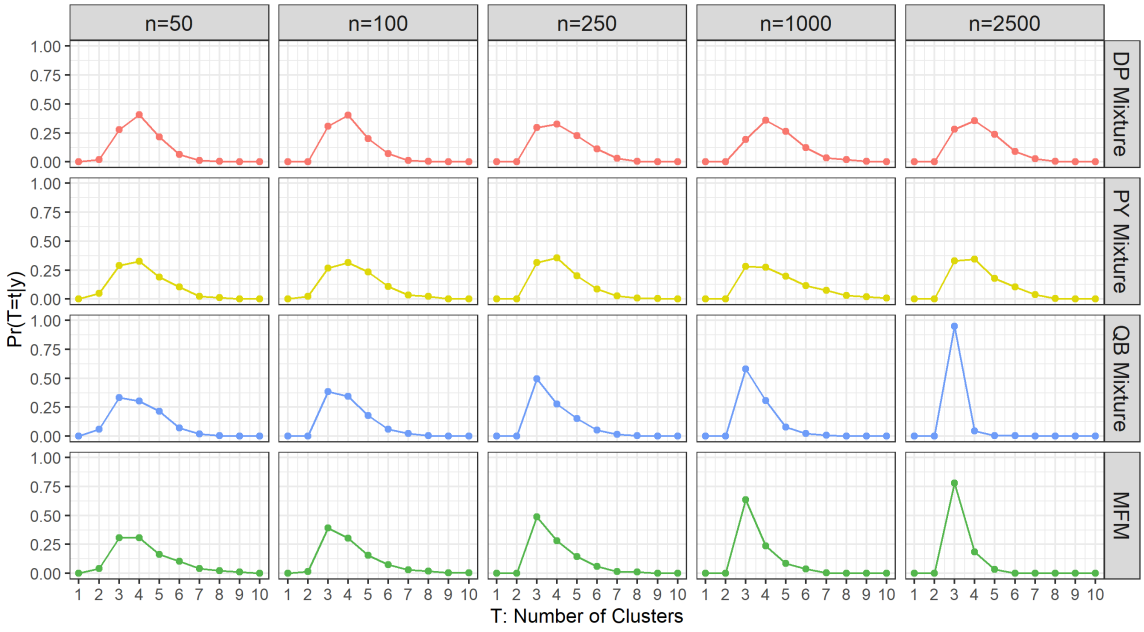
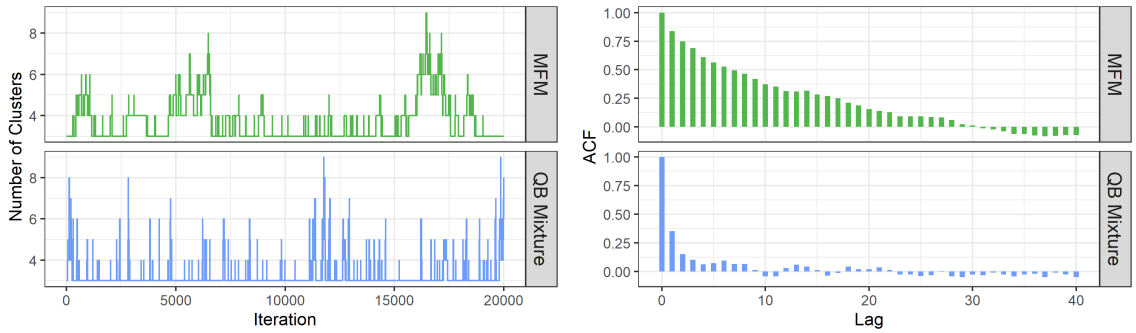


Figure 2: Posterior distribution of the number of clusters (T) for data from a three-component univariate Gaussian mixture. The quasi-Bernoulli mixture model correctly concentrates on three clusters, and its posterior distribution of T concentrates to a point mass at $k_0 = 3$. Coherent with our theory, at large n, the posterior distributions become almost identical to the ones using the MFM model. However, the posterior distributions of calibrated DP mixture model and PY mixture model do not concentrate to a point mass at $k_0 = 3$.

Figure 2 plots the posterior distribution of the number of clusters T at each n. Under the quasi-Bernoulli mixture model (shown in blue), the posterior of T concentrates to a point mass at the true number of components ($k_0 = 3$) as n grows, in accordance with our theory. Further, for both small and large n, the posterior mode of T coincides with the true number of components. Clearly, our model yields almost the same results (blue) as the MFM model (green), especially at large n. This is coherent with our theory (Equation (7)). On the other hand, the Dirichlet process mixture model and the Pitman–Yor process mixture model fail to concentrate on the true number of components.

Despite similar performances in achieving consistency, a major strength of our model is its computational efficiency gained via the slice sampling (Kalli et al., 2011). In comparison, the existing MFM model requires a combinatorial search via the split-merge sampler (Jain and Neal, 2007), which suffers from slow mixing with high auto-correlation. As shown in Figure 3, with thinning, quasi-Bernoulli mixture model using slice sampler shows a drop in the auto-correlation (effective sample size 16.0%, on average of five experiments with sample size 1000), while MFM model using the split-merge sampler shows a much slower drop (effective sample size 7.8%).



(a) Trace on number of clusters (b) Auto-correlation (based on thinning at 50)

Figure 3: The trace of the Markov chain on T and auto-correlation functions for univariate Gaussian mixture data with sample size 1000. Quasi-Bernoulli mixture model using the slice sampler shows much better mixing in the Markov chain, compared to the MFM model using the split-merge sampler. We discard the first 5,000 iterations as burn-ins and record the following 20,000 samples. The slice sampling is not available to the MFM model, since the change to K needs to satisfy the constraint that a non-empty cluster should not have a zero mixture weight.

Next, we consider a multivariate simulation scenario in which we generate data sets of size $n \in \{250, 1000, 2500\}$ from a three-component bivariate Gaussian mixture: $0.3\,\mathrm{N}((-4\ 1)^{\mathrm{T}}, I_2) + 0.3\,\mathrm{N}((0\ \ 2)^{\mathrm{T}}, I_2) + 0.4\,\mathrm{N}((5\ \ 3)^{\mathrm{T}}, I_2)$. We use the data-dependent prior $\mu \sim \mathrm{N}(m, C)$, $\Sigma \sim \mathrm{Wishart}_2^{-1}(C^{-1}/2, 2)$ on the component parameters, where m is the sample mean and C is the sample covariance. The results are similar to the univariate simulation scenario; see Appendix B.1.

4.2 Simulations with Non-Gaussian Mixtures

The quasi-Bernoulli mixture model can easily be extended to mixture models with non-Gaussian components. To illustrate that the consistency result still holds, we consider data generated from a mixture of Laplace distributions.

We generate data from a three-component Laplace mixture: $0.35\mathrm{Lap}(-10, 1) + 0.3\mathrm{Lap}(0, 1.5) + 0.35\mathrm{Lap}(10, 0.5)$, where $\mathrm{Lap}(\mu, \lambda)$ denotes a Laplace distribution with mean μ and scale λ. We use a data-dependent prior (base measure $\mathcal{G}$) on (μ, λ): $\mu \sim \mathrm{N}(m_\mu, \sigma_\mu^2)$ and $\lambda \sim \mathrm{Gamma}^{-1}(2, 1)$, where $m_\mu = (\max\{y_{1:n}\} + \min\{y_{1:n}\})/2$ and $\sigma_\mu = \max\{y_{1:n}\} - \min\{y_{1:n}\}$. Figure 4 shows that the quasi-Bernoulli process successfully recovers the true number of components, while the Dirichlet process (red) and the Pitman–Yor process (yellow) fail to do so. Under a mixture distribution like the Laplace distribution having heavier tails than the Gaussian distribution, the Dirichlet process and the Pitman–Yor process tend to overestimate the number of clusters to a greater extent.

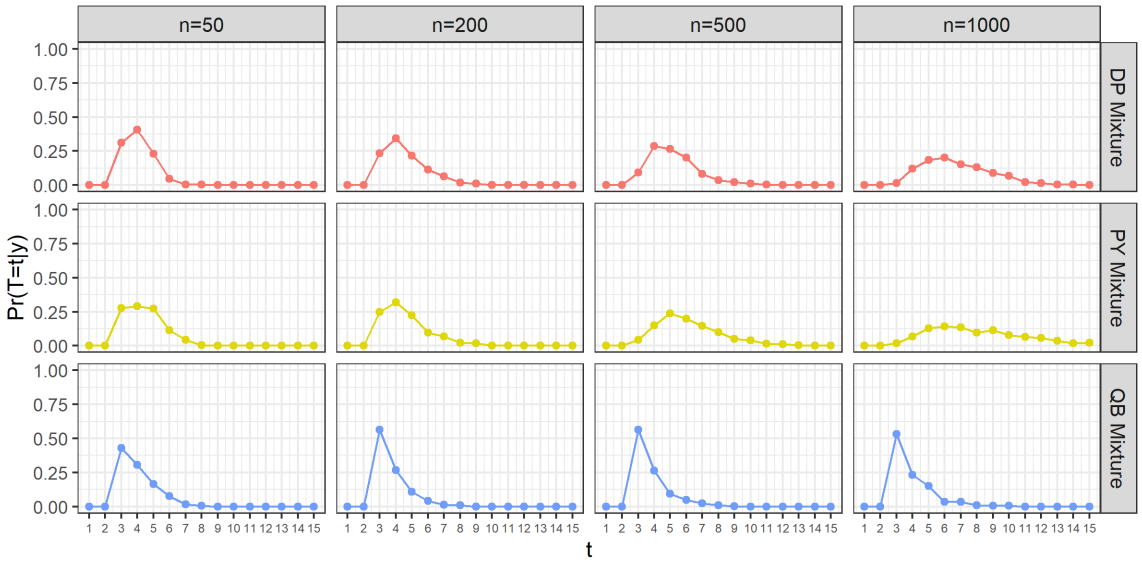


Figure 4: Quasi-Bernoulli mixture model correctly recovers three clusters as the ground truth when each component is from a Laplace distribution. The Dirichlet process and Pitman–Yor process overestimate the number of clusters, due to having small spurious clusters.

5. Data Application: Clustering Brain Networks

To demonstrate the ease of using our model in an advanced data analysis, we apply it to cluster multiple brain networks, collected from $n = 812$ subjects in the human connectome project (Marcus et al., 2011).

For each subject, resting-state functional magnetic resonance imaging (fMRI) signals were collected from $R = 50$ regions of the brain, indexed by $r = 1, \ldots, R$. The data were processed and transformed into a connectivity graph on R vertices, represented as a symmetric binary adjacency matrix $Y^{(i)} \in \{0,1\}^{R \times R}$ such that $Y^{(i)} = Y^{(i)\mathrm{T}}$.

