Bernoulli **29**(2), 2023, 1195–1218
https://doi.org/10.3150/22-BEJ1495

Optimal Bayesian estimation of Gaussian mixtures with growing number of components

ILSANG OHN[1,a] and LIZHEN LIN[2,b]

[1]*Department of Statistics, Inha University, Incheon, Republic of Korea,* [a]*ilsang.ohn@inha.ac.kr*
[2]*Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, Indiana, USA,* [b]*lizhen.lin@nd.edu*

We study Bayesian estimation of finite mixture models in a general setup where the number of components is unknown and allowed to grow with the sample size. An assumption on growing number of components is a natural one as the degree of heterogeneity present in the sample can grow and new components can arise as sample size increases, allowing full flexibility in modeling the complexity of data. This however will lead to a high-dimensional model which poses great challenges for estimation. We novelly employ the idea of a sample size dependent prior in a Bayesian model and establish a number of important theoretical results. We first show that under mild conditions on the prior, the posterior distribution concentrates around the true mixing distribution at a near optimal rate with respect to the Wasserstein distance. Under a separation condition on the true mixing distribution, we further show that a better and adaptive convergence rate can be achieved, and the number of components can be consistently estimated. Furthermore, we derive optimal convergence rates for the higher-order mixture models where the number of components diverges arbitrarily fast. In addition, we suggest a simple recipe for using Dirichlet process (DP) mixture prior for estimating the finite mixture models and provide theoretical guarantees. In particular, we provide a novel solution for adopting the number of clusters in a DP mixture model as an estimate of the number of components in a finite mixture model. Simulation study and real data applications are carried out demonstrating the utilities of our method.

Keywords: Gaussian mixtures; finite mixture models; growing number of components; mixing distribution estimation; posterior contraction rates; Dirichlet processes

1. Introduction

Finite mixture models are powerful tools for modeling heterogeneous data, which have been used in a wide range of applications in statistics and machine learning including density estimation [26], clustering [12], document modeling [4], image generation [40] and designing generative adversarial networks [9], just to name a few. To date, a large number of methods, both frequentist and Bayesian, have been proposed in the literature for various estimation problems related to finite mixture models. Rather than listing a large body of related work here, we refer the readers to the book [13] and a review paper [29] for recent advances on finite mixture modeling. Our work focuses on the estimation of the finite mixture itself, i.e., estimating the parameters of a mixture model such as the mixing distribution and the number of mixing components, from a Bayesian perspective. Although a number of important Bayesian methods have dealt with the problem of finite mixture estimation, many interesting questions remain open. Most of the Bayesian work in the literature assume the number of components is either known or fixed. The minimax optimal convergence rate for estimating the mixing distribution has not been achieved by Bayesian methods even for the fixed set up. Further, posterior consistency on the number of components has not been established except for some special cases. This paper aims to bridge these gaps through establishing a number of new theoretical results under the general framework of finite mixture modeling with growing number of components. Allowing the number of components $k^\star$ to grow is a natural assumption and even required in many situations, for instance, in topic modelling [3]

and computer vision [18], where we expect that when the size of the sample grows so will the degree of heterogeneity present in the sample.

To have a better understanding of some of the theoretical gaps, it is important to review some of the major developments in the literature. A pioneering work on characterizing convergence rates for mixing distribution estimation in finite mixture models is due to Chen [7] which established a *point-wise* convergence rate $C_{\nu^\star} n^{-1/4}$ for estimating the mixing distribution under the L_1 distance, where n denotes the sample size and the $C_{\nu^\star}$ is a constant depending on the true mixing distribution $\nu^\star$.[1] This convergence result holds for the so-called *strongly identifiable* mixtures which include the Gaussian location mixtures as special cases, and so do those stated below. Nguyen [36] and Scricciolo [45] derived the $n^{-1/4}$ point-wise posterior contraction rate under the second-order Wasserstein distance. Ho and Nguyen [21] proved that the maximum likelihood estimator can also achieve this point-wise rate. Under the first-order Wasserstein distance, a better point-wise convergence rate $C_{\nu^\star} n^{-1/2}$ can be obtained. Heinrich and Kahn [20], Ho, Nguyen and Ritov [22] and Guha, Ho and Nguyen [19] established the $n^{-1/2}$ point-wise rate for the minimum Kolmogorov distance estimator, minimum Hellinger distance estimator and Bayesian procedure with the mixture of finite mixtures (MFM) prior, respectively. On the other hand, for the continuous mixtures where the mixing distribution admits a density function, Martin [27] derived a near $n^{-1/2}$ point-wise rate of the mixing density estimation for their predictive recursion algorithm [35,48].

