5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–11

Clustering inconsistency for Pitman–Yor mixture models with a prior on the precision but fixed discount parameter

Anonymous Authors
Anonymous Institution

Abstract

Bayesian nonparametric (BNP) mixture models such as Dirichlet process (DP) and Pitman–Yor process (PY) mixture models are popular to model complex data. Their posterior distributions exhibit nice theoretical properties, converging at the optimal minimax rate to the true data-generating distribution, and extensive research has been devoted to developing this theory. However, consistency of the posterior distribution does not imply consistency of the number of clusters, and asymptotic guarantees for the posterior number of clusters of these BNP mixture models have been lacking until recently. Recent research has shown that these models can be inconsistent for the number of clusters. In the case of DP mixture models, this problem can be avoided when a prior is put on the model's concentration hyperparameter α, as is common practice. In this work, we prove that PY mixture models remain inconsistent for the number of clusters when a prior is put on α, in the special case where the true number of components in the data generating mechanism is equal to 1 and the discount parameter $\sigma \in (0, 1)$ is a fixed constant.

1. Introduction

Mixture models, popular for their flexibility and simplicity, are commonly used in the statistical analysis of heterogeneous data where observations are assumed to come from an unknown number of different populations. Since in a mixture each observation is assumed to come from one population, such models naturally induce clustering in the data: two data points are associated with the same cluster if they come from the same population. We focus on the problem of inferring the number of clusters in the data.

One solution is to fit the data with a number of different mixture models, each with a different number of components and to select the best model in terms of some criterion, such as the Akaike information criterion (AIC) or the Bayes information criterion (BIC). This method, however, may be computationally expensive since many models must be fitted. A Bayesian approach could alternatively be taken by putting a parametric prior (such as a Poisson) on the number of components, but inference can be challenging when the dimensionality or the amount of data becomes large (although new strategies have been proposed recently Miller and Harrison, 2018).

In this work, we consider infinite mixture models where the mixing measure is modelled with a nonparametric prior. In such models, the number of components possible has no upper bound. Inference may be performed in a unified way without the need for strong assumptions on the number of components and with no need to fit multiple models.

While the most standard nonparametric prior remains the Dirichlet process (DP) introduced by Ferguson (1973), many extensions now exist. In this work, we focus on the Pitman–Yor process (PY, Pitman and Yor, 1997), which is a natural extension of the DP with an extra parameter allowing for an increase in model flexibility. Compared with the

© A. Authors.

DP mixtures, PY mixtures are better suited for mixtures where the sizes of clusters are more evenly distributed. Due to the interpretability of their parameters, their ease of implementation, and their nice mathematical properties, DP and PY priors are widely used in practice, and in the last two decades a huge amount of research has focused on their properties (see for example Ghosal and Van der Vaart, 2017; Müller et al., 2018). The use of the DP as a mixing measure was first introduced by Lo (1984). Thanks to the wide variety of efficient computational methods which have been introduced for their inference (Escobar and West, 1998; MacEachern and Müller, 1998; Neal, 2000; Blei et al., 2006), nonparametric mixture models have become common in a wide range of modeling applications.

In the context of density estimation, it has been shown that under certain conditions, the posterior distribution of DP mixture models concentrates at the true data-generating density at the minimax-optimal rate (Ghosal and Van der Vaart, 2017; Ghosal et al., 1999). This result has then been extended to other types of Bayesian nonparametric priors, such as PY priors (Lijoi et al., 2005). Nguyen (2013) further proved posterior consistency of the mixing distribution in the Wasserstein metric DP and PY mixture models.

It is important to realize that consistency of the posterior distribution and even of the mixing measure does not imply consistency of the inferred number of clusters. Empirically, many researchers have observed this, noticing that DP mixture posteriors tend to overestimate the number of clusters (West and Escobar, 1993; Lartillot and Philippe, 2004; Onogi et al., 2011). More recently, Miller and Harrison (2013, 2014) proved theoretically non-consistency for the number of components in DP and PY mixtures, respectively. Alamichel et al. (2022) extend this result to the case of Gibbs-type processes and finite-dimensional representations of BNP priors. A possible explanation for this inconsistency result can be found in a result proved by Rousseau and Mengersen (2011), that in overfitted finite or infinite mixture models, the weight attributed to extra cluster goes to zero as the number of observations grows. Provided that the weights for the extra components are infinitesimally small, any mixture can be approximated arbitrarily well by a mixture with a larger number of components.