We model these adjacency matrices using a probit-Bernoulli mixture model in which each component distribution has a low-rank latent structure. The goal of this model is to cluster the networks into disjoint groups of similar subjects, and to find a meaningful low-dimensional representation of networks within each sub-group/cluster.

Given a matrix of probabilities $\theta \in [0,1]^{R \times R}$, we write $Y \sim \text{Bernoulli}(\theta)$ to denote that $Y_{rs} \sim \text{Bernoulli}(\theta_{rs})$ independently for $r, s \in \{1, \ldots, R\}$. In this notation, we use the following infinite mixture model with a quasi-Bernoulli stick-breaking prior on the mixture weights $w = (w_1, w_2, \ldots)$,

$$\begin{aligned} Y^{(i)} \mid c_i, \mu, M &\overset{indep}{\sim} \text{Bernoulli}(\Phi(\mu + M_{c_i})), \\ c_i \mid w &\overset{iid}{\sim} \text{Categorical}(w) \text{ for } i = 1, \ldots, n, \\ M_k &= Q_k \Lambda_k Q_k^{\mathrm{T}}, \quad \text{for } k \geq 1, \\ w &\sim \text{Quasi-Bernoulli}(\tilde{p}, \epsilon, \alpha), \end{aligned} \tag{8}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian distribution, applied element-wise, and the weights $w = (w_1, w_2, \ldots)$ are drawn from Equation (2) which is denoted by Quasi-Bernoulli$(\tilde{p}, \epsilon, \alpha)$. The model enforces symmetry for the binary matrices $Y^{(i)}$ by only modeling the lower triangle part of them. Here, μ is a scalar that is shared by all components, and each $M_k = Q_k \Lambda_k Q_k^{\mathrm{T}}$ is a component-specific matrix such that $\Lambda_k = \mathrm{diag}(\lambda_{k,1}, \ldots, \lambda_{k,d})$, and Q_k belongs to the Stiefel manifold $\mathcal{V}^{d,R} := \{Q \in \mathbb{R}^{R \times d} : Q^{\mathrm{T}} Q = I_d\}$ (Hoff, 2009). The role of M_k is to provide a low-rank representation for mixture component k, and the μ is a nuisance parameter that captures departures from this assumed low-rank structure.

For the other priors, we assign $Q_k \sim \mathrm{Uniform}(\mathcal{V}^{d,R})$ for $k = 1, 2, \ldots$ with $d = 2$, use a truncated Gaussian prior $\pi(\lambda_{k,l}) \propto \mathrm{N}(\lambda_{k,l} \mid 0, 50)$ for $l = 1, \ldots, d$, and assign a Gaussian prior $\mathrm{N}(0, 10^2)$ on μ, following Hoff (2009). For the quasi-Bernoulli prior on w, we use $\tilde{p} = 0.9$, $\alpha = 1$ and $\epsilon = 1/n^{2.1}$. We run the MCMC sampler from Section 3 for $30,000$ iterations and discard the first $10,000$ as burn-ins.

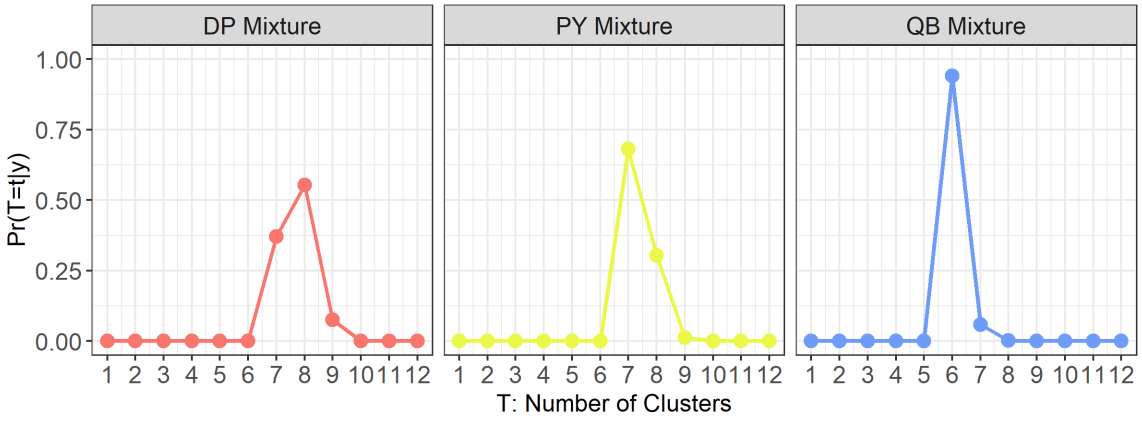
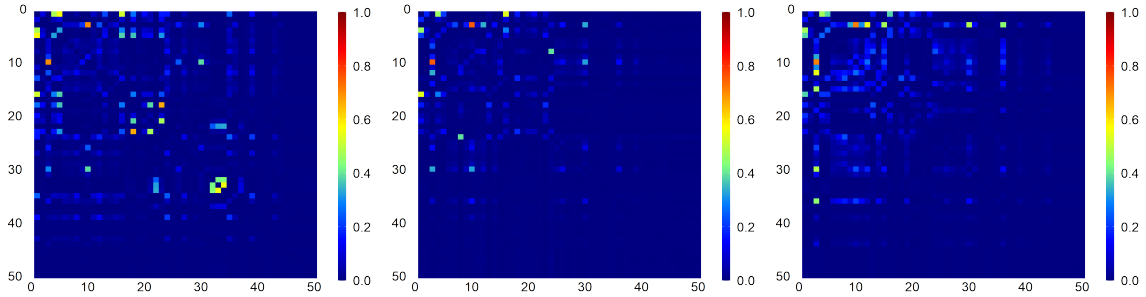


Figure 5: The quasi-Bernoulli model concentrates on $T = 6$ clusters on the brain connectivity data. For comparison, the posterior mode of the corresponding Dirichlet process and Pitman–Yor process mixture models are at $T = 8$ and $T = 7$.

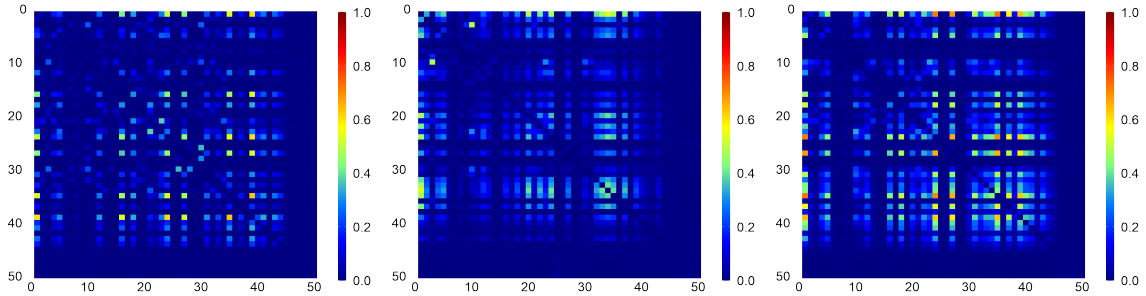
Figure 5 (right) shows the posterior of the number of clusters T for the quasi-Bernoulli mixture model, which is highly concentrated on $T = 6$ clusters. For comparison, we also consider a Dirichlet process mixture and a Pitman–Yor process mixture. We use the same prior for the component parameters Q_k, Λ_k, as the one in the quasi-Bernoulli model. We set the Dirichlet process concentration parameter α so that the expected number of clusters under the Dirichlet process prior is as close as possible to the one under the quasi-Bernoulli process prior for $n = 812$. Similarly, for the Pitman–Yor process prior $\mathrm{PY}(\alpha, d)$, we choose α and d to match both the expectation and the variance of the number of clusters with the quasi-Bernoulli process prior. The α for the Dirichlet process is chosen to be 0.63, and the parameters of the Pitman–Yor process are $\alpha = 0.30$, $d = 0.11$. As the results, these two models yield posterior modes of $T = 8$ and $T = 7$ clusters (Figure 5, left and middle) and produce several very small clusters (the Dirichlet process mixture model produces four small groups with 5.8%, 5.5%, 5.2% and 0.1% of subjects; the Pitman–Yor process mixture model produces three small groups with 3.3%, 2.5% and 1.7% of subjects—these proportions are



(a) Group 1 (46.0% of the subjects)

(b) Group 2 (18.9% of the subjects)

(c) Group 3 (12.9% of the subjects)



(d) Group 4 (9.0% of the subjects)

(e) Group 5 (7.9% of the subjects)

(f) Group 6 (5.2% of the subjects)

Figure 6: The posterior means of the edge connectivity probabilities $\Phi(\mu + M_k)$ over the six groups.

calculated as the average cluster sizes divided by n, with average taken over those posterior samples with T equal to the posterior mode of T). The result from the quasi-Bernoulli model leads to a more parsimonious representation.

In the quasi-Bernoulli posterior, the subjects are clustered into six groups with high probability. Figure 6 shows the posterior means of the edge probabilities $\Phi(\mu + M_k)$ for each group. These results indicate that sparse connectivity is exhibited by the first three groups of subjects (accounting for 77.8% of subjects), whereas the other three groups have denser connectivity. Specifically, for each group, we examine the posterior mean proportion of node pairs with edge connectivity probabilities greater than 0.05. The proportions for the six groups are 8.6%, 4.4%, 8.6%, 15.7%, 21.1% and 28.2%, respectively.

In addition, we conduct additional experiments using two common Bayesian clustering methods: (i) Dirichlet process mixture of high-dimensional probit-Bernoulli model (without low-rank latent structure); (ii) approximation using the mixture of factor analyzers (McLachlan et al., 2003) and treating the data as if they were continuous, and selecting the number of clusters using Bayesian information criterion (BIC). The method (i) is a popular solution, however, it is found to suffer from a curse of dimensionality (Chandra et al., 2022). Indeed, applying (i) on the data leads to only $T = 1$ cluster in the result. The method (ii) selects 3 clusters and 2 latent dimensions under BIC. The three clusters of the maximum a

posteriori (MAP) estimates have 36.6%, 20.6% and 42.8% of the subjects. In the appendix, Figure 10 shows the MAP estimate of the mean of each Gaussian component.

6. Discussion

In this article, we propose a modification to the canonical stick-breaking construction, leading to an infinite mixture model that provides consistency for the number of clusters (like the MFM model) as well as easy implementation in posterior MCMC computation (like the Dirichlet process mixture model). Heiner et al. (2019) similarly proposes a tweak to the breaking proportion v_k under the framework of the Bayesian finite mixture prior, while we focus on the infinite mixture model and construct the theoretical properties on the number of clusters.