However, due to a lack of uniformity in the constant $C_{\nu^\star}$, their analysis has been restricted to the fixed truth setup, with the number of components assumed to be either known or fixed. Also note that these point-wise rates are not upper bounds of the actual minimax optimal rates of mixing distribution estimation, which were later derived by Heinrich and Kahn [20]. It was shown that the minimax optimal convergence rate of mixing distribution estimation for strongly identifiable mixtures, is of order $n^{-1/(4(k^\star - k_0)+2)}$, where $k^\star$ and k_0 denote the *total* number of components and the number of *well-separated* components of the true mixing distribution, respectively. In other words, the minimax rate deteriorates with the factor $k^\star - k_0$ which can be viewed as the degree of overspecification. Heinrich and Kahn [20] also proposed a minimax optimal minimum Kolmogorov distance estimator which however can be computationally expensive. More recently, Wu and Yang [49] proposed a computationally tractable estimator called the denoised method of moments estimator for Gaussian mixture models, and showed that this estimator achieves the minimax rate. However, these minimax optimal estimators require the knowledge of the number of components $k^\star$, which is not practical. On the other hand, no Bayesian procedure has yet been able to yield a minimax optimal rate.

In general, one does not have the prior knowledge on the number of components, and selecting an appropriate value of the number of components is a crucial step in providing accurate estimates of the true mixing distribution. With too many components, one may suffer from large variances whereas too few components may lead to biased estimators. Also estimating the number of components may be of interest itself in practice especially when each component has a physical interpretation. A widely used approach to choose the number of components is based on a model selection criterion before estimating parameters, and a few consistent model selection criteria are available in the literature such as complete likelihood [2], the Bayesian information criteria (BIC) [25], the singular Bayesian information criteria (sBIC) [8] and the Bayes factor [6].

A Bayesian approach is an attractive alternative due to its ability to estimate both the number of components and parameters in a unified manner. A natural strategy to infer a mixture model with an unknown number of components is to also impose a prior on the number of components k. By doing so, it provides a way of not only choosing the *best* number of components (i.e., model selection), but also

[1] Chen [7] did not realize that the multiplicative constant $C_{\nu^\star}$ depends on the true mixing distribution, thus they argued that the rate $n^{-1/4}$ is the minimax optimal rate. This mistake was later corrected by Heinrich and Kahn [20].

combining results from different mixture models with possibly varying number of components (i.e., model averaging). One notable disadvantage for such models is that posterior computations may be challenging, since it requires developing Monte Carlo Markov chain (MCMC) algorithms for sampling from a parameter space of varying dimensions, which often results in poor mixing or slow convergence of the Markov chain to the stationary distribution. Several MCMC methods have been proposed to circumvent this issue including [32,37,41,47]. On the theoretical side, Guha, Ho and Nguyen [19] derived the $n^{-1/2}$ point-wise posterior contraction rate for this type of prior distribution. They also obtained posterior consistency of the *fixed* number of components under the strong identifiability condition. Another promising approach is to use over-fitted mixtures. This approach considers a mixture model with the number of components larger than the true one and estimates the true model by discarding spurious components. Rousseau and Mengersen [43] studied asymptotic properties of the over-fitted mixtures and proved with a prior on weights of a mixture using a Dirichlet distribution with a suitably selected hyperparameter, the spurious components vanishes asymptotically at the rate $n^{-1/2}\log^a n$ for some $a > 0$ under the posterior distribution.

Our work considers a Bayesian procedure which imposes appropriate priors on both the number of components and the mixing parameters in a general setup where the number of the mixing components is allowed to grow with sample size. We consider a general class of priors and provide assumptions on the prior on the number of components, the mixing weights as well as the atoms of the mixing distribution, that lead to optimal convergence of the posterior. Our work contributes in both methodological and theoretical development, and obtains collection of important results which can be summarized in the following.