Despite the above inconsistency results, it can be possible to achieve posterior consistency for the number of clusters in the mixture models we consider. Guha et al. (2021) introduce a fast and simple post-processing procedure which, once applied to samples from the posterior distribution of a DP mixture, results in clustering consistency. Alamichel et al. (2022) extend this result to PY mixtures. Ascolani et al. (2022) show that posterior consistency can be achieved by putting a prior on the DP concentration parameter α. DP mixtures modeled this way can be considered as mixtures of DP mixtures (Antoniak, 1974), and are more common in practice than DP mixture models with fixed α.

We show that Ascolani et al. (2022)'s result can not be directly extended to PY mixtures: we prove clustering inconsistency for Pitman–Yor process mixture models when a prior is put on the concentration parameter, in the special case where the true number of clusters in the data generating mechanism, t, is equal to one, and when the discount parameter $\sigma \in (0, 1)$ is a fixed constant.

2. Preliminaries

Formally, we assume data $X_{1:n} \in \mathbb{X}^n$ is generated by a mechanism of the following form:

$$X_i \overset{\text{iid}}{\sim} P = \sum_{j=1}^{t} p_j k(\cdot|\theta_j^\star), \tag{1}$$

where the p_j are probability weights in $(0,1)$ summing to one, and where the $k(\cdot|\theta_j^\star)$ are probability kernels, each depending on some parameter $\theta_j^\star$. The above may alternatively be expressed as a convolution of the component-specific kernel $k(\cdot|\theta)$ with the discrete mixing measure $G = \sum_{j=1}^{t} p_j \delta_{\theta_j^\star}$:

$$P(x) = \int k(x|\theta) G(\mathrm{d}\theta).$$

We consider the well-specified case where the kernel density $k(\cdot|\theta)$ is known, but where the integer t, the weights p_j, and the latent variables $\theta_j^\star$ in Equation (1) are all unknown. In order to allow for an unbounded number of components t in the mixture, we consider nonparametric mixture models, with nonparametric priors on the mixing measure G.

Ascolani et al. (2022) consider Dirichlet process mixture models with a prior on the concentration parameter α:

$$X_i|\theta_i \overset{\text{ind}}{\sim} k(\cdot|\theta_i), \quad \theta_i|\tilde{P} \overset{\text{iid}}{\sim} \tilde{P} \quad \tilde{P}|\alpha \sim \mathrm{DP}(\alpha, Q_0), \quad \alpha \sim \pi_1, \tag{2}$$

where π_1 is a prior distribution on α, and Q_0 is the DP base measure.

We consider an extension of Ascolani et al. (2022)'s model, which are Pitman–Yor mixture models with a prior on the concentration parameter α and with a fixed discount parameter $\sigma \in (0,1)$:

$$X_i|\theta_i \overset{\text{ind}}{\sim} k(\cdot|\theta_i), \quad \theta_i|\tilde{P} \overset{\text{iid}}{\sim} \tilde{P} \quad \tilde{P}|\alpha, \sigma \sim \mathrm{PY}(\alpha, \sigma, Q_0), \quad \alpha \sim \pi_1. \tag{3}$$

Following the notation of Ascolani et al. (2022), for every pair of numbers $(n, s) \in \mathbb{N}^2$ with $s \leq n$, we let $\tau_s(n)$ denote the set of partitions of $\{1, \ldots, n\}$ into s non empty subsets. Conditional on parameters α and σ, a Pitman–Yor mixture model induces the following prior distribution on the space of partitions on n, for any $n \in \mathbb{N}$, and any $A = \{A_1, \ldots, A_s\} \in \tau_s(n), s \leq n$,

$$p(A|\alpha, \sigma) = \frac{\sigma^{s-1}(1 + \frac{\alpha}{\sigma})_{(s)}}{(1+\alpha)_{(n)}} \prod_{j=1}^{s} (1-\sigma)_{(a_j)}, \tag{4}$$

where $\alpha_{(n)} = \alpha \cdots (\alpha + n - 1)$ is the ascending factorial and $a_j = |A_j|$ stands for the cardinality of the set A_j. Conditionally on the partition A, the probability distributions of the data $X_{1:n} = (X_1, \ldots, X_n)$ and of the cluster-specific parameters $\hat{\theta}_{1:s} = (\hat{\theta}_1, \ldots, \hat{\theta}_s)$ are

$$p(X_{1:n}|\hat{\theta}_{1:s}, A) = \prod_{j=1}^{s} \prod_{i \in A_j} k(X_i|\hat{\theta}_j), \quad p(\hat{\theta}_{1:s}|A, \theta) = p(\hat{\theta}_{1:s}|A) = \prod_{j=1}^{s} q_0(\hat{\theta}_j).$$