There are several extensions worth further pursuing. First, recovery of the true number of clusters under a *misspecified* model is still an open problem. In a recent work (Cai et al., 2021), it is theoretically shown that even a small amount of misspecification of the components will tend to result in over-estimation of the number of clusters in overfitted finite mixture models. Intuitively, this suggests that in addition to controlling the mixture weights to avoid small spurious clusters, it is also important to ensure that the family of component distributions is flexible enough to avoid severe model misspecification issues. Second, it would be interesting to investigate whether the combination of the quasi-Bernoulli infinite mixture framework and distance clustering approaches, such as the Laplacian-based approach (Rohe et al., 2011), can lead to a consistency result for the number of clusters.

The popular mixture models (such as Dirichlet process and Pitman–Yor process mixtures) are completely fine for the task of density estimation; nevertheless in some sense, a non-parametric mixing measure entails some misspecification under an indefinitely increasing n, which inherently assumes the number of clusters growing with n, hence conflicting with the popular modeling view where there is a fixed ground-truth k_0. Our insight is that for any fixed n, we have an exchangeable partition probability function (which is calculated by integrating out the mixture weights of those non-occupied components) that gives a discrete distribution for $K \in \{1, \dots, n\}$. Therefore, we can calibrate the asymptotic behavior of the probability function to produce a consistent estimator, via either controlling α in the Dirichlet process mixture or ϵ as we do in our quasi-Bernoulli model.

Appendix A. Proofs

Proof of Theorem 1
The conditional probability mass function of the assignment variables $c = (c_1, \dots, c_n)$ is

$$p(c \mid b_1, b_2, \dots, \beta_1, \beta_2, \dots) = \prod_{k=1}^{\infty} (1 - \beta_k b_k)^{n_k} (\beta_k b_k)^{m_k},$$

where $n_k = \sum_{i=1}^n \mathbb{1}(c_i = k)$ and $m_k = \sum_{i=1}^n \mathbb{1}(c_i > k)$. Define $M(c) := \max\{c_1, \ldots, c_n\}$ and $Q_k := \tilde{p} + (1-\tilde{p}) I_\epsilon(m_k + \alpha, n_k + 1)/\epsilon^\alpha$. Then

$$\begin{aligned}
p(c) &= \prod_{k=1}^{\infty} \int_0^1 \Big(\tilde{p}(1-\beta_k)^{n_k}(\beta_k)^{m_k} + (1-\tilde{p})(1-\epsilon\beta_k)^{n_k}(\epsilon\beta_k)^{m_k}\Big)\alpha\beta_k^{\alpha-1}\, d\beta_k \\
&= \prod_{k=1}^{\infty} \Big(\tilde{p}\alpha \mathrm{B}(n_k+1, m_k+\alpha) + (1-\tilde{p})\frac{\alpha}{\epsilon^\alpha}\mathrm{B}(n_k+1, m_k+\alpha)\int_0^\epsilon \frac{(1-x)^{n_k}(x)^{m_k+\alpha-1}}{\mathrm{B}(n_k+1, m_k+\alpha)}\, dx\Big) \\
&\overset{\text{(a)}}{=} \prod_{k=1}^{M(c)} \Big(\tilde{p}\alpha \mathrm{B}(n_k+1, m_k+\alpha) + (1-\tilde{p})\frac{\alpha}{\epsilon^\alpha}\mathrm{B}(n_k+1, m_k+\alpha) I_\epsilon(m_k+\alpha, n_k+1)\Big) \\
&= \prod_{k=1}^{M(c)} \frac{\alpha\Gamma(n_k+1)\Gamma(m_k+\alpha)}{\Gamma(n_k+m_k+\alpha+1)} Q_k \\
&\overset{\text{(b)}}{=} \Big(\prod_{k=1}^{M(c)} \Gamma(n_k+1)\Big) \prod_{k=1}^{M(c)} \frac{\alpha Q_k \Gamma(m_k+\alpha)}{(m_{k-1}+\alpha)\Gamma(m_{k-1}+\alpha)} \\
&= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\Big(\prod_{k=1}^{M(c)} \Gamma(n_k+1)\Big) \prod_{k=1}^{M(c)} \frac{\alpha Q_k}{m_{k-1}+\alpha}.
\end{aligned}$$

In step (a), we use the fact that for all $k > M(c)$, we have $n_k = 0$ and $m_k = 0$, and thus, everything cancels for such k since $B(1, \alpha) = 1/\alpha$ and $I_\epsilon(\alpha, 1) = \epsilon^\alpha$. In step (b), we use the fact that $n_k + m_k = m_{k-1}$, and thus, $\Gamma(n_k + m_k + \alpha + 1) = (m_{k-1} + \alpha)\Gamma(m_{k-1} + \alpha)$.

Define $g_k := \sum_{i=1}^n \mathbb{1}(c_i \geq k) = \sum_{l=k}^\infty n_l$, and note that $g_k = m_{k-1}$. Let $\mathcal{A}_c$ be the partition of $\{1, \ldots, n\}$ induced by c. Fix a partition $\mathcal{A} = \{A_1, \ldots, A_t\}$. When $\mathcal{A}_c = \mathcal{A}$, there are exactly t unique values among $c_1, \ldots, c_n$. Let $k_1 < k_2 < \cdots < k_t$ denote these unique values, and set $k_0 = 0$. For $k_{j-1} < k < k_j$, we have $n_k = 0$ and $g_k = g_{k+1} = g_{k_j}$, and thus $Q_k = \tilde{p} + (1-\tilde{p})\epsilon^{g_{k_j}}$. Meanwhile, for $k = k_j$, we have $n_k = n_{k_j}$ and $g_{k+1} = g_{k_{j+1}}$, and thus it follows that $Q_k = \tilde{p} + (1-\tilde{p}) I_\epsilon(g_{k_{j+1}} + \alpha, n_{k_j} + 1)/\epsilon^\alpha$. Hence, for all c such that $\mathcal{A}_c = \mathcal{A}$, we have

$$p(c) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\Big(\prod_{j=1}^{t} \Gamma(n_{k_j}+1)\Big) \prod_{j=1}^{t} U_j(c)$$

where

$$U_j(c) = \left(\frac{\alpha\tilde{p} + \alpha(1-\tilde{p}) I_\epsilon(g_{k_{j+1}} + \alpha, n_{k_j} + 1)/\epsilon^\alpha}{g_{k_j} + \alpha}\right)\left(\frac{\alpha\tilde{p} + \alpha(1-\tilde{p})\epsilon^{g_{k_j}}}{g_{k_j} + \alpha}\right)^{d_j}$$

where $d_j = k_j - k_{j-1} - 1$. Since there is a unique permutation $\sigma = (\sigma_1, \ldots, \sigma_t)$ of $\{1, \ldots, t\}$ such that $A_{\sigma_j} = \{i : c_i = k_j\}$ for all $j \in \{1, \ldots, t\}$, the mapping between $\{c : \mathcal{A}_c = \mathcal{A}\}$ and $\{(\sigma, d_1, \ldots, d_t) : \sigma \in S_t,\ d_1, \ldots, d_t \in \mathbb{N}\}$ is a bijection. (Here, S_t is the set of all permutations of $\{1, \ldots, t\}$, and $\mathbb{N} := \{0, 1, 2, \ldots\}$.) Let $n_j^* := |A_j|$ and $g_j^*(\sigma) := \sum_{l=j}^t n_{\sigma_l}^*$. For the value of $(\sigma, d_{1:t})$ that corresponds to c, we have $g_{k_j} = g_j^*(\sigma)$ and $n_{k_j} = |A_{\sigma_j}| = n_{\sigma_j}^*$, and thus, $U_j(c) = U_j^*(\sigma, d_{1:t})$ where

$$U_j^*(\sigma, d_{1:t}) := \left(\frac{\alpha\tilde{p} + \alpha(1-\tilde{p}) I_\epsilon(g_{j+1}^*(\sigma) + \alpha, n_{\sigma_j}^* + 1)/\epsilon^\alpha}{g_j^*(\sigma) + \alpha}\right)\left(\frac{\alpha\tilde{p} + \alpha(1-\tilde{p})\epsilon^{g_j^*(\sigma)}}{g_j^*(\sigma) + \alpha}\right)^{d_j}.$$

Summing over d_j and using $\sum_{d=0}^{\infty} x^d = 1/(1-x)$ for $|x| < 1$, we have

$$\sum_{d_j=0}^{\infty} U_j^*(\sigma, d_{1:t}) = \frac{\alpha\tilde{p} + \alpha(1-\tilde{p})I_\epsilon(g_{j+1}^*(\sigma)+\alpha, n_{\sigma_j}^*+1)/\epsilon^\alpha}{g_j^*(\sigma)+\alpha(1-\tilde{p})(1-\epsilon^{g_j^*(\sigma)})}.$$

Note that $U_j^*(\sigma, d_{1:t})$ depends on $d_{1:t}$ only through d_j. Therefore,

$$\begin{aligned}
p_{\epsilon,n}(\mathcal{A}) &= \sum_{c\,:\,\mathcal{A}_c=\mathcal{A}} p(c) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\Big(\prod_{j=1}^{t}\Gamma(n_j^*+1)\Big)\sum_{\sigma\in S_t}\sum_{d_1=0}^{\infty}\cdots\sum_{d_t=0}^{\infty}\prod_{j=1}^{t}U_j^*(\sigma,d_{1:t}) \\
&= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\Big(\prod_{j=1}^{t}\Gamma(n_j^*+1)\Big)\sum_{\text{all }\sigma}\prod_{j=1}^{t}\sum_{d_j=0}^{\infty}U_j^*(\sigma,d_{1:t}) \\
&= \frac{\alpha^t\Gamma(\alpha)}{\Gamma(n+\alpha)}\Big(\prod_{j=1}^{t}\Gamma(n_j^*+1)\Big)\sum_{\text{all }\sigma}\prod_{j=1}^{t}\frac{\tilde{p}+(1-\tilde{p})I_\epsilon(g_{j+1}^*(\sigma)+\alpha, n_{\sigma_j}^*+1)/\epsilon^\alpha}{g_j^*(\sigma)+\alpha(1-\tilde{p})(1-\epsilon^{g_j^*(\sigma)})}.
\end{aligned}$$

