Bayesian nonparametric mixtures inconsistency for the number of clusters

Louise Alamichel, Julyan Arbel, Daria Bystrova, Guillaume Kon Kam King

Bayesian nonparametric mixtures inconsistency for the number of clusters

Louise Alamichel[1,*], Julyan Arbel[1], Daria Bystrova[1] & Guillaume Kon Kam King[2]

[1] *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France*
{louise.alamichel, julyan.arbel, daria.bystrova}@inria.fr
[2] *Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France*
guillaume.kon-kam-king@inrae.fr

Résumé. Les modèles de mélange bayésiens non paramétriques sont souvent utilisés pour modéliser des données complexes. Si ces modèles sont bien adaptés à l'estimation de densité, leur application au clustering, bien que courante, reste plus discutée. En effet, Miller and Harrison (2014) montrent l'inconsistance a posteriori du nombre de clusters lorsque le nombre réel de clusters est fini pour les modèles de mélange à processus de Dirichlet et à processus de Pitman–Yor. Dans ce travail, nous étendons ce résultat à d'autres priors bayésiens non paramétriques tels que les processus de type Gibbs et la représentation en dimension finie de ces différents priors. Ces représentations finies comprennent le processus multinomial de Dirichlet, de Pitman–Yor et celui de gamma généralisé normalisé, récemment proposés. Plus précisément, nous montrons que les modèles de mélange basés sur tous ces processus sont également inconsistants quant au nombre de clusters.

Mots-clés. Bayésien non-paramétrique, consistance, mélange fini

Abstract. Bayesian nonparametric mixture models are often employed for modelling complex data. While these models are well-suited for density estimation, their application for clustering has some limitations. Miller and Harrison (2014) proved posterior inconsistency in the number of clusters when the true number of clusters is finite for Dirichlet process and Pitman–Yor process mixture models. In this work, we extend this result to additional Bayesian nonparametric priors such as Gibbs-type processes and finite-dimensional representations of them. The latter include the Dirichlet multinomial process and the recently emerged Pitman–Yor and normalized generalized gamma multinomial processes. We show that mixture models based on these processes are also inconsistent in the number of clusters.

Keywords. Bayesian nonparametric, consistency, finite mixture

*This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissements d'avenir

1 Introduction

Finite mixture models are hugely popular in applications, but choosing the appropriate number of components remains challenging. Using Bayesian nonparametric (BNP) priors (such as the Dirichlet process) for mixture modelling allows to avoid this choice, by assuming an infinite number of components, which implies that the number of clusters found in a dataset is flexibly estimated depending on the amount and the structure of the data. While BNP mixture models are well-suited for density estimation, they have some limitations when employed for clustering data known to originate from a finite mixture. Indeed, Miller and Harrison (2014) states that when using a Dirichlet process or Pitman–Yor process mixture the posterior will not be consistent for the number of clusters if the data is generated according to a finite mixture. Here we show that this type of inconsistency result generalizes to Gibbs-type process mixtures and finite-dimensional representations of BNP priors.

We start by introducing the notion of a partition-based mixture model, before stating the inconsistency results of Miller and Harrison (2014) on Dirichlet process mixtures and Pitman–Yor process mixtures. We then present our generalization of this result to Gibbs-type processes and to several finite dimensional representations of BNP processes. For clarity, we omit the proofs of our results. However, the proofs are presented in article in preparation, which is a part of PhD project of Louise Alamichel.

1.1 Partition-based mixture model

We consider partition-based mixture model as defined in Miller and Harrison (2014). Firstly, we define the distribution on partitions. Let $\mathcal{A}_k(n)$ be the set of ordered partitions of $\{1, \dots, n\}$ into k ($k \in \{1, \dots, n\}$) nonempty sets:

$$\mathcal{A}_k(n) := \left\{ (A_1, \dots, A_k) : A_1, \dots, A_k \text{ disjoint}, \bigcup_{i=1}^{k} A_i = \{1, \dots, n\}, \ |A_i| \geq 1 \ \forall i \right\}.$$

We denote $n_i := |A_i|$. Then, we consider a distribution $p(A)$ on $\bigcup_{k=1}^{n} \mathcal{A}_k(n)$, which induces a distribution on k as $p(k \mid A) = \mathbb{I}(A \in \mathcal{A}_k(n))$ where $\mathbb{I}$ is the indicator function.