We use the standard notation K_n to denote the number of clusters in a sample of size n. Under our model (3), K_n has the following prior distribution

$$p(K_n = s|\sigma) = \int \sum_{A \in \tau_s(n)} p(A|\alpha, \sigma)\pi_1(\mathrm{d}\alpha)$$

where $p(A|\alpha, \sigma)$ is as above.

In order to study the asymptotic behavior of the number of clusters, we will be interested in the quantity $p(K_n = s|X_{1:n}, \sigma)$. We start by looking at the joint distribution of $(X_{1:n}, K_n|\sigma)$ which, for every $x_{1:n} = (x_1, \ldots, x_n) \in \mathbb{X}^n$, is given by

$$p(X_{1:n} = x_{1:n}, K_n = s|\sigma) = \sum_{A \in \tau_s(n)} p(A|\sigma) \prod_{j=1}^{s} m(x_{A_j})$$

where $p(A|\sigma) = \int p(A|\alpha, \sigma)\pi_1(\mathrm{d}\alpha)$ and $m(x_{A_j}) = \int \prod_{i \in A_j} k(x_i|\theta)q_0(\theta)\mathrm{d}\theta$ is the marginal likelihood for the subset of observations identified by A_j, given that they are clustered together.

3. Theoretical result

Throughout, we make the same assumptions as Ascolani et al. (2022). The first set of assumptions referred to as A1, A2, and A3, regard the prior π_1 for the precision parameter α: it is assumed to be absolutely continuous with respect to the Lebesgue measure, to have a polynomial behaviour around the origin, and to have subfactorial moments. Ascolani et al. (2022) prove in their Lemma 1 that common families of prior such as distributions with bounded support or the gamma distribution satisfy these assumptions. The second set of assumptions regards the type of mixture kernels $k(\cdot|\cdot)$ considered: attention is restricted to location families, where the kernel is of the form $k(x|\theta) = g(x - \theta)$ for some density function g on $\mathbb{R}$. More specifically, assumptions B1 and B2 of Ascolani et al. (2022) require that g be strictly positive and differentiable with bounded derivative on some interval, and zero elsewhere. Finally, assumption B3 requires that the BNP process base measure Q_0 is absolutely continuous with respect to Lebesgue measure, with bounded density q_0.

Theorem 1 *Suppose that the prior π_1 over the concentration parameter α, the kernel k, and density q_0 satisfy assumptions of Ascolani et al. (2022) recalled above. For every P as in (1), for $t = 1$, we have*

$$p(K_n = 1|X_{1:n}) \not\to 1 \text{ as } n \to \infty.$$

The proof of Theorem 1 rests on analysing the ratio $\frac{p(K_n=s|X_{1:n})}{p(K_n=1|X_{1:n})}$, as consistency cannot hold if it does not converge to 0 as $n \to \infty$. Following the strategy of Ascolani et al. (2022), this ratio can be split into the product of two quantities, one capturing the impact of the prior distribution on the concentration parameter α, and the other independent of the prior on α. In the Dirichlet process case with a prior on α, the first quantity goes to 0 and the second remains bounded, which plays a key role for consistency. We show in the proof that in the Pitman–Yor case, the σ parameter enters the first quantity and prevents it from vanishing as $n \to \infty$, destroying consistency and highlighting a fundamental difference between the DP and PY processes.

4. Simulation study

To illustrate the different results on the Pitman–Yor process mixture model, we consider a simulation study. The simulated data are generated using a Gaussian location mixture. More precisely, the data came from $t = 3$ components with Gaussian kernels such that:

$$P(x) = \sum_{i=1}^{3} p_i \, \mathcal{N}(x|\mu_i, \Sigma),$$

where $p = (p_1, p_2, p_3)$ are the weights fixed as $p = (0.5, 0.3, 0.2)$ and $\mathcal{N}(x|\mu_i, \Sigma)$ is a multivariate Gaussian with mean μ_i and covariance matrix Σ. The following parameters are considered for the mean and the covariance matrix,

$$\mu_1 = (0.8, 0.8),\ \mu_2 = (0.8, -0.8),\ \mu_3 = (-0.8, 0.8) \quad \text{and} \quad \Sigma = 0.05\, I_2.$$