1. We design sample size dependent priors and provide mild and explicit conditions on them, based on which near-optimal posterior contraction rate of the mixing distribution estimation is derived with respect to the Wasserstein distance (Theorem 2.2). Under a separation condition on the mixing components, we further show that a better and adaptive optimal posterior contraction rate can be obtained (Theorem 2.3 and Corollary 2.5). To our knowledge, this is the first minimax optimality result in the Bayesian literature.
2. We derive the posterior consistency of the number of components even when the true number of components diverges (Theorem 2.6). To the best of our knowledge, this is the first result on the posterior consistency of the number of components in a general setup where the true mixing distribution varies as the sample size grows.
3. We propose an optimal Bayesian procedure for estimating higher-order mixture models in which the number of components diverges arbitrarily fast (Theorem 2.7).
4. We extend our analysis to general mixture models beyond Gaussian location mixtures with growing number of components. We show that the proposed Bayesian procedure maintain the same theoretical properties even in this setup.
5. We investigate some theoretical properties of the Dirichlet process (DP) mixture models and provide a pathway for using DP models for inference of the finite mixture models (Section 3). The DP prior for the mixing distribution, which only generates infinite mixtures, cannot provide a meaningful posterior distribution for the number of components. We suggest a recipe for using the posterior distribution of *the number of clusters* as the estimate of the true number of components and provide theoretical guarantees. (Theorem 3.1). For mixing distribution estimation, the performance of the DP is inferior in view of the convergence rate (Theorem 3.2).

The rest of this paper is organized as follows. In Section 2, we introduce the notation, finite Gaussian location mixture models, and the prior distribution. Then we present the main results of the paper, including optimal posterior contraction rates of the mixing distribution, and posterior consistency of the number of components. Moreover, we study theoretical properties of the proposed Bayesian procedures

for estimating general mixture models. In Section 3, we analyze theoretical properties of DP mixture models for estimating the finite Gaussian mixtures. In Section 4, numerical studies including both simulation study and real data analysis are conducted for illustrating our theory. Proofs are deferred to Appendix A. In Appendix B, we provide another theoretical analysis for general mixture models with a fixed number of components under different conditions and proof techniques.

2. Main results

2.1. Notations

We first introduce some notation that will be used throughout the paper. We denote by $\mathbb{1}(\cdot)$ the indicator function. For a positive integer $n \in \mathbb{N}$, we let $[n] := \{1, 2, \ldots, n\}$. For a d-dimensional real vector $\mathbf{x} := (x_1, \ldots, x_d) \in \mathbb{R}^d$, we denote $\|\mathbf{x}\|_0 := \sum_{j=1}^d \mathbb{1}(x_j \neq 0)$ and $\|\mathbf{x}\|_\infty := \max_{1 \leq j \leq d} |x_j|$. For two positive sequences $\{a_n\}_{n \in \mathbb{N}}$ and $\{b_n\}_{n \in \mathbb{N}}$, we write $a_n \lesssim b_n$ if there exists a positive constant $C > 0$ such that $a_n \leq C b_n$ for any $n \in \mathbb{N}$. Moreover, we write $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For n random variables $X_1, \ldots, X_n$, we use the shorthand notation $X_{1:n} := (X_1, \ldots, X_n)$. Let δ_θ denote a Dirac measure at θ.

Let $(\mathfrak{X}, \mathcal{X})$ be a measurable space equipped with a Lebesgue measure λ. Let $\mathcal{P}(\mathfrak{X})$ be the set of all distributions supported on $\mathfrak{X}$. For $G \in \mathcal{P}(\mathfrak{X})$, let P_G denote the probability or the expectation under the probability measure G. We denote by p_G the probability density function of G with respect to the Lebesgue measure λ if it exists. For $n \in \mathbb{N}$, let $\mathsf{P}_G^{(n)}$ be the probability or the expectation under the product measure and $p_G^{(n)}$ its density function. For two probability densities p_1 and p_2, we denote by $\mathsf{KL}(p_1, p_2)$ the Kullback-Leibler (KL) divergence from p_2 to p_1 and by $\mathsf{KL}_2(p_1, p_2)$ the KL variations, i.e., $\mathsf{KL}(p_1, p_2) := \int \log\left(\frac{p_1(x)}{p_2(x)}\right) p_1(x)\lambda(\mathrm{d}x)$ and $\mathsf{KL}_2(p_1, p_2) := \int \left\{\log\left(\frac{p_1(x)}{p_2(x)}\right)\right\}^2 p_1(x)\lambda(\mathrm{d}x)$. For $\zeta > 0$, a space of certain distributions $\mathcal{G}$ and a distribution $G_0 \in \mathcal{G}$, we define a ζ-KL neighborhood of G_0 by

