POSTERIOR CLUSTERING FOR DIRICHLET PROCESS MIXTURES OF GAUSSIANS WITH CONSTANT DATA

Filippo Ascolani[1] and Valentina Ghidini[2]

[1] Bocconi University, Milan (filippo.ascolani@phd.unibocconi.it)

[2] Bocconi University, Milan (valentina.ghidini@phd.unibocconi.it)

ABSTRACT: Dirichlet process mixtures, obtained by convolving the law of a Dirichlet process with a suitable kernel, are popular methods for density estimation. Due to the almost sure discreteness of the mixing measure, they automatically provide a latent clustering which is often of great interest for applied researchers. However, despite its relevance, little is known about the posterior properties of clustering, even with a large sample. We contribute by considering a simple data generating mechanism and showing the asymptotic properties of the maximum a posteriori clustering with Gaussian kernel.

KEYWORDS: Bayesian nonparametrics; clustering; maximum a posteriori; asymptotic analysis.

1 Introduction

Bayesian nonparametric methodologies have witnessed a growing popularity in the last decades, mainly due to the their flexibility: see Ghosal & Van Der Vaart (2017) for a recent review. A popular model for density estimation is given by Dirichlet process mixtures (Lo (1984)), which can be summarized as follows

$$Y_i \mid \theta_i \sim k(y \mid \theta_i), \quad \theta_i \mid P \overset{\text{i.i.d.}}{\sim} P, \quad P \sim DP(P_0, \alpha), \tag{1}$$

where $k(y \mid \theta)$ is a density function with parameter θ and $DP(P_0, \theta)$ is the law of a Dirichlet process (DP, Ferguson (1973)) with baseline distribution P_0 and concentration parameter $\alpha > 0$. It can be shown that the realizations of P are almost surely discrete probability measures, so that the θ_i's will present ties with positive probability, leading to a latent clustering of the observed datapoints $Y_{1:n} = (Y_1, \dots, Y_n)$.

Models as in (1) are provided with good asymptotic properties in terms of density estimation (Ghosal & Van der Vaart, 2007), when the data are generated i.i.d. from a "true" distribution P^*, but the clustering behavior a posteriori is less understood. As a positive note, it has been shown that, under suitable assumptions, the posterior on the mixing measure converges to the

"true" one in Wasserstein distance (Nguyen, 2013), but the metric is too weak to prove *per se* results on the clustering. More recently, Miller & Harrison (2013, 2014) showed that the posterior distribution on the number of clusters is often inconsistent, in the sense that it places positive mass to a larger number of clusters, even asymptotically. However, such results are not as bad as they sound: indeed, Ascolani *et al.* (2023) suggested that the issue is alleviated by placing a suitable hyperprior on the concentration parameter α, while Wade (2023) empirically showed that different estimators for the partition (obtained by minimizing different loss functions) lead to considerably different estimates for the number of clusters. Beyond this framework, Rajkowski (2019) proved interesting geometric properties of the maximum a posteriori partition.

In this work we consider a Gaussian kernel for model (1) and a purposely simple data generating mechanism, so that computation of posterior quantities becomes easier. We show that in this context the maximum a posteriori clustering converges to the "natural" partition of the observations.

2 Dirichlet process mixtures with Gaussian kernel

As discussed in Section 1, by the discreteness of the DP the set $(\theta_1, \dots, \theta_n)$, corresponding to observations $Y_{1:n}$, yields ties with positive probability. Therefore model (1) induces a distribution over the space of partitions of $[n] = \{1, \dots, n\}$. If $A = \{A_1, \dots, A_s\} \in \tau_s(n)$, where $\tau_s(n)$ is the space of partitions of $[n]$ in s non-empty and disjoint subsets, it is possible to show (Miller & Harrison, 2013; Ascolani *et al.* , 2023) that

$$\mathbb{P}(A \mid Y_{1:n}) \propto \alpha^s \prod_{j=1}^{s} \Gamma(a_j) \prod_{j=1}^{s} m\left(Y_{A_j}\right), \tag{2}$$

where $a_j = |A_j|$, $Y_{A_j} = \{Y_i \mid i \in A_j\}$ and $m\left(Y_{A_j}\right) = \int \prod_{i \in A_j} k(Y_i \mid \theta) P_0(\mathrm{d}\theta)$ denotes the marginal distribution of cluster j. We call the *maximum a posteriori clustering*, the partition $A^*(Y_{1:n})$ which maximizes the above posterior distribution, i.e. $A^*(Y_{1:n}) = \operatorname{argmax}_A \mathbb{P}(A \mid Y_{1:n})$. In this work we assume to observe scalar data points and

$$k(y \mid \theta) = N(\theta, \sigma^2) \quad \text{and} \quad P_0(\mathrm{d}\theta) = N(\mu_0, \sigma_0^2)\mathrm{d}\theta, \tag{3}$$

where $N(\mu, \tau^2)$ denotes the density of a normal distribution with mean μ and variance τ^2, while $(\mu_0, \sigma_0^2, \sigma^2)$ are fixed hyperparameters. With standard com-

putations it is easy to obtain

$$m\left(Y_{A_j}\right) = \sqrt{\frac{\sigma^2}{\sigma_0^2 a_j + \sigma^2}} (2\pi\sigma^2)^{-a_j} e^{\frac{\sigma_0^2 a_j + \sigma^2}{2\sigma^2\sigma_0^2}\left(\frac{\sigma_0^2 a_j}{\sigma_0^2 a_j + \sigma^2}\frac{1}{a_j}\sum_{i\in A_j} Y_i + \frac{\sigma^2}{\sigma_0^2 a_j + \sigma^2}\mu_0\right)^2}. \quad (4)$$