Next we introduce the hierarchical mixture model. We denote π a prior density on the d-dimensional parameters $\theta \in \Theta \subset \mathbb{R}^d$ and p_θ a parametrized component density. The hierarchical structure of *partition-based mixture model* is then:

$$p(A, k) = p(A)\mathbb{I}(A \in \mathcal{A}_k(n)), \quad p(\theta_{1:k}|A, k) = \prod_{i=1}^{k} \pi(\theta_i), \quad p(x_{1:n}|A, k, \theta_{1:k}) = \prod_{i=1}^{k} \prod_{j \in A_i} p_{\theta_i}(x_j),$$

where $x_{1:n} = (x_1, \dots, x_n)$ with $x_i \in \mathcal{X}$, $\theta_{1:k} = (\theta_1, \dots, \theta_k)$ with $\theta_i \in \Theta$, and $A \in \mathcal{A}_k(n)$. In

particular, it is a Dirichlet process mixture model when

$$p(A) = \frac{\alpha^k}{k!(\alpha)_n} \prod_{i=1}^{k} (n_i - 1)!$$

for $A \in \mathcal{A}_k(n)$, where $\alpha > 0$ and $(x)_n = x(x+1)\cdots(x+n-1)$, with $(x)_0 = 1$ by convention.

Finally, we denote by K_n the number of clusters.

1.2 Inconsistency theorem

The central result of Miller and Harrison (2014, Theorem 6) is reproduced below as Theorem 1.1. This result depends on two conditions which are given thereafter.

Firstly, we introduce some notations for Condition 1. For $A \in \mathcal{A}_k(n)$, we define $R_A = \bigcup_{i:|A_i|>2} A_i$, the union of all clusters except singletons. For $j \in R_A$, we define $B(A, j)$ to be the ordered partition $B \in \mathcal{A}_{k+1}(n)$ obtained by removing j from its cluster and creating a new singleton for it. Then $B_i = A_i \setminus \{j\}$, $i = 1, \ldots, k$, and $B_{k+1} = \{j\}$. Let $\mathcal{Z}_A := \{B(A, j) : j \in R_A\}$, for $n > k \geq 1$, we define

$$c_n(k) := \frac{1}{n} \max_{A \in \mathcal{A}_k(n)} \max_{B \in \mathcal{Z}_A} \frac{p(A)}{p(B)},$$

with the convention that $0/0 = 0$ and $y/0 = \infty$ for $y > 0$.

Condition 1. *Assume* $\limsup_{n\to\infty} c_n(k) < \infty$, *given some particular* $k \in \{1, 2, \ldots\}$.

The second condition induces a control on the likelihood through the control of single-cluster marginals. A single-cluster marginal of the model is $m(x_{A_i}) = \int_\Theta \left(\prod_{j \in A_i} p_\theta(x_j)\right) \pi(\theta) \mathrm{d}\theta$. Given $c \in [0, \infty)$, we introduce

$$\varphi_k(x_{1:n}, c) := \min_{A \in \mathcal{A}_k(n)} \frac{1}{n} \left| S_A(x_{1:n}, c) \right|,$$

where $S_A(x_{1:n}, c)$ is the set of indices $j \in \{1, \ldots, n\}$ such that the part A_ℓ containing j satisfies $m(x_{A_\ell}) \leq c m(x_{A_\ell \setminus j}) m(x_j)$, i.e. the set of observations for which the marginals of the new clusters obtained after taking out that observation and creating a new singleton cluster dominates the marginal of the original cluster up to a constant c.

Condition 2. *Given a sequence of random variables* $X_1, X_2, \ldots \in \mathcal{X}$, *and* $k \geq 1$, *assume* $\sup_{c \in [0,\infty)} \liminf_{n\to\infty} \varphi_k(X_{1:n}, c) > 0$ *a.s.*

For $c = 1$, Condition 2 means that as $n \to \infty$, there is always a non-vanishing proportion of the observations for which creating a singleton cluster will increase its cluster marginal.