We use an adaptation of the code in Alamichel et al. (2022), which allows to sample from a Dirichlet multinomial process mixture model with a prior on the concentration parameter α. As we need to sample from a Pitman–Yor mixture model, we adapt the Importance Conditional Sampler introduced by Canale et al. (2022). We use parameters with the following prior distributions:

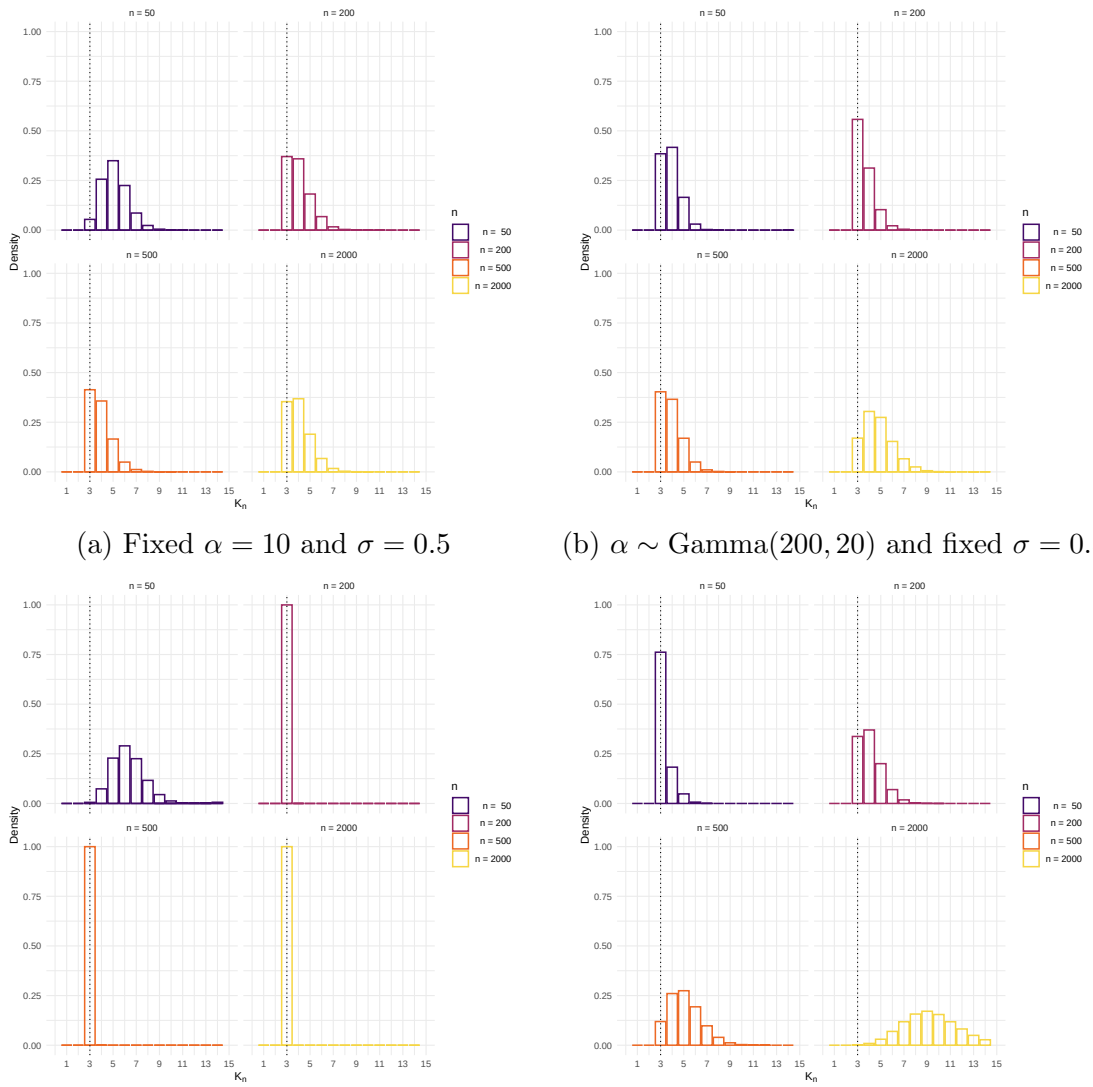
$$\begin{aligned} \tilde{P} &\sim \mathrm{PY}(\alpha, \sigma, Q_0), \\ \mu_i &\sim \mathcal{N}(b_0, B_0), \quad i = 1, \ldots, t, \\ \Sigma^{-1} &\sim \mathcal{W}(c_0, C_0), \quad C_0 \sim \mathcal{W}(g_0, G_0). \end{aligned}$$

The parameters for the prior on Σ^{-1}, the Wishart distribution, and the prior on μ_i are defined as in Malsiner-Walli et al. (2016).