$$\mathcal{B}_{\mathsf{KL}}(\zeta, G_0, \mathcal{G}) := \left\{G \in \mathcal{G} : \mathsf{KL}(p_{G_0}, p_G) < \zeta^2, \mathsf{KL}_2(p_{G_0}, p_G) < \zeta^2\right\}.$$

For a real-valued function f on $\mathfrak{X}$, let $\|f\|_q := \left(\int |f(x)|^q \lambda(\mathrm{d}x)\right)^{1/q}$ for $q > 0$ and $\|f\|_\infty := \sup_{x \in \mathfrak{X}} |f(x)|$. For $G \in \mathcal{P}(\mathbb{R})$, we denote by $m_h(G)$ the h-th moment of G, i.e., $m_h(G) := \int x^h \mathsf{P}_G(\mathrm{d}x)$. The r-th moment vector is defined by $m_{1:r}(G) := \big(m_1(G), \cdots, m_r(G)\big)$.

2.2. Gaussian location mixtures

In this paper, we initially consider the Gaussian location mixture model in one dimension:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \sum_{j=1}^k w_j \mathrm{N}(\theta_j, \sigma^2), \tag{2.1}$$

where $\theta_1, \ldots, \theta_k \in \mathbb{R}$ are the *atoms* and $(w_1, \ldots, w_k) \in \Delta_k$ are the mixing *weights*. Here we define

$$\Delta_k := \left\{(w_1, \ldots, w_k) \in [0,1]^k : \sum_{j=1}^k w_j = 1\right\}$$

for $k \in \mathbb{N}$. We assume that the variance σ^2 is known and without loss of generality $\sigma^2 = 1$. With the convolution denoted with the symbol $*$, we simply write

$$p * \Phi = \sum_{j=1}^{k} w_j \mathrm{N}(\theta_j, 1)$$

for the mixing distribution $p := \sum_{j=1}^{k} w_j \delta_{\theta_j}$, where Φ denotes the standard normal distribution. For a set $\Theta \subset \mathbb{R}$ and $k \in \mathbb{N}$, we define the set of k-atomic distributions

$$\mathcal{M}_k(\Theta) := \left\{ \sum_{j=1}^{k} w_j \delta_{\theta_j} : (w_1, \ldots, w_k) \in \Delta_k, \theta_1, \ldots, \theta_k \in \Theta \right\}.$$

Note that $\mathcal{M}_k(\Theta) \subset \mathcal{M}_{k+1}(\Theta)$ for every $k \in \mathbb{N}$. The parameter space is given by $\mathcal{M}(\Theta) := \bigcup_{k \in \mathbb{N}} \mathcal{M}_k(\Theta)$. Note that $\mathcal{M}(\Theta) \subset \mathcal{P}(\Theta)$.

For mixture models, the Wasserstein distance is widely used as a performance measure for the mixing distribution estimation. To define the Wasserstein distance between two atomic distributions, we first define

$$Q(w, w') := \left\{ (p_{jh})_{j \in [k], h \in [k']} \in [0,1]^{k \times k'} : \sum_{h=1}^{k'} p_{jh} = w_j, \sum_{j=1}^{k} p_{jh} = w'_h, \forall j \in [k], h \in [k'] \right\},$$

for given two weight vectors $w \in \Delta_k$ and $w' \in \Delta_{k'}$, which is a set of joint distributions on $[k] \times [k']$ with marginal distributions w and w'. For any $q \geq 1$, the q-th order Wasserstein distance between two atomic distributions $p := \sum_{j=1}^{k} w_j \delta_{\theta_j}$ and $p' := \sum_{h=1}^{k'} w'_h \delta_{\theta'_h}$ is defined as

$$\mathsf{W}_q(p, p') := \inf_{p \in Q(w, w')} \left(\sum_{j=1}^{k} \sum_{h=1}^{k'} p_{jh} |\theta_j - \theta'_h|^q \right)^{1/q}.$$