Theorem 1.1 (Miller and Harrison, 2014). *Let $X_1, X_2, \ldots \in \mathcal{X}$ be a sequence of r.v., and consider a partition-based model. Then, if Conditions 1 and 2 hold for any $k \geq 1$, we have*

$$\limsup_{n\to\infty} p(K_n = k|X_{1:n}) < 1 \quad \textit{with probability } 1.$$

The first condition is only related to partition distribution, while the second condition only involves the data-distribution and single cluster marginals. Hence, to generalize this inconsistency result to other processes, it is enough to show that Condition 1 also holds for these different processes.

2 Gibbs-type process

Gibbs-type processes are a natural generalization of the Dirichlet and Pitman–Yor processes. Gibbs-type processes of order: $\sigma, \sigma < 1$, can be characterized through the probability distribution of the induced random partition, which has the following form:

$$p(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}, \tag{1}$$

where $V_{n,k}$ are nonnegative numbers that satisfy the recurrence relation

$$V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}, \quad V_{1,1} = 1. \tag{2}$$

In this case, we have a probability distribution for ordered partition $A \in \mathcal{A}_k(n)$ which can be deduced from (1) by dividing on $k!$ to adjust for order: $p(A) = \frac{V_{n,k}}{k!} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}$.

Proposition 2.1. *Consider a Gibbs-type process with $0 < \sigma < 1$, then Condition 1 holds for any $k \in \{1, 2, \ldots\}$, so does inconsistency of Theorem 1.1.*

Idea of proof. For $B = B(A, j)$ as defined in section 1.2, we want to prove that

$$c_n(k) = \frac{1}{n} \max_{A\in\mathcal{A}_k(n)} \max_{B\in\mathcal{Z}_A} \frac{p(A)}{p(B)} < \infty$$

As $\frac{1}{n}\frac{p(A)}{p(B)} \leq \frac{V_{n,k}}{V_{n,k+1}}(k+1)$, we just have to prove that the sequence $\left(\frac{V_{n,k}}{V_{n,k+1}}\right)_{n\geq 0}$ is bounded. Coefficients $V_{n,k}$ can be equivalently defined through the density f_σ of a positive stable random variable

$$V_{n,k} = \frac{\sigma^k}{\Gamma(n-k\sigma)} \int_0^{+\infty} \int_0^1 t^{-k\sigma} p^{n-k\sigma-1} h(t) f_\sigma((1-p)t) \mathrm{d}t\mathrm{d}p.$$

Applying the Laplace approximation method as in Arbel and Favaro (2021), we obtain that the sequence $\left(\frac{V_{n,k}}{V_{n,k+1}}\right)_{n\geq 0}$ is bounded and Condition 1 is satisfied. □

3 Finite dimensional representation

The recent works by Lijoi et al. (2020a,b) develop finite-dimensional versions of the Pitman–Yor process and normalized random measures with independent increments (NRMI). The latter includes the Dirichlet and normalized generalized gamma multinomial processes as special cases. These processes can be seen as approximations for the corresponding infinite dimensional processes.

3.1 Pitman–Yor multinomial process

The Pitman–Yor multinomial process is defined in Lijoi et al. (2020b) as a discrete random probability measure $\tilde{p}_H$ such that $\tilde{p}_H|\tilde{p}_{0,H} \sim \mathrm{PY}(\sigma, \alpha; \tilde{p}_{0,H})$, $\tilde{p}_{0,H} = \frac{1}{H}\sum_{h=1}^{H} \delta_{\tilde{\theta}_h}$ where $H \geq 1$, $\tilde{\theta}_h \overset{\text{iid}}{\sim} P$, $\sigma \in [0, 1)$ and $\alpha > -\sigma$. The distribution on ordered partitions for the Pitman–Yor multinomial process is as follows