This proves the result. ■

Proof of Lemma 2

The general approach of the proof follows the technique of Miller (2019). The conditional probability mass function of the assignment variables is

$$p(c \mid b_1, b_2, \ldots, \beta_1, \beta_2, \ldots) = \prod_{k=1}^{\infty}(1-\beta_k b_k)^{n_k}(\beta_k b_k)^{g_{k+1}},$$

where $n_k = \sum_{i=1}^{n}\mathbb{1}(c_i = k)$ and $g_k = \sum_{i=1}^{n}\mathbb{1}(c_i \geq k)$. Let $M(c) = \max\{c_1, \ldots, c_n\}$. Then

$$\begin{aligned}
p(c) &= \prod_{k=1}^{\infty}\int_0^1 \Big(\tilde{p}(1-\beta_k)^{n_k}(\beta_k)^{g_{k+1}} + (1-\tilde{p})\mathbb{1}(g_{k+1}=0)\Big)\alpha\beta_k^{\alpha-1}\,d\beta_k \\
&= \prod_{k=1}^{M(c)}\Big(\tilde{p}\alpha\mathrm{B}(n_k+1, g_{k+1}+\alpha) + (1-\tilde{p})\mathbb{1}(g_{k+1}=0)\Big) \\
&= \Bigg(\prod_{k=1}^{M(c)-1}\frac{\tilde{p}\alpha\Gamma(n_k+1)\Gamma(g_{k+1}+\alpha)}{\Gamma(n_k+g_{k+1}+\alpha+1)}\Bigg)\big(\tilde{p}\alpha\mathrm{B}(n_{M(c)}+1,\alpha)+1-\tilde{p}\big) \\
&= \Bigg(\prod_{k=1}^{M(c)-1}\Gamma(n_k+1)\Bigg)\Bigg(\prod_{k=1}^{M(c)-1}\frac{\tilde{p}\alpha\Gamma(g_{k+1}+\alpha)}{\Gamma(g_k+\alpha+1)}\Bigg)\big(\tilde{p}\alpha\mathrm{B}(n_{M(c)}+1,\alpha)+1-\tilde{p}\big) \\
&= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\Big(\prod_{k=1}^{M(c)}\Gamma(n_k+1)\Big)\Bigg(\prod_{k=1}^{M(c)-1}\frac{\tilde{p}\alpha}{g_k+\alpha}\Bigg)\Bigg(\frac{\tilde{p}\alpha+(1-\tilde{p})/\mathrm{B}(n_{M(c)}+1,\alpha)}{n_{M(c)}+\alpha}\Bigg).
\end{aligned}$$

Let $\mathcal{A}(c)$ denote the partition of $\{1, \ldots, n\}$ corresponding to c. For fixed $\mathcal{A} = \{A_1, \ldots, A_t\}$, when $\mathcal{A}(c) = \mathcal{A}$, there are exactly t unique values among $c_1, \ldots, c_n$. Let $k_1 < k_2 < \cdots < k_t$

denote these unique values, and set $k_0 = 0$. For $k_{j-1} < k \le k_j$, we have $g_k = g_{k_j}$. Hence, for c satisfying $\mathcal{A}(c) = \mathcal{A}$, we have $p(c) = (\Gamma(\alpha)/\Gamma(n+\alpha))\big(\prod_{j=1}^t \Gamma(n_{k_j}+1)\big)\prod_{j=1}^t U_j(c)$, where $U_j(c) = (\tilde{p}\alpha/(g_{k_j}+\alpha))^{d_j}$ for $1 \le j < t$ and

$$U_t(c) = \frac{\tilde{p}\alpha + (1-\tilde{p})/\mathrm{B}(n_{k_t}+1,\alpha)}{n_{k_t}+\alpha}\left(\frac{\tilde{p}\alpha}{g_{k_t}+\alpha}\right)^{d_t-1} = \left(1 + \frac{(1-\tilde{p})/\tilde{p}\alpha}{\mathrm{B}(n_{k_t}+1,\alpha)}\right)\left(\frac{\tilde{p}\alpha}{g_{k_t}+1}\right)^{d_t}$$

where $d_j = k_j - k_{j-1}$ for $j = 1, \ldots, t$.

Since there is a unique permutation $\sigma = (\sigma_1, \ldots, \sigma_t) \in S_t$ such that $A_{\sigma_j} = \{i : c_i = k_j\}$, the mapping between $\{c : \mathcal{A}(c) = \mathcal{A}\}$ and $\big\{(\sigma, d_1, \ldots, d_t) : \sigma \in S_t,\ d_1, \ldots, d_t \in \{1, 2, \ldots\}\big\}$ is a bijection. Letting $n_j = |A_j|$, we have

$$p_{0,n}(\mathcal{A}) = \sum_{c\,:\,\mathcal{A}(c)=\mathcal{A}} p(c) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}\left(\prod_{j=1}^t \Gamma(n_j+1)\right)\sum_{\sigma\in S_t}\sum_{d_1=1}^{\infty}\cdots\sum_{d_t=1}^{\infty}\prod_{j=1}^t U_j(c), \tag{9}$$

treating c as a function of $\sigma, d_1, \ldots, d_t$. Changing the order of summations and multiplication, defining $g_j(\sigma) = \sum_{l=j}^t n_{\sigma_l}$, and using the geometric series $\sum_{d=1}^{\infty} x^d = x/(1-x)$ for $x \in (0,1)$,

$$\begin{aligned}
\sum_{d_1=1}^{\infty}\cdots\sum_{d_t=1}^{\infty}\prod_{j=1}^t U_j(c) &= \left(1 + \frac{(1-\tilde{p})/\tilde{p}\alpha}{\mathrm{B}(n_{\sigma_t}+1,\alpha)}\right)\prod_{j=1}^t\sum_{d_j=1}^{\infty}\left(\frac{\tilde{p}\alpha}{g_j(\sigma)+\alpha}\right)^{d_j} \\
&= \left(1 + \frac{(1-\tilde{p})/\tilde{p}\alpha}{\mathrm{B}(n_{\sigma_t}+1,\alpha)}\right)\prod_{j=1}^t \frac{\tilde{p}\alpha}{g_j(\sigma)+\alpha(1-\tilde{p})} \\
&= \alpha^t\prod_{j=1}^t \frac{\tilde{p} + \mathbb{1}(j=t)(1-\tilde{p})/(\alpha\mathrm{B}(n_{\sigma_t}+1,\alpha))}{g_j(\sigma)+\alpha(1-\tilde{p})}.
\end{aligned}$$

Combining with Equation (9), this proves the result. ■

We provide the complete statements of the two results from Nobile (1994).

Theorem 7 *Nobile (1994, Corollary 3.1) Assume $\phi \in \Omega'$ is identifiable up to permutation of the mixture components. Let Π_0 be the prior on Ω under the model defined by Equations (3) and (7), and assume $\Pi_0(\{\phi : \exists\ i \ne j$ such that $\theta_i = \theta_j\}) = 0$. Let Π_0' be the corresponding prior on Ω' induced by η. Then there is a subset $\Omega_0' \subset \Omega'$ with $\Pi_0'(\Omega_0') = 1$ such that for any $\phi_0 = (k_0, w_1^0, \ldots, w_{k_0}^0, \theta_1^0, \ldots, \theta_{k_0}^0) \in \Omega_0'$, if $y_1, y_2, \ldots \mid \phi_0 \overset{iid}{\sim} P_{\phi_0}$, then as $n \to \infty$, we have*

$$p_{\epsilon=0}(\phi \in D \mid y_{1:n}) \to \mathbb{1}(\phi_0 \in D') \quad \text{a.s.}[P_{\phi_0}],$$

for any measurable subset $D' \subset \Omega'$ and $D = \{\phi \in \Omega : \eta(\phi) \in D'\}$.

Theorem 8 *Nobile (1994, Proposition 3.5) Under the same assumptions of Theorem 7,*

$$p_{\epsilon=0}(K = k \mid y_{1:n}) \to \mathbb{1}(k_0 = k) \quad \text{a.s.}[P_{\phi_0}].$$

Proof of Theorem 3

The first result is proved by Theorem 8. The second result can be proved as follows. Using the Theorem 7, there is a subspace Ω_0' as described in the theorem such that if $\phi_0 \in \Omega_0'$ then the posterior distribution of (w, θ) given $K = k_0$ and $y_{1:n}$ will converge to a uniform distribution in which with equal probability the (w, θ) is one of the permutations of $(w_1^0, \theta_1^0), \ldots, (w_{k_0}^0, \theta_{k_0}^0)$. This is because the transformation η maps all of the permutations of $(w_1^0, \theta_1^0), \ldots, (w_{k_0}^0, \theta_{k_0}^0)$ into the same one with a specific order. Define the random variables $N_k = \sum_{i=1}^n \mathbb{1}(c_i = k)$ for $k = 1, \ldots, k_0$. Then

$$
\begin{aligned}
&p_{\epsilon=0}(N_k = 0 \mid K = k_0, y_{1:n}) \\
&= p_{\epsilon=0}(\cap_{i=1}^n \{c_i \neq k\} \mid K = k_0, y_{1:n}) \\
&= \int p_{\epsilon=0}(\cap_{i=1}^n \{c_i \neq k\} \mid K = k_0, w, \theta, y_{1:n}) \mathbb{P}_{\epsilon=0}(dw, d\theta \mid K = k_0, y_{1:n}) \\
&= \int \prod_{i=1}^n \Big(1 - \frac{w_k f_{\theta_k}(y_i)}{\sum_{l=1}^{k_0} w_l f_{\theta_l}(y_i)}\Big) \mathbb{P}_{\epsilon=0}(dw, d\theta \mid K = k_0, y_{1:n}) \\
&\leq \int \prod_{i=1}^{n_0} \Big(1 - \frac{w_k f_{\theta_k}(y_i)}{\sum_{l=1}^{k_0} w_l f_{\theta_l}(y_i)}\Big) \mathbb{P}_{\epsilon=0}(dw, d\theta \mid K = k_0, y_{1:n}) \qquad (10)
\end{aligned}
$$

for any given positive integer n_0 and $n \geq n_0$. Using the weak convergence of the posterior distribution of the (w, θ) and the fact that the integrand is bounded and f_θ is continuous at all θ_k^0's, the Equation (10) converges to

$$
\sum_{k=1}^{k_0} \frac{1}{k_0} \prod_{i=1}^{n_0} \Big(1 - \frac{w_k^0 f_{\theta_k^0}(y_i)}{\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_i)}\Big)
$$

a.s.$[P_{\phi_0}]$ when $n \to \infty$. Since Equation (10) holds for any positive integer n_0, we have

$$
\limsup_{n \to \infty} p_{\epsilon=0}(N_k = 0 \mid K = k_0, y_{1:n}) \leq \sum_{k=1}^{k_0} \frac{1}{k_0} \prod_{i=1}^{\infty} \Big(1 - \frac{w_k^0 f_{\theta_k^0}(y_i)}{\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_i)}\Big).
$$