2.3. Prior distribution

We first assume that the true data generating process is given as $p^\star * \Phi$ where $p^\star \in \mathcal{M}_{k^\star}([-L, L])$, $L > 0$ for some $k^\star \in \mathbb{N}$, which is the true number of mixing components. For simplicity, we write $\mathcal{M}_k := \mathcal{M}_k([-L, L])$ for each $k \in \mathbb{N}$ and $\mathcal{M} := \mathcal{M}([-L, L]) := \cup_{k=1}^{\infty} \mathcal{M}_k([-L, L])$. We consider a general model in which the true mixing distribution $p^\star \in \mathcal{M}_{k^\star}$ can vary with sample size n as well as the true number of components $k^\star$ can vary with n. This is a critical difference from the existing Bayesian literature on mixture models which assumed a fixed true mixing distribution [19,36,45].

We assume an upper bound $\bar{k}_n \lesssim \log n / \log \log n$ on the true number of components $k^\star$. This assumption alleviates some technical difficulties, and can be justified by the following remark. Since the minimax optimal convergence rate of mixing distribution estimation for large mixtures $p^\star \in \mathcal{M}_{k^\star}$ with $k^\star \asymp \log n / \log \log n$ is the same as the one for mixtures $p^\star \in \mathcal{M}$ with any order, which is a slow rate of $\log \log n / \log n$ (See Proposition 9 of [49]), without assuming $k^\star \lesssim \log n / \log \log n$, we may not obtain improved convergence rates. We will also show that one can develop a Bayesian procedure that attains this minimax rate without knowing the upper bound of the true number of components. See Theorem 2.7 in Section 2.5.

We now introduce our prior distribution on the finite Gaussian mixture model. The prior distribution first samples the number of components k from $\Pi(k)$ and then samples the atoms $\theta \in [-L, L]^k$ and weights $w \in \Delta_k$ from $\Pi(\theta|k)$ and $\Pi(w|k)$, respectively. Thus the prior distribution induces a distribution on $\mathcal{M} = \cup_{k\in\mathbb{N}}\mathcal{M}_k$.

We impose the following conditions on the prior.

Assumption P. Recall that $\bar{k}_n$ is the known upper bound on the true number of components. The prior distribution Π satisfies the following conditions:

(P1) The prior distribution on the number of components k is sample size dependent. There are a constant $c_1 > 0$ and a sufficiently large constant $A > 0$ such that for any sample size $n \in \mathbb{N}$ and any $k^\circ \in \mathbb{N}$,

$$\frac{\Pi(k = k^\circ + 1)}{\Pi(k = k^\circ)} \leq c_1 \mathrm{e}^{-A\bar{k}_n \log n}. \tag{2.2}$$

Additionally, there are constants $c_2 > 0$ and $c_3 > 0$ such that for any $n \in \mathbb{N}$ and any $k^\dagger \in [\bar{k}_n]$,

$$\Pi(k = k^\dagger) \geq c_2 \mathrm{e}^{-(c_3\bar{k}_n \log n)k^\dagger}. \tag{2.3}$$

(P2) For any $k \in \mathbb{N}$ and any $(w_1^0, \ldots, w_k^0) \in \Delta_k$, there are positive constants c_4 and c_5 such that for any $\eta \in (0, 1/k)$,

$$\Pi\left(\sum_{j=1}^{k} |w_j - w_j^0| \leq \eta \middle| k\right) \geq c_4 \eta^{c_5 k}. \tag{2.4}$$

(P3) For any $k \in \mathbb{N}$ and any $\theta^0 \in [-L, L]^k$, there are positive constants c_6 and c_7 such that for any $\eta > 0$,

$$\Pi\left(\max_{1\leq j\leq k} |\theta_j - \theta_j^0| \leq \eta \middle| k\right) \geq c_6 \eta^{c_7 k}. \tag{2.5}$$

In (P1), we require a prior distribution to heavily penalize mixture models with a large number of components, and further assume that the degree of the penalization becomes more severe of an appropriate order as sample size grows. This enables the resulting posterior distribution not to overestimate the number of components.