$$p(A) = \frac{H!}{(H-k)!k!} \frac{1}{(\alpha+1)_{n-1}} \sum_{(\ell_1,\ldots,\ell_k)} \frac{\Gamma(\alpha/\sigma + |\ell^{(k)}|)}{\sigma\Gamma(\alpha/\sigma+1)} \prod_{i=1}^{k} \frac{C(n_i, \ell_i; \sigma)}{H^{\ell_i}},$$

where the sum runs over the vectors $\ell^{(k)} = (\ell_1, \ldots, \ell_k)$ such that $\ell_i \in \{1, \ldots, n_i\}$ and $|\ell^{(k)}| = \ell_1 + \cdots + \ell_k$. The $C(n, k; \sigma)$ are the generalised factorial coefficients defined as

$$C(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n. \tag{3}$$

Proposition 3.1. *Consider a Pitman–Yor multinomial process, then Condition 1 holds for any $k < \min(n, H)$, so does the inconsistency of Theorem 1.1.*

3.2 Normalized infinitely divisible multinomial processes

Normalized infinitely divisible multinomial (NIDM) processes were introduced in Lijoi et al. (2020a). NIDM processes can be described through NRMI measures using following hierarchical structure: $(\tilde{p}_H \mid \tilde{p}_{0,H}) \sim \mathrm{NRMI}(c, \rho; \tilde{p}_{0,H})$, $\quad \tilde{p}_{0,H} = \frac{1}{H}\sum_{h=1}^{H} \delta_{\tilde{\theta}_h}$.

If the NRMI process in the definition of the NIDM process is a Dirichlet process, then we can we obtain the Dirichlet multinomial process and the distribution on the ordered partitions is defined as:

$$p(A) = \Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!k!} \frac{\prod_{j=1}^{k} (c/H)_{n_j}}{(c)_n}$$

Proposition 3.2. *Consider a Dirichlet multinomial process, then Condition 1 holds for any $k < \min(n, H)$, so does inconsistency of Theorem 1.1.*

Similarly, when we consider normalized generalized gamma as NRMI process in the definition, we get NGG multinomial process.

In this case the probability on the ordered partition is slightly more involved: We have with $k \leq \min(n, H)$ (see Example 2 Lijoi et al., 2020a):

$$P(A) = \frac{H!}{(H-k)!k!} \sum_{(\ell_1,\ldots,\ell_k)} \frac{V_{n,|\ell^{(k)}|}}{H^{|\ell^{(k)}|}} \prod_{i=1}^{k} \frac{C(n_i, \ell_i; \sigma)}{\sigma^{\ell_i}},$$

Here, the $C(n, k; \sigma)$ are again the generalized factorial coefficients defined in (3). The $V_{n,k}$ are the parameters defined in (2) for the particular case of NGG processes.

Proposition 3.3. *Consider a normalized generalized gamma multinomial process, then Condition 1 holds for any $k < \min(n, H)$, so does inconsistency of Theorem 1.1.*

4 Discussion

We have proved that Gibbs-type process mixtures are inconsistent a posteriori for the number of clusters, when the true number of components is finite. It is also the case, for some finite-dimensional representations of BNP priors as the Dirichlet multinomial process. However, we did not prove inconsistency in general for NIDM. A recent proposal by Guha et al. (2021) shows that this inconsistency problem can be resolved for Dirichlet process mixtures using a simple post-processing procedure based on merging neighbouring clusters. This procedure is effective as long as the posterior contracts at a known rate under the Wasserstein metric to the true mixing measure, which raises hope concerning its applicability to a larger class of processes than the Dirichlet process mixture.

References

Arbel, J. and Favaro, S. (2021). "Approximating predictive probabilities of Gibbs-type priors." *Sankhya A*, 83(1), 496–519.

Guha, A., Ho, N., and Nguyen, X. (2021). "On posterior contraction of parameters and interpretability in Bayesian mixture modeling." *Bernoulli*, 27(4), 2159–2188.

Lijoi, A., Prünster, I., and Rigon, T. (2020a). "Finite-dimensional discrete random structures and Bayesian clustering." *Preprint*.

— (2020b). "The Pitman–Yor multinomial process for mixture modelling." *Biometrika*, 107(4), 891–906.

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