For every $k = 1, \ldots, k_0$, there exists a measurable set D_k with non-zero measure such that $f_{\theta_k^0}(y) \geq \delta_k > 0$ when $y \in D_k$, and all $f_{\theta_l^0}(y)$ $(l = 1, \ldots, k_0)$ are finite on D_k. For every $k = 1, \ldots, k_0$, there exists a sequence $n_{k1}, n_{k2}, \ldots$ such that $y_{n_{ki}} \in D_k$ $(i = 1, 2, \ldots)$, a.s.$[P_{\phi_0}]$. Hence, for all $i \geq 1$, $f_{\theta_k^0}(y_{n_{ki}}) \geq \delta_k$ and $\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_{n_{ki}}) \leq M$ for some M. Therefore,

$$
\prod_{i=1}^{\infty} \Big(1 - \frac{w_k^0 f_{\theta_k^0}(y_i)}{\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_i)}\Big) \leq \prod_{i=1}^{\infty} \Big(1 - \frac{w_k^0 f_{\theta_k^0}(y_{n_{ki}})}{\sum_{l=1}^{k_0} w_l^0 f_{\theta_l^0}(y_{n_{ki}})}\Big) \leq \prod_{i=1}^{\infty} \Big(1 - \frac{w_k^0 \delta}{M}\Big) = 0,
$$

which leads to $p_{\epsilon=0}(N_k = 0 \mid K = k_0, y_{1:n}) \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. Hence, given $K = k_0$, the posterior probability of having k_0 clusters is

$$
\begin{aligned}
&p_{\epsilon=0}(T = k_0 \mid K = k_0, y_{1:n}) = p_{\epsilon=0}(N_1 > 0, \ldots, N_{k_0} > 0 \mid K = k_0, y_{1:n}) \\
&= 1 - p_{\epsilon=0}(\cup_{k=1}^{k_0} \{N_k = 0\} \mid K = k_0, y_{1:n}) \\
&\geq 1 - \sum_{k=1}^{k_0} p_{\epsilon=0}(N_k = 0 \mid K = k_0, y_{1:n}) \to 1
\end{aligned}
$$

a.s.$[P_{\phi_0}]$ as $n \to \infty$. Therefore,

$$\begin{aligned} p_{\epsilon=0}(T = k_0 \mid y_{1:n}) &= \sum_{k=k_0}^{\infty} p_{\epsilon=0}(T = k_0 \mid K = k, y_{1:n}) p_{\epsilon=0}(K = k \mid y_{1:n}) \\ &\geq p_{\epsilon=0}(T = k_0 \mid K = k_0, y_{1:n}) p_{\epsilon=0}(K = k_0 \mid y_{1:n}) \to 1 \end{aligned}$$

a.s.$[P_{\phi_0}]$ as $n \to \infty$. ■

Proof of Theorem 4

For a given partition $\mathcal{A}$, let $t = |\mathcal{A}|$. Define the following notation to represent the factors in $p_{0,n}(\mathcal{A})$ and $p_{\epsilon,n}(\mathcal{A})$, respectively:

$$\begin{aligned} U_j(\sigma) &:= \frac{\tilde{p} + \mathbb{1}(j = t)(1 - \tilde{p})/(\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1))}{g_j(\sigma) + \alpha(1 - \tilde{p})} \\ V_j(\sigma) &:= \frac{\tilde{p} + (1 - \tilde{p}) I_\epsilon(g_{j+1}(\sigma) + \alpha, n_{\sigma_j} + 1)/\epsilon^\alpha}{g_j(\sigma) + \alpha(1 - \tilde{p})(1 - \epsilon^{g_j(\sigma)})} \end{aligned}$$

for $j = 1, \ldots, t$. When $j < t$, we have

$$U_j(\sigma) \leq V_j(\sigma) \tag{11}$$

since $(1-\tilde{p}) I_\epsilon(g_{j+1}(\sigma)+\alpha, n_{\sigma_j}+1)/\epsilon^\alpha > 0$ and $g_j(\sigma)+\alpha(1-\tilde{p}) > g_j(\sigma)+\alpha(1-\tilde{p})(1-\epsilon^{g_j(\sigma)}) > 0$.

Meanwhile, for the case of $j = t$, we have

$$\begin{aligned} \frac{U_t(\sigma)}{V_t(\sigma)} &\overset{\text{(a)}}{\leq} \frac{\tilde{p} + (1 - \tilde{p})/(\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1))}{\tilde{p} + (1 - \tilde{p}) I_\epsilon(\alpha, n_{\sigma_t} + 1)/\epsilon^\alpha} \\ &\overset{\text{(b)}}{\leq} \frac{\tilde{p}\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) + 1 - \tilde{p}}{\tilde{p}\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) + (1 - \tilde{p})(1 - \alpha\epsilon n_{\sigma_t}/(\alpha + 1))} \\ &= 1 + \frac{(1 - \tilde{p})\alpha\epsilon n_{\sigma_t}/(\alpha + 1)}{\tilde{p}\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) + (1 - \tilde{p})(1 - \alpha\epsilon n_{\sigma_t}/(\alpha + 1))} \\ &\overset{\text{(c)}}{\leq} 1 + \frac{\alpha n \epsilon}{\alpha + 1 - \alpha\epsilon n}, \end{aligned} \tag{12}$$

where (a) uses $g_{\sigma_k} + \alpha(1 - \tilde{p}) > g_{\sigma_k} + \alpha(1 - \tilde{p})(1 - \epsilon^{g_{\sigma_k}}) > 0$, (b) uses

$$\begin{aligned} &\frac{\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) I_\epsilon(\alpha, n_{\sigma_t} + 1)}{\epsilon^\alpha} = \frac{\alpha}{\epsilon^\alpha} \int_0^\epsilon x^{\alpha - 1}(1 - x)^{n_{\sigma_t}} \, dx \geq \frac{\alpha}{\epsilon^\alpha} \int_0^\epsilon x^{\alpha - 1}(1 - n_{\sigma_t} x) \, dx \\ &= 1 - \frac{\alpha\epsilon n_{\sigma_t}}{\alpha + 1}, \end{aligned}$$

and (c) uses $\tilde{p}\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1) > 0$, $n_{\sigma_t} \leq n$, and the assumption that $\epsilon \leq 1/n$.

Using the exchangeable partition probability functions in Theorem 1 and Lemma 2, along with Equations (11), (12) and (15), we have

$$\frac{p_{0,n}(\mathcal{A})}{p_{\epsilon,n}(\mathcal{A})} = \frac{\sum_{\sigma \in S_t} \prod_{j=1}^t U_j(\sigma)}{\sum_{\sigma \in S_t} \prod_{j=1}^t V_j(\sigma)} \leq 1 + \frac{\alpha n \epsilon}{\alpha + 1 - \alpha\epsilon n} \tag{13}$$

for all partitions $\mathcal{A}$. Therefore, the Kullback–Leibler divergence satisfies

$$D_{\mathrm{KL}}(p_{0,n}\|p_{\epsilon,n}) = \sum_{\mathcal{A}\in\cup_{t=1}^{n}\mathcal{H}_t(n)} p_{0,n}(\mathcal{A})\log\frac{p_{0,n}(\mathcal{A})}{p_{\epsilon,n}(\mathcal{A})} \leq \sum_{\mathcal{A}} p_{0,n}(\mathcal{A})\frac{\alpha n\epsilon}{\alpha+1-\alpha\epsilon n} = \frac{\alpha n\epsilon}{\alpha+1-\alpha\epsilon n},$$

where the sum is over all partitions of $\{1,\ldots,n\}$ and the inequality uses $\log(x) \leq x-1$. The result follows by Pinsker's inequality. ■

In the proof of Theorem 5, we employ the following two inequalities. Let $x_i, y_i \geq 0$ for all i in some countable set I, such that $\sum_{i\in I} x_i > 0$ and $\sum_{i\in I} y_i > 0$. First, if $A \subseteq I$ then

$$\begin{aligned}\left|\frac{\sum_{i\in A} x_i}{\sum_{i\in I} x_i} - \frac{\sum_{i\in A} y_i}{\sum_{i\in I} y_i}\right| &= \frac{|\sum_{i'\in A}\sum_{i\in I} x_{i'}y_i - \sum_{i'\in A}\sum_{i\in I} x_i y_{i'}|}{\sum_{i',i\in I} x_{i'}y_i} \\ &= \frac{|\sum_{i'\in A}\sum_{i\in A^c} x_{i'}y_i - \sum_{i'\in A}\sum_{i\in A^c} x_i y_{i'}|}{\sum_{i',i\in I} x_{i'}y_i} \\ &\leq \frac{\sum_{i'\in A}\sum_{i\in A^c}|x_{i'}y_i - x_i y_{i'}|}{\sum_{i',i\in I} x_{i'}y_i}\end{aligned} \tag{14}$$

where $A^c = I \setminus A$. Second, if $a_i \geq 0$, $y_i > 0$ for all $i \in I$, $A \subseteq I$, and $\sum_{i\in I} a_i y_i > 0$, then

$$\frac{\sum_{i\in A} a_i x_i}{\sum_{i\in I} a_i y_i} = \frac{\sum_{i\in A} a_i (x_i/y_i) y_i}{\sum_{i\in I} a_i y_i} \leq \frac{\sum_{i\in A} a_i (\max_{j\in A} x_j/y_j) y_i}{\sum_{i\in I} a_i y_i} \leq \max_{i\in A} x_i/y_i. \tag{15}$$

Proof of Theorem 5
In Theorem 3, we proved posterior consistency of the number of clusters in the case of $\epsilon = 0$, so we only need to show that $|p_{\epsilon(n)}(T = t \mid y_{1:n}) - p_{\epsilon=0}(T = t \mid y_{1:n})| \to 0$ as $n \to \infty$. We abbreviate $y = y_{1:n}$ to reduce notational clutter. First, using Equation (6), for any integer $1 \leq t \leq n$,

$$\begin{aligned}&|p_\epsilon(T = t \mid y) - p_{\epsilon=0}(T = t \mid y)| \\ &= \left|\frac{\sum_{\mathcal{A}\in\mathcal{H}_t(n)} p(y \mid \mathcal{A}) p_{\epsilon,n}(\mathcal{A})}{\sum_{\mathcal{A}\in\cup_{t=1}^{n}\mathcal{H}_t(n)} p(y \mid \mathcal{A}) p_{\epsilon,n}(\mathcal{A})} - \frac{\sum_{\mathcal{A}\in\mathcal{H}_t(n)} p(y \mid \mathcal{A}) p_{0,n}(\mathcal{A})}{\sum_{\mathcal{A}\in\cup_{t=1}^{n}\mathcal{H}_t(n)} p(y \mid \mathcal{A}) p_{0,n}(\mathcal{A})}\right| \\ &\overset{(a)}{\leq} \frac{\sum_{\mathcal{A}'\in\mathcal{H}_t(n)}\sum_{\mathcal{A}\in\cup_{l\neq t}\mathcal{H}_l(n)} p(y \mid \mathcal{A}')p(y \mid \mathcal{A})|p_{\epsilon,n}(\mathcal{A}')p_{0,n}(\mathcal{A}) - p_{0,n}(\mathcal{A}')p_{\epsilon,n}(\mathcal{A})|}{\sum_{\mathcal{A}',\mathcal{A}\in\cup_{t=1}^{n}\mathcal{H}_t(n)} p(y \mid \mathcal{A}')p(y \mid \mathcal{A})p_{\epsilon,n}(\mathcal{A}')p_{0,n}(\mathcal{A})} \\ &\overset{(b)}{\leq} \max_{\mathcal{A}'\in\mathcal{H}_t(n)}\max_{\mathcal{A}\in\cup_{l\neq t}\mathcal{H}_l(n)}\left|1 - \frac{p_{0,n}(\mathcal{A}')p_{\epsilon,n}(\mathcal{A})}{p_{\epsilon,n}(\mathcal{A}')p_{0,n}(\mathcal{A})}\right|,\end{aligned} \tag{16}$$

where (a) and (b) are by Equations (14) and (15), respectively. To ensure that the denominators in the preceding display are nonzero, there needs to exist at least one partition $\mathcal{A} \in \cup_{t=1}^{n}\mathcal{H}_t(n)$ such that $p(y \mid \mathcal{A}) > 0$, and this is indeed the case since (1) with probability 1, $p(y \mid \phi_0) > 0$, and (2) we assume the support of the prior $\mathcal{G}(\theta)$ contains a neighborhood of each θ_k^0 and the component density f_θ is continuous (with respect to θ) at each θ_k^0.