Figure 1 (a) illustrates the inconsistency of Pitman–Yor mixture models for the number of clusters proved in Miller and Harrison (2014) when the parameters α and σ of PY are fixed. The result proved in this paper stated that PY mixture models are also inconsistent for the number of clusters if σ is fixed and there is a prior on α when $t = 1$, here Figure 1 (b) illustrates that this is also the case for the more realistic case of $t = 3$: we can observe that the number of clusters does not converge around the true number of components $t = 3$ (the case $t = 1$ is illustrated in Appendix B). Figure 1 (c) and (d) illustrate cases beyond the scope of this paper, in which a hyperprior is placed on σ and α is either fixed or random. When α is fixed and there is a hyperprior on σ, the model seems to consistently recover the true number of components. In the second scenario, the simulations appear to show inconsistency for the model. Both cases are not covered by current theoretical results and constitute interesting future research topics.

5. Discussion

We have proved inconsistency for the number of clusters when modeling mixtures consisting of one single component with Pitman–Yor mixture models, with a prior on the concentration parameter α and with the discount parameter σ kept fixed. Our result holds only for the case where the true number of clusters in the data-generating mechanism equals one. While it gives us a good indication of what could be expected, further study would be needed to



(a) Fixed $\alpha = 10$ and $\sigma = 0.5$

(b) $\alpha \sim \text{Gamma}(200, 20)$ and fixed $\sigma = 0.5$

(c) Fixed $\alpha = 10$ and $\sigma \sim \text{Unif}(0, 1)$

(d) $\alpha \sim \text{Gamma}(200, 20)$ and $\sigma \sim \text{Unif}(0, 1)$

Figure 1: Posterior distribution of the number of clusters K_n under a Pitman–Yor process mixture for various choices of n and with (a) fixed parameters α and σ; (b) $\alpha \sim \text{Gamma}(200, 20)$ and fixed σ; (c) fixed α and $\sigma \sim \text{Unif}(0, 1)$; and (d) $\alpha \sim \text{Gamma}(200, 20)$ and $\sigma \sim \text{Unif}(0, 1)$.

fully understand if clustering consistency holds when the number of clusters in the data-generating mechanism may be arbitrary.

While our result is limited to the setting where the discount parameter σ is kept fixed, it is common in practice to put a prior on both PY parameters α and σ in PY mixture models. Simulation study seems to suggest inconsistency for this extension, and this is the subject of current investigations. We are also investigating the extension which consists in keeping α fixed and putting a prior on σ. This extension seems consistent in the simulation study.

References

Louise Alamichel, Daria Bystrova, Julyan Arbel, and Guillaume Kon Kam King. Bayesian mixture models (in)consistency for the number of clusters. arXiv preprint arXiv:2210.14201, 2022.

Charles E Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics, pages 1152–1174, 1974.

Filippo Ascolani, Antonio Lijoi, Giovanni Rebaudo, and Giacomo Zanella. Clustering consistency with Dirichlet process mixtures. arXiv preprint arXiv:2205.12924, 2022.

David M Blei, Michael I Jordan, et al. Variational inference for Dirichlet process mixtures. Bayesian analysis, 1(1):121–144, 2006.

Antonio Canale, Riccardo Corradin, and Bernardo Nipoti. Importance conditional sampling for Pitman–Yor mixtures. Statistics and Computing, 32(3):40, May 2022. ISSN 1573-1375. doi: 10.1007/s11222-022-10096-0.

Michael D Escobar and Mike West. Computing nonparametric hierarchical models. Practical Nonparametric and Semiparametric Bayesian Statistics, pages 1–22, 1998.

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

Subhashis Ghosal and Aad Van der Vaart. Fundamentals of nonparametric Bayesian inference, volume 44. Cambridge University Press, 2017.