3 Data generating mechanism and main result

As it is commonly done in asymptotic analysis, we assume that the observed datapoints are not generated according to model (1), but rather are independent and identically distributed from a "true" distribution P^*. In the following we assume $P^*(\mathrm{d}y) = \delta_{c^*}(\mathrm{d}y)$, that is all the observations are equal to a fixed real value c^*. This is a stylized setting, where we expect the partition generated by model (1) to converge to $[n]$, i.e. all observations clustered together. However, Theorem 4.1 in Miller & Harrison (2013) implies that the posterior on the number of clusters does not converge to 1 as $n \to \infty$. Notice that Theorem 3 in Ascolani *et al.* (2023) shows instead that consistency holds with a prior on the concentration parameter α. In the following theorem we prove that, even with α fixed, the maximum a posteriori clustering converges to $[n]$, as expected.

Theorem 1. *Consider model (1) with kernel as in (3). Let $Y_i \overset{i.i.d.}{\sim} \delta_{c^*}$, with $i = 1,\dots,n$. Then, for every $(\mu_0, \sigma_0^2, \sigma^2)$ there exists N such that for every $n \geq N$ it holds $A^*(Y_{1:n}) = [n]$.*

Proof. Fix a triplet $(\mu_0, \sigma_0^2, \sigma^2)$. The statement is proved by showing that there exists N such that for every $n \geq N$ it holds

$$\sup_{2\leq s\leq n} \sup_{A\in\tau_s(n)} \frac{\mathbb{P}(A \mid Y_{1:n})}{\mathbb{P}([n] \mid Y_{1:n})} < 1.$$

By (4) it is easy to show that there exists a constant $K > 0$, which does not depend on s and n, such that $\alpha^{1-s}\prod_{j=1}^{s} m\left(Y_{A_j}\right)/m(Y_{1:n}) \leq e^{Ks}$ for every $A \in \tau_s(n)$. Therefore, by (2) we can give the following bound

$$\sup_{2\leq s\leq n} \sup_{A\in\tau_s(n)} \frac{\mathbb{P}(A \mid Y_{1:n})}{\mathbb{P}([n] \mid Y_{1:n})} \leq \sup_{2\leq s\leq n} \sup_{a\in\sigma_s(n)} e^{Ks} \frac{\prod_{j=1}^{s}\Gamma(a_j)}{\Gamma(n)},$$

where $\sigma_s(n) = \left\{a \in \{1,\dots,n\}^s \mid \sum_{j=1}^{s} a_j = n\right\}$. Moreover, it is not difficult to show that $\sup_{a\in\sigma_s(n)} \prod_{j=1}^{s}\Gamma(a_j) = \Gamma(n-s+1)$, which implies

$$\sup_{2\leq s\leq n} \sup_{A\in\tau_s(n)} \frac{\mathbb{P}(A \mid Y_{1:n})}{\mathbb{P}([n] \mid Y_{1:n})} \leq \sup_{2\leq s\leq n} e^{Ks}\frac{\Gamma(n-s+1)}{\Gamma(n)} =: \sup_{2\leq s\leq n} f(s).$$

Notice that $f(s+1) > f(s)$ if and only if $s > n - e^K$, so that $f(s)$ attains its maximum either at 2 or n. Therefore we conclude

$$\sup_{2\leq s\leq n} \sup_{A\in\tau_s(n)} \frac{\mathbb{P}(A \mid Y_{1:n})}{\mathbb{P}([n] \mid Y_{1:n})} \leq f(2) + f(n) = \frac{e^{2K}}{n-1} + \frac{e^{Kn}}{(n-1)!} \to 0$$

as $n \to \infty$, as desired. □

4 Discussion

We showed that, with constant data, despite inconsistency for the number of clusters, the maximum a posteriori clustering converges to the "true" partition. It would be of great interest to extend this result beyond such simple data generating mechanism, even if the identification of a "true" clustering becomes less clear: see Section 3 of Rajkowski (2019) for some examples. This will be object of future work.

References

ASCOLANI, F., LIJOI, A., REBAUDO, G., & ZANELLA, G. 2023. Clustering consistency with Dirichlet process mixtures. *Biometrika*, **forthcoming**.

FERGUSON, T. S. 1973. A Bayesian analysis of some nonparametric problems. *Ann. Stat.*, **1**, 209–230.

GHOSAL, S., & VAN DER VAART, A. W. 2007. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Stat.*, **35**, 697–723.

GHOSAL, S., & VAN DER VAART, A. W. 2017. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

LO, A. Y. 1984. On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.*, **12**, 351–357.

MILLER, J. W., & HARRISON, M. T. 2013. A simple example of Dirichlet process mixture inconsistency for the number of components. *Adv. Neural Inf. Process. Syst.*, **26**, 199–206.

MILLER, J. W., & HARRISON, M. T. 2014. Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.*, **15**, 3333–3370.

NGUYEN, X. 2013. Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Stat.*, **41**, 370–400.

RAJKOWSKI, L. 2019. Analysis of the Maximal a Posteriori Partition in the Gaussian Dirichlet Process Mixture Model. *Bayesian Anal.*, **14**.

WADE, S. 2023. Bayesian cluster analysis. *Phil. Trans. R. Soc. A*, **381**.