We use the same notation as in the proof of Theorem 4, where we have already proved that

$$\frac{p_{0,n}(\mathcal{A})}{p_{\epsilon,n}(\mathcal{A})} \leq 1 + \frac{\alpha n \epsilon}{\alpha + 1 - \alpha \epsilon n} \tag{17}$$

for any partition $\mathcal{A}$. Next, we construct an upper bound on the reciprocal, $p_{\epsilon,n}(\mathcal{A})/p_{0,n}(\mathcal{A})$, by upper bounding $V_j(\sigma)/U_j(\sigma)$ for each j. We split the analysis into two cases: $j \leq t-1$ and $j = t$.

Case 1: $j \leq t-1$. This implies $g_{j+1}(\sigma) \geq t - j \geq 1$ since every cluster has at least one element. Letting r denote the integer such that $0 < \alpha - r \leq 1$ (or equivalently, $\max(\alpha - 1, 0) \leq r < \alpha$),

$$\begin{aligned}
&\frac{1}{\epsilon^\alpha} I_\epsilon(g_{j+1}(\sigma) + \alpha, n_{\sigma_j} + 1) \\
&\leq \frac{1}{\epsilon^\alpha \mathrm{B}(g_{j+1}(\sigma) + \alpha, n_{\sigma_j} + 1)} \int_0^\epsilon x^{g_{j+1}(\sigma)+\alpha-1}\, dx \\
&= \frac{\Gamma(g_j(\sigma) + \alpha + 1)\epsilon^{g_{j+1}(\sigma)}}{\Gamma(g_{j+1}(\sigma) + \alpha)\Gamma(n_{\sigma_j} + 1)(g_{j+1}(\sigma) + \alpha)} \\
&= \epsilon^{g_{j+1}(\sigma)} \frac{(g_j(\sigma) + \alpha)(g_j(\sigma) + \alpha - 1)\cdots(\alpha - r)}{(g_{j+1}(\sigma) + \alpha)(g_{j+1}(\sigma) + \alpha - 1)\cdots(\alpha - r)} \frac{1}{n_{\sigma_j}!} \\
&= \epsilon^{g_{j+1}(\sigma)} \left(\prod_{m=0}^{g_{j+1}(\sigma)+r} \frac{n_{\sigma_j} + \alpha - r + m}{\alpha - r + m} \right) \left(\prod_{m=1}^{n_{\sigma_j}} \frac{m + \alpha - r - 1}{m} \right) \\
&\overset{\text{(a)}}{\leq} \epsilon^{g_{j+1}(\sigma)} \left(1 + \frac{n_{\sigma_j}}{\alpha - r} \right)^{g_{j+1}(\sigma)+r+1} \\
&\overset{\text{(b)}}{\leq} \epsilon^{t-j} \left(1 + \frac{n}{\alpha - r} \right)^{t-j+r+1}
\end{aligned}$$

for all n sufficiently large, where (a) results from $(n_{\sigma_j} + \alpha - r + m)/(\alpha - r + m) \leq (n_{\sigma_j} + \alpha - r)/(\alpha - r)$ for $m = 0, 1, \ldots, g_{j+1}(\sigma) + r$ and $(m + \alpha - r - 1)/m \leq 1$ for $m = 1, \ldots, n_{\sigma_j}$, and (b) uses $n_{\sigma_j} \leq n$, $\epsilon(1 + n/(\alpha - r)) < 1$ for all n sufficiently large since $\epsilon = \epsilon(n) = o(1/n)$, and $g_{j+1}(\sigma) \geq t - j$.

Combining this with

$$\frac{g_j(\sigma) + \alpha(1 - \tilde{p})}{g_j(\sigma) + \alpha(1 - \tilde{p})(1 - \epsilon^{g_j(\sigma)})} = 1 + \frac{\alpha(1 - \tilde{p})\epsilon^{g_j(\sigma)}}{g_j(\sigma) + \alpha(1 - \tilde{p})(1 - \epsilon^{g_j(\sigma)})} \leq 1 + \alpha(1 - \tilde{p})\epsilon, \tag{18}$$

we have

$$\frac{V_j(\sigma)}{U_j(\sigma)} \leq \left(1 + \frac{1 - \tilde{p}}{\tilde{p}} \epsilon^{t-j} \Big(1 + \frac{n}{\alpha - r} \Big)^{t-j+r+1} \right) \big(1 + \alpha(1 - \tilde{p})\epsilon \big).$$

Case 2: $j = t$. We have

$$\frac{V_t(\sigma)}{U_t(\sigma)} \leq 1 + \alpha(1 - \tilde{p})\epsilon$$

since Equation (18) holds when $j = t$ and

$$\frac{I_\epsilon(\alpha, n_{\sigma_t} + 1)}{\epsilon^\alpha} \leq \frac{1}{\epsilon^\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1)} \int_0^\epsilon x^{\alpha-1}\, dx = \frac{1}{\alpha \mathrm{B}(\alpha, n_{\sigma_t} + 1)}.$$

Thus, using the same expression as in Equation (13),

$$\begin{aligned}
\frac{p_{\epsilon,n}(\mathcal{A})}{p_{0,n}(\mathcal{A})} &\leq \max_{\sigma \in S_t} \prod_{j=1}^{t} \frac{V_j(\sigma)}{U_j(\sigma)} \\
&\leq \big(1+\alpha(1-\tilde{p})\epsilon\big)^t \prod_{j=1}^{t-1} \left(1 + \frac{1-\tilde{p}}{\tilde{p}} \epsilon^{t-j} \Big(1 + \frac{n}{\alpha - r}\Big)^{t-j+r+1}\right) \\
&\leq \big(1+\alpha(1-\tilde{p})\epsilon\big)^n \left(1 + \frac{1-\tilde{p}}{\tilde{p}} \epsilon^2 \Big(1 + \frac{n}{\alpha - r}\Big)^{3+r}\right)^n \left(1 + \frac{1-\tilde{p}}{\tilde{p}} \epsilon \Big(1 + \frac{n}{\alpha - r}\Big)^{2+r}\right) \qquad (19)
\end{aligned}$$

since $t = |\mathcal{A}| \leq n$ and $\epsilon(1+n/(\alpha-r)) < 1$ for n sufficiently large, because $\epsilon = \epsilon(n) = o(1/n)$.

Since $\epsilon(n) = o(1/n^{2+r})$, the upper bound on $p_{\epsilon(n),n}(\mathcal{A})/p_{0,n}(\mathcal{A})$ in Equation (19) converges to 1 as $n \to \infty$. Meanwhile, Equation (17) provides a lower bound that also converges to 1. Since these upper and lower bounds depend only on α, r, $\tilde{p}$, ϵ, and n, they hold uniformly over all $\mathcal{A} \in \cup_{t=1}^{n} \mathcal{H}_t(n)$. Hence, $p_{\epsilon(n),n}(\mathcal{A})/p_{0,n}(\mathcal{A}) \to 1$ uniformly over $\mathcal{A}$, as $n \to \infty$. Therefore, along with Equation (16), this implies that

$$|p_{\epsilon(n)}(T = t \mid y) - p_{\epsilon=0}(T = t \mid y)| \leq \max_{\mathcal{A}' \in \mathcal{H}_t(n)} \max_{\mathcal{A} \in \cup_{l \neq t} \mathcal{H}_l(n)} \left| 1 - \frac{p_{0,n}(\mathcal{A}') p_{\epsilon(n),n}(\mathcal{A})}{p_{\epsilon(n),n}(\mathcal{A}') p_{0,n}(\mathcal{A})} \right| \longrightarrow 0$$

as $n \to \infty$. ■

Proof of Lemma 6

To study the posterior distribution of the number of clusters $p(T = t \mid y_{1:n})$, we will focus on

$$p(y_{1:n}, T = t) = \sum_{\mathcal{A} \in \mathcal{H}_t(n)} p(y_{1:n} \mid \mathcal{A}) p(\mathcal{A}) \qquad (20)$$

where $\mathcal{A} = \{A_1, \ldots, A_t\}$, $p(y_{1:n} \mid \mathcal{A}) = \prod_{A \in \mathcal{A}} m(y_A)$, $y_A = (y_i : i \in A)$, and $m(y_A) = \int_\Theta \big(\prod_{i \in A} f_\theta(y_i)\big)\, d\mathcal{G}(\theta)$, as described at Equation (6). In this lemma, $f_\theta(y_i) = \mathrm{N}(y_i \mid \theta, 1)$ and

$$m(y_A) = \frac{1}{\sqrt{|A|+1}} f_0(y_A) \exp\left(\frac{(\sum_{i \in A} y_i)^2}{2(|A|+1)}\right), \qquad (21)$$

where $f_0(y_A) = \prod_{j \in A} \mathrm{N}(y_j \mid 0, 1)$. Under the Dirichlet process prior, the probability mass function on partitions is $p(\mathcal{A}) = \frac{\alpha^t}{\alpha^{(n)}} \prod_{i=1}^{t} (|A_i| - 1)!$ where $\alpha^{(n)} = \alpha(\alpha+1)\cdots(\alpha+n-1)$.