Subhashis Ghosal, Jayanta K Ghosh, and RV Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. The Annals of Statistics, 27(1):143–158, 1999.

Aritra Guha, Nhat Ho, and XuanLong Nguyen. On posterior contraction of parameters and interpretability in Bayesian mixture modeling. Bernoulli, 27(4):2159–2188, 2021.

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

Antonio Lijoi, Igor Prünster, and Stephen G Walker. On consistency of nonparametric normal mixtures for Bayesian density estimation. Journal of the American Statistical Association, 100(472):1292–1296, 2005.

Albert Y Lo. On a class of Bayesian nonparametric estimates: I. density estimates. The Annals of Statistics, pages 351–357, 1984.

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

Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter, and Bettina Grün. Model-based clustering based on sparse finite Gaussian mixtures. Statistics and Computing, 26(1-2):303–324, 2016.

Jeffrey W Miller and Matthew 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.

Jeffrey W Miller and Matthew T Harrison. Inconsistency of Pitman-Yor process mixtures for the number of components. The Journal of Machine Learning Research, 15(1):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.

Peter Müller, Fernando A Quintana, and Garritt Page. Nonparametric Bayesian inference in applications. Statistical Methods & Applications, 27:175–206, 2018.

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

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

Akio Onogi, Masanobu Nurimoto, and Mitsuo Morita. Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. BMC Bioinformatics, 12(1):1–16, 2011.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. The Annals of Probability, pages 855–900, 1997.

Judith Rousseau and Kerrie Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(5):689–710, 2011.

Mike West and Michael D Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. Institute of Statistics and Decision Sciences, Duke University, 1993.

Appendix A. Proof of results

The proof of our result relies on the following simple lemma, used by and proved by Ascolani et al. (2022). It justifies working with ratios, which allows us to avoid calculations of marginal likelihoods of the observed data.

Lemma 2 *The convergence $p(K_n = t|X_{1:n}) \to 1$ as $n \to \infty$ holds if and only if one has*

$$\sum_{s \neq t} \frac{p(K_n = s|X_{1:n})}{p(K_n = t|X_{1:n})} \to 0 \quad \text{as } n \to \infty.$$

Proof of Theorem 1. By Lemma 2, it will be sufficient to prove that $\frac{p(K_n=s|X_{1:n})}{p(K_n=1|X_{1:n})} \not\to 0$ as $n \to \infty$, for some $s > 1$. We will prove this using $s = 2$.

In order to prove our result, we make use of results of the asymptotic behavior of certain quantities under the Dirichlet process mixture model of Ascolani et al. (2022). Throughout this proof will thus use the subscript DP to indicate that a quantity is related to the Dirichlet process model, and we will use the subscript PY to indicate that a quantity is related to the Pitman–Yor model, whenever there is ambiguity.

Under our Pitman–Yor mixture model, by applying Equation (4), we have

$$\begin{aligned}\frac{p(K_n = 2|X_{1:n})_{\mathrm{PY}}}{p(K_n = 1|X_{1:n})_{\mathrm{PY}}} &= \frac{\int \sigma\left(1 + \frac{\alpha}{\sigma}\right) \frac{\pi_1(\alpha)}{(1+\alpha)_{(n)}} \mathrm{d}\alpha}{\int \frac{\pi_1(\alpha)}{(1+\alpha)_{(n)}} \mathrm{d}\alpha} \frac{\sum_{A \in \tau_2(n)} \prod_{j=1}^{2} (1-\sigma)_{(a_j)} m(X_{A_j})}{(1-\sigma)_{(n)} m(X_{1:n})} \\ &= C_{\mathrm{PY}}(n,1,2) R_{\mathrm{PY}}(n,1,2)\end{aligned}$$

where

$$C_{\mathrm{PY}}(n,1,2) = \frac{\int \sigma\left(1 + \frac{\alpha}{\sigma}\right) \frac{\pi_1(\alpha)}{(1+\alpha)_{(n)}} \mathrm{d}\alpha}{\int \frac{\pi_1(\alpha)}{(1+\alpha)_{(n)}} \mathrm{d}\alpha}$$

and

$$R_{\mathrm{PY}}(n,1,2) = \frac{\sum_{A \in \tau_2(n)} \prod_{j=1}^{2} (1-\sigma)_{(a_j)} m(X_{A_j})}{(1-\sigma)_{(n)} m(X_{1:n})}.$$