Hence,

$$\begin{aligned}
\frac{p(y_{1:n}, T=2)}{p(y_{1:n}, T=1)} &\overset{\text{(a)}}{=} \frac{\sum_{\mathcal{A}=\{A_1,A_2\}\in\mathcal{H}_2(n)} p(\mathcal{A}) m(y_{A_1}) m(y_{A_2})}{\mathbb{P}(\mathcal{A}=\{\{1,\dots,n\}\}) m(y_{1:n})} \\
&\overset{\text{(b)}}{=} \sum_{\mathcal{A}=\{A_1,A_2\}\in\mathcal{H}_2(n)} \Bigg[\frac{\alpha^2(|A_1|-1)!(|A_2|-1)!}{\alpha(n-1)!} \times \frac{\sqrt{n+1}}{\sqrt{|A_1|+1}\sqrt{|A_2|+1}} \\
&\qquad \times \exp\left(\frac{(\sum_{i\in A_1} y_i)^2}{2(|A_1|+1)}\right) \exp\left(\frac{(\sum_{i\in A_2} y_i)^2}{2(|A_2|+1)}\right) \exp\left(-\frac{(\sum_{i=1}^n y_i)^2}{2(n+1)}\right)\Bigg] \\
&\overset{\text{(c)}}{\leq} \frac{\alpha\sqrt{n+1}}{(n-1)!} \sum_{\mathcal{A}\in\mathcal{H}_2(n)} \frac{(|A_1|-1)!(|A_2|-1)!}{\sqrt{|A_1|+1}\sqrt{|A_2|+1}} \exp\left(\frac{\sum_{i\in A_1} y_i^2}{2} + \frac{\sum_{i\in A_2} y_i^2}{2}\right) \\
&\overset{\text{(d)}}{\leq} \frac{\alpha\sqrt{n+1}}{2(n-1)!} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \sum_{\mathcal{A}\in\mathcal{H}_2(n)} (|A_1|-1)!(|A_2|-1)! \\
&= \frac{\alpha\sqrt{n+1}}{2(n-1)!} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \sum_{i=1}^{n-1} \frac{1}{2}\binom{n}{i} (i-1)!(n-i-1)! \\
&= \frac{\alpha\sqrt{n+1}}{2} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \sum_{i=1}^{n-1} \frac{n}{2i(n-i)} \\
&= \frac{\alpha\sqrt{n+1}}{2} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1}\right) \\
&\leq \frac{\alpha\sqrt{n+1}}{2} \exp\left(\frac{\sum_{i=1}^n y_i^2}{2}\right) \big(\log(n-1)+1\big)
\end{aligned}$$

where (a) is using Equation (20) for both numerator and denominator, (b) is using Equation (21) and the probability mass function of $\mathcal{A}$, (c) follows from $(\sum_{j=1}^n y_j)^2 \geq 0$ and Jensen's inequality, (d) is using $A_1 \cup A_2 = \{1,\dots,n\}$ and both $|A_1|$ and $|A_2|$ are greater than or equal to 1, and the last inequality is induced from $1/k \leq \log(k) - \log(k-1)$ for $k = 2,\dots,n-1$.

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n y_i^2 \xrightarrow[n\to\infty]{\text{a.s.}} E(y_1^2) = 1 + \kappa^2/2.$$

Let $\delta = C - (1/2 + \kappa^2/4)$, where by the assumption of the theorem, $C > 1/2 + \kappa^2/4$ such that $\alpha = \alpha(n) = o(\exp(-C\,n))$. Then, almost surely, for all n sufficiently large, $\frac{1}{2n}\sum_{i=1}^n y_i^2 \leq 1/2 + \kappa^2/4 + \delta/2 = C - \delta/2$. Hence, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{p(y_{1:n}, T=2)}{p(y_{1:n}, T=1)} &\leq \frac{\alpha(n)\sqrt{n+1}}{2} \exp\big(n(C-\delta/2)\big)\big(\log(n-1)+1\big) \\
&= \big(\alpha(n)\exp(C\,n)\big)\exp(-n\delta/2)\frac{\sqrt{n+1}}{2}\big(\log(n-1)+1\big) \longrightarrow 0
\end{aligned}$$

as $n \to \infty$. Therefore, we have the conclusion,

$$p(T=2 \mid y_{1:n}) = \frac{p(y_{1:n}, T=2)}{\sum_{t=1}^\infty p(y_{1:n}, T=t)} \leq \frac{p(y_{1:n}, T=2)}{p(y_{1:n}, T=1)} \xrightarrow{\text{a.s.}} 0.$$

■

Appendix B. Additional Simulation Results

In this section, we first provide the simulation results when the component distribution is the bivariate Gaussian distribution. Then we show a comparison between the quasi-Bernoulli mixture and the Dirichlet process mixture with $\alpha(n) \to 0$. Finally, we provide information on convergence diagnostics and the running time of the algorithm.

B.1 Simulation Results with Bivariate Gaussian Mixtures

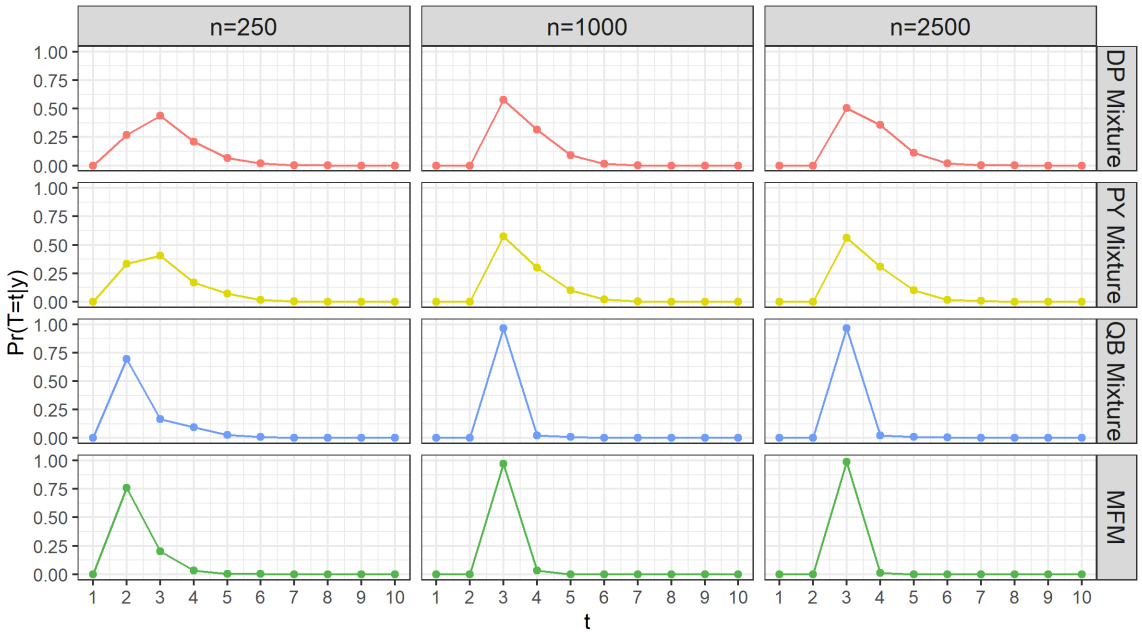
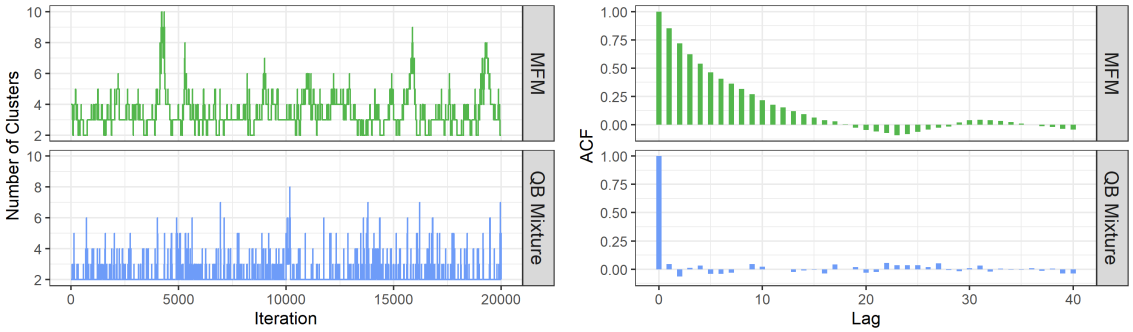


Figure 7: Posterior distribution on the number of clusters for data generated from a three-component Gaussian mixture in $\mathbb{R}^2$. Using the quasi-Bernoulli mixture model, the posterior probabilities of T concentrate to a point mass at $k_0 = 3$ for large n. However, the posterior distributions of the calibrated Dirichlet process mixture model and Pitman–Yor process mixture model do not concentrate to a point mass at $k_0 = 3$.

Figure 7 plots the posterior distribution of the number of clusters T at each n. Under the quasi-Bernoulli mixture model (shown in blue), the posterior of T concentrates to a point mass at the true number of components ($k_0 = 3$) as n grows, in accordance with our theory. On the other hand, the posterior distributions of T with the Dirichlet process mixture model and the Pitman–Yor process mixture model fail to concentrate to a point mass at the true number of components.

As shown in Figure 8, the MFM model suffers from slow mixing with high auto-correlation even after thinning (effective sample size 15.3% on average of five experiments

with sample size 250); whereas the quasi-Bernoulli mixture model quickly shows a much faster drop in the auto-correlation within a few lags (effective sample size 57.2%).



(a) Trace on number of clusters (b) Auto-correlation (based on thinning at 25)

Figure 8: The trace of the Markov chain on T and auto-correlation functions for bivariate Gaussian mixture data with sample size 250. Quasi-Bernoulli mixture model shows much better mixing in the Markov chain, compared to the MFM model. We discard the first 5,000 iterations as burn-ins and record the following 20,000 samples.

B.2 Simulations on the Dirichlet Process Mixture with $\alpha(n) \to 0$

To compare the quasi-Bernoulli mixture model with the Dirichlet process mixture model under different rates of $\alpha(n) \to 0$, we conduct additional experiments with univariate Gaussian mixtures. The experimental settings are similar to the ones in Section 4.1, except that we generate data from mixtures with smaller distances between the component centers: $0.3\,\mathrm{N}(-2, 1^2) + 0.4\,\mathrm{N}(0, 1^2) + 0.3\,\mathrm{N}(2, 1^2)$ under sample sizes $n \in \{100, 250, 1000, 2500\}$. The purpose is to examine if each model shows the trend of converging to the ground truth $T = k_0$ as n increases, when the component distribution $\mathcal{F}$ has an unbounded support and clusters have large overlaps.

For Dirichlet process mixture models, we use three rates $\alpha_1(n) = \exp(-n/10)$, $\alpha_2(n) = 4/\log(n)$ and $\alpha_3(n) = 20/n$. For quasi-Bernoulli mixture models, we use two rates $\epsilon_1(n) = n^{-2.1}$ and $\epsilon_2(n) = n^{-3.1}$. Figure 9 shows the posterior distributions of the number of clusters.

This empirical result suggests that one may be able to obtain posterior consistency on estimating T for the Dirichlet process mixture model with general $\mathcal{F}$, by choosing an $\alpha \to 0$ faster than $4/\log(n)$ but slower than $20/n$, although the theory remains an open question. On the other hand, the quasi-Bernoulli mixture models show almost no difference in the trend of convergence. This is as expected, since both $n^{-2.1}$ and $n^{-3.1}$ satisfy the rate condition that guarantees consistency on estimating T.

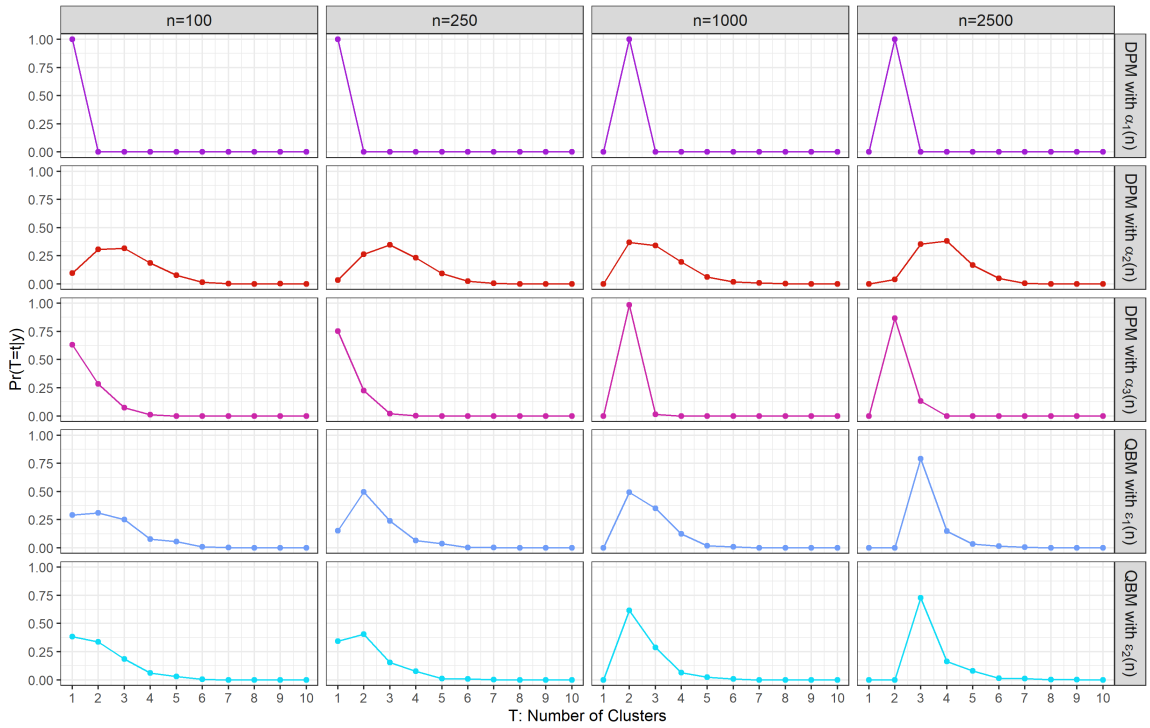
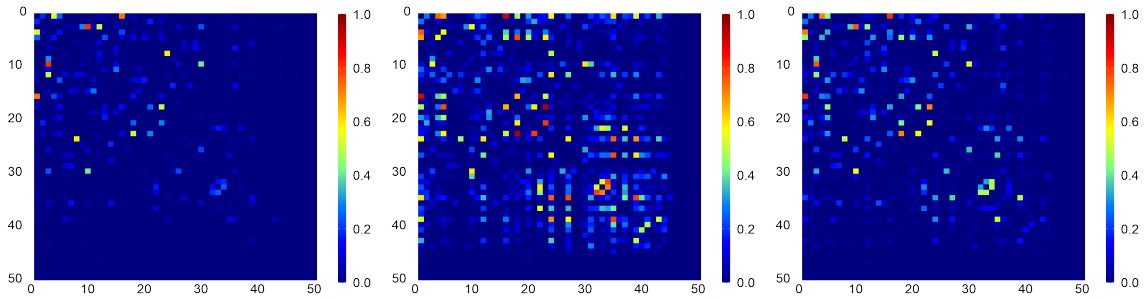


Figure 9: Posterior distribution on the number of clusters for data generated from a three-component univariate Gaussian mixture.

B.3 Convergence Diagnostics and Timing Information

We use the Markov chain sample of T (the number of clusters) for convergence diagnosis. We choose T because it is often the slowest-changing variable. As shown in Figure 3, the auto-correlations for T show a quick drop to an insignificant level, within as few lags after thinning at 50. For each experiment, we run multiple chains from 5 randomly initialized points, and compute the $\hat{R}$ statistic (Gelman and Rubin, 1992). All of the experiments get $\hat{R}$ close to 1, which means that the Markov chains have converged.

We provide the timing information of our posterior sampling algorithms. The algorithms are implemented in R, and run on a 4.0 GHz processor. For the model with univariate Gaussian components, each iteration costs $0.0005, 0.0006, 0.0010, 0.0035, 0.0078$ seconds for the sample size $50, 100, 250, 1000, 2500$, respectively. For the bivariate Gaussian case, the algorithm runs $0.0029, 0.0236, 0.0945$ seconds for each iteration for sample size $250, 1000, 2500$, respectively. When the component distribution is the Laplace distribution, the algorithm takes $0.0055, 0.1018, 0.2269, 0.3083$ seconds for each iteration for sample size $50, 200, 500, 1000$, respectively. For the network model used in the data application, each iteration takes around 0.3 seconds.



(a) Group 1 (36.6% of the subjects) (b) Group 2 (20.6% of the subjects) (c) Group 3 (42.8% of the subjects)

Figure 10: The MAP estimation of the mean of each Gaussian component under the mixture of factor analyzers model.

Appendix C. Other Useful Results

	Dirichlet process		Pitman–Yor process				quasi-Bernoulli process	
	α	$\mathbb{E}(T)$	α	d	$\mathbb{E}(T)$	$\mathrm{Var}(T)$	$\mathbb{E}(T)$	$\mathrm{Var}(T)$
50	0.71	3.62	0.48	0.09	3.62	3.07	3.64	3.09
100	0.69	4.04	0.46	0.09	4.10	3.95	4.06	3.97
250	0.67	4.58	0.38	0.10	4.56	5.39	4.59	5.40
1000	0.63	5.25	0.29	0.11	5.25	8.25	5.28	7.92
2500	0.61	5.69	0.29	0.10	5.74	9.51	5.69	9.79

Table 1: Settings of the hyper-parameters for the Dirichlet processes and the Pitman–Yor processes when sample sizes $n \in \{50, 200, 500, 1000, 2500\}$. We match the expectations of the number of clusters T under the three priors, and also make the variances of T close under the Pitman-Yor process prior and the quasi-Bernoulli process prior (with $\tilde{p} = 0.9$). Each expectation and variance are approximated based on 2×10^5 samples from the prior.

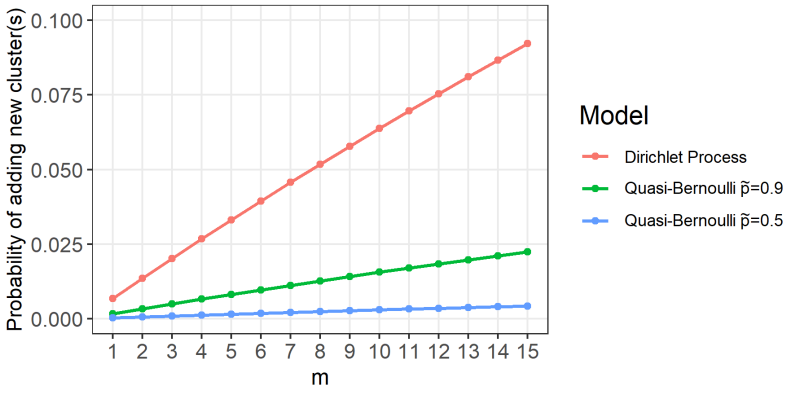


Figure 11: The probability of adding one or more new clusters for m future data points ($n = 100$). All of the parameters are chosen to be the same as Figure 1 except for $\epsilon = \epsilon(n, m) = 1/(n + m)^5$. Under the prior, the Dirichlet process exhibits rapid growth in this probability, favoring the creation of additional clusters *a priori*. Meanwhile, the quasi-Bernoulli process exhibits much slower growth of this probability.

References

Filippo Ascolani, Antonio Lijoi, Giovanni Rebaudo, and Giacomo Zanella. Clustering consistency with dirichlet process mixtures. *Biometrika*, 110(2):551–558, 2023.

David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

Diana Cai, Trevor Campbell, and Tamara Broderick. Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning*, pages 1158–1169. PMLR, 2021.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

Noirrit Kiran Chandra, Antonio Canale, and David B Dunson. Escaping the curse of dimensionality in bayesian model based clustering. *arXiv preprint arXiv:2006.02700*, 2022.

David B Dunson and Ju-Hyun Park. Kernel stick-breaking processes. *Biometrika*, 95(2): 307–323, 2008.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Matthew Heiner, Athanasios Kottas, and Stephan Munch. Structured priors for sparse probability vectors with application to model selection in markov chains. *Statistics and Computing*, 29(5):1077–1093, 2019.

Peter D. Hoff. Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009.

Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

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

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

Daniel Marcus, John Harwell, Timothy Olsen, Michael Hodge, Matthew Glasser, Fred Prior, Mark Jenkinson, Timothy Laumann, Sandra Curtiss, and David Van Essen. Informatics and data mining tools and strategies for the human connectome project. *Frontiers in Neuroinformatics*, 5:4, 2011.

Geoffrey J McLachlan, David Peel, and Richard W Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3-4):379–388, 2003.

Jeffrey W Miller. An elementary derivation of the chinese restaurant process from sethuraman's stick-breaking process. *Statistics & Probability Letters*, 146:112–117, 2019.

Jeffrey W Miller and Matthew T Harrison. A simple example of dirichlet process mixture inconsistency for the number of components. *Advances in neural information processing systems*, 26, 2013.

Jeffrey W. Miller and Matthew T. Harrison. Inconsistency of pitman-yor process mixtures for the number of components. *Journal of Machine Learning Research*, 15(96):3333–3370, 2014.

Jeffrey W. Miller and Matthew T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

Agostino Nobile. *Bayesian Analysis of Finite Mixture Distributions*. PhD thesis, PhD Thesis. Carnegie Mellon University, Pittsburgh, 1994.

Ilsang Ohn and Lizhen Lin. Optimal bayesian estimation of gaussian mixtures with growing number of components. *Bernoulli*, 29(2):1195–1218, 2023.

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

Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

Lu Ren, Lan Du, Lawrence Carin, and David B Dunson. Logistic stick-breaking process. *Journal of Machine Learning Research*, 12(1), 2011.

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

Abel Rodríguez, David B Dunson, and Alan E Gelfand. Latent stick-breaking processes. *Journal of the American Statistical Association*, 105(490):647–659, 2010.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Jayaram Sethuraman. A constructive definition of the dirichlet prior. *Statistica Sinica*, 4: 639–650, 1994.

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

Chiao-Yu Yang, Eric Xia, Nhat Ho, and Michael I Jordan. Posterior distribution for the number of clusters in dirichlet process mixture models. *arXiv preprint arXiv:1905.09959*, 2020.