Similarly, under the Dirichlet process mixture model of Ascolani et al. (2022), one gets

$$\frac{p(K_n = s|X_{1:n})_{\mathrm{DP}}}{p(K_n = t|X_{1:n})_{\mathrm{DP}}} = C_{\mathrm{DP}}(n,t,s) R_{\mathrm{DP}}(n,s,t),$$

where $C_{\mathrm{DP}}(n,t,s)$ is an integral in α over all of the terms involving α, and $R_{\mathrm{DP}}(n,t,s)$ contains all of the remaining factors:

$$C_{\mathrm{DP}}(n,t,s) := \frac{\int [\alpha^s \pi_1(\alpha)/\alpha_{(n)}] \mathrm{d}\alpha}{\int [\alpha^t \pi_1(\alpha)/\alpha_{(n)}] \mathrm{d}\alpha}$$

and

$$R_{\mathrm{DP}}(n,t,s) := \frac{\sum_{A \in \tau_s(n)} \prod_{j=1}^{s} (a_j - 1)! \prod_{j=1}^{s} m(X_{A_j})}{\sum_{B \in \tau_t(n)} \prod_{j=1}^{t} (b_j - 1)! \prod_{j=1}^{t} m(X_{B_j})}.$$

Ascolani et al. (2022) prove that

$$C_{\mathrm{DP}}(n,t,s) \to 0 \text{ as } n \to \infty \quad \forall 0 < t < s. \tag{5}$$

Now, since our expression $R_{\mathrm{PY}}(n,1,2)$ above does not depend on α, it is identical to the corresponding expression in the setup of Miller and Harrison (2014), who prove that it does not converge to zero as $n \to \infty$. What is left to show is that our expression $C_{\mathrm{PY}}(n,1,2)$ above does not converge to zero as $n \to \infty$.

We then have,

$$\begin{aligned} C_{\mathrm{PY}}(n,1,2) &= \frac{\int \sigma\left(1+\frac{\alpha}{\sigma}\right)\frac{\pi_1(\alpha)}{(1+\alpha)_{(n)}}\mathrm{d}\alpha}{\int \frac{\pi_1(\alpha)}{(1+\alpha)_{(n)}}\mathrm{d}\alpha} \\ &= \sigma + \frac{\int \alpha\frac{\pi_1(\alpha)}{(1+\alpha)_{(n)}}\mathrm{d}\alpha}{\int \frac{\pi_1(\alpha)}{(1+\alpha)_{(n)}}\mathrm{d}\alpha} = \sigma + \frac{\int \alpha^2\frac{\pi_1(\alpha)}{(\alpha)_{(n+1)}}\mathrm{d}\alpha}{\int \alpha\frac{\pi_1(\alpha)}{(\alpha)_{(n+1)}}\mathrm{d}\alpha} \\ &= \sigma + C_{\mathrm{DP}}(n+1,2,3) \to \sigma \text{ as } n \to \infty, \end{aligned}$$

where the final line above comes from the special case of Equation (5) where $t = 2$ and $s = 3$. ■

Appendix B. Details on the simulation study

To be closer to the results stated in Section 3, we now consider data generated from a mixture model with only one component. It means that the data are generated from the same Gaussian model,

$$P(x) = \mathcal{N}(x|\mu,\Sigma).$$

The different datasets are of size $n \in \{50, 200, 500, 2000\}$ as in Section 4. The following parameters are considered for the mean and the covariance matrix,

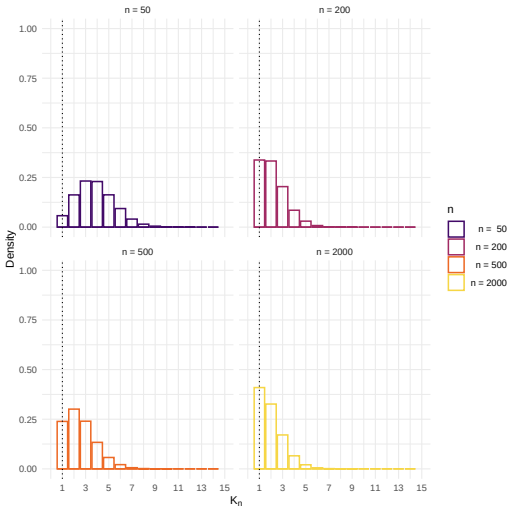
$$\mu = (0.8, 0.8) \quad \text{and} \quad \Sigma = 0.05\, I_2.$$

The model and the algorithms used are similar as in Section 4. We choose the hyperparameters of the priors on σ and α such that the mean of these distributions correspond to the fixed values chosen for α and σ. The prior on α also satisfies the conditions introduced in Ascolani et al. (2022).

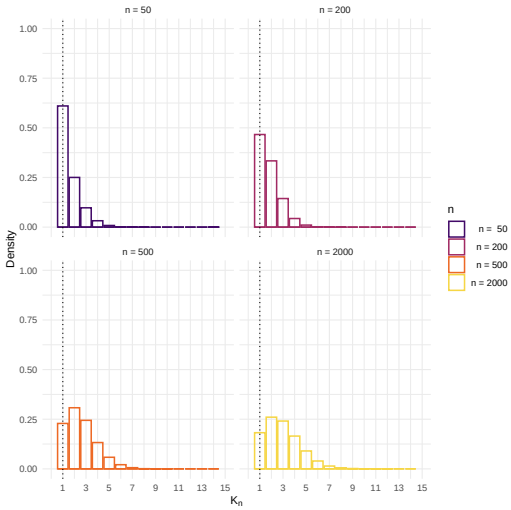
It can be noted that we need a larger number of iterations to achieve MCMC convergence when we put a prior on σ.

As we choose here a unique component the results are less graphic than in Section 4. Still, we can observe that in Figure 2 (b), the model seems to overestimate K_n as the size increase. Hence, Figure 2 (b) illustrates the theoretical result in Section 3.

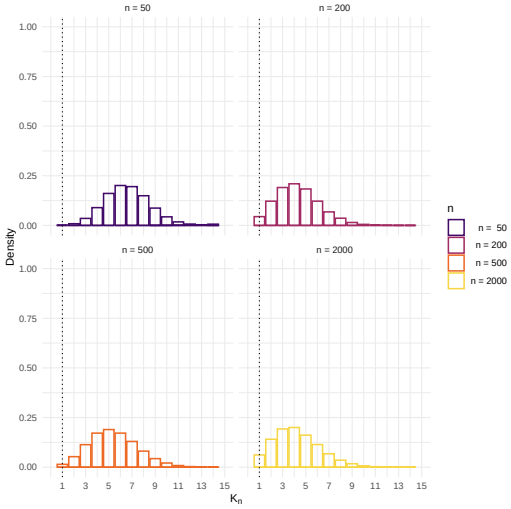
Also, the distribution in Figure 2 (c) seems to slowly concentrate on the left, but concentration of the posterior on a single component is probably slower than on three components and it would be necessary to use simulated data with a larger sample size to obtain more striking figures.



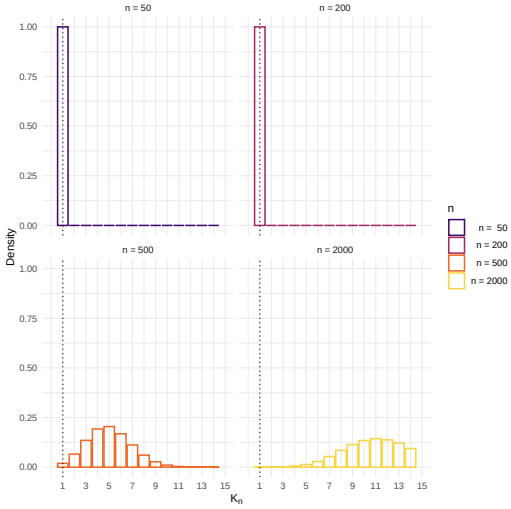
(a) Fixed $\alpha = 10$ and $\sigma = 0.5$



(b) $\alpha \sim \text{Gamma}(200, 20)$ and fixed $\sigma = 0.5$



(c) Fixed $\alpha = 10$ and $\sigma \sim \text{Unif}(0, 1)$



(d) $\alpha \sim \text{Gamma}(200, 20)$ and $\sigma \sim \text{Beta}(0.5, 0.5)$

Figure 2: Posterior distribution of the number of clusters K_n under a Pitman–Yor process mixture for various choices of n and with (a) fixed parameters α and σ; (b) $\alpha \sim \text{Gamma}(200, 20)$ and fixed σ; (c) fixed α and $\sigma \sim \text{Unif}(0, 1)$; and (d) $\alpha \sim \text{Gamma}(200, 20)$ and $\sigma \sim \text{Unif}(0, 1)$.