Remark 1. The idea of using a same size dependent prior to control model complexity is not new and have frequently appeared in the Bayesian literature, e.g., prior distributions on sparsity in the multivariate normal mean model [5], sparsity in nonparametric regression [24], the number of communities in the stochastic block model [14] and the number of factors in the factor model [38].

We now provide some examples of prior distributions satisfying Assumption P. In the following examples, the constant $A > 0$ is the same as the one appearing in (2.2).

Example 1. The mixture of finite mixture (MFM) prior considered in [19,26,32] is a hierarchical prior consisting of distributions on the number of components, the weights and the atoms. Assumption P is met by the MFM prior with appropriate choices of each distribution. The geometric distribution with probability mass function $\Pi(k) = (1 - p_n)^{k-1} p_n$ on k with $p_n := 1 - a\exp(-A\bar{k}_n \log n)$ for arbitrary $a > 0$, satisfies (P1). Also, the Poisson-like distribution on $\mathbb{N}$ such that $\Pi(k) = \mathrm{e}^{-\lambda_n}\lambda_n^{k-1}/(k-1)!$ with

$\lambda_n := a\exp(-A\bar{k}_n \log n)$ for arbitrary $a > 0$ satisfies (P1). The Dirichlet distribution $\mathrm{DIR}(\kappa_1, \ldots, \kappa_k)$ on the mixing weights with $\kappa_j \in (\kappa_0, 1)$ for every $j \in [k]$ and some $\kappa_0 \in (0,1)$ satisfies (P2), see Lemma A.6. If the prior distribution on θ behaves like a uniform distribution up to a multiplicative constant, then (P3) holds.

Example 2. Consider a Binomial prior distribution on the number of components such that $k - 1 \sim \mathrm{BINOM}(\bar{k}_n - 1, p_n)$ with $p_n := a\exp(-2A\bar{k}_n \log n)$ for arbitrary $a > 0$. Then this prior satisfies (2.2) since $\binom{\bar{k}_n - 1}{k^\circ} / \binom{\bar{k}_n - 1}{k^\circ - 1} \le \bar{k}_n \lesssim \mathrm{e}^{\log\log n}$ and $1 - p_n \le 1$. Also it satisfies (2.3) since $1 - p_n \gtrsim 1$. The MFM prior with this Binomial prior distribution satisfies Assumption P.

Example 3. The spike and slab prior distribution on the *unnormalized* weights can satisfy (P1) and (P2). Suppose that we consider an over-fitted mixture model $\nu = \sum_{j=1}^{\bar{k}_n} w_j \delta_{\theta_j}$. Let $S := \{j \in [\bar{k}_n] : w_j > 0\}$, a set of indices corresponding to nonzero weights. Then we can write $\nu = \sum_{j \in S} w_j \delta_{\theta_j}$. Let $\tilde{w} \equiv (\tilde{w}_j)_{j \in [\bar{k}_n]}$ be the independent random variables where $\tilde{w}_1$ is generated from $\mathrm{GAMMA}(\kappa, b)$ and the other variables, i.e., $\tilde{w}_2, \ldots, \tilde{w}_{\bar{k}_n}$, are generated from a spike and slab distribution $(1-p_n)\delta_0 + p_n \mathrm{GAMMA}(\kappa, b)$ with $p_n := a\exp(-2A\bar{k}_n \log n)$ for $a > 0$, $b > 0$ and $\kappa \in (0,1)$. If we define the number of components as the number of nonzero elements in $\tilde{w}$ and the weights as a normalized version of $(\tilde{w}_j)_{j \in S}$, i.e., $k := \|\tilde{w}\|_0$ and $w_j := \tilde{w}_j / \|\tilde{w}\|_1$ for $j \in S$, then $k - 1$ follows $\mathrm{BINOM}(\bar{k}_n - 1, p_n)$ and $(w_j)_{j \in S}$ follows $\mathrm{DIR}(\kappa, \ldots, \kappa)$. Thus Assumption P holds by Examples 1 and 2.

2.4. Posterior concentration

In this section, we present concentration properties of the posterior distribution $\Pi(\cdot|X_{1:n})$ defined below, with the prior given in Section 2.3 and the data from the Gaussian mixture model in (2.1):

$$\Pi(\mathrm{d}\nu|X_{1:n}) := \frac{p^{(n)}_{\nu * \Phi}(X_{1:n})\Pi(\mathrm{d}\nu)}{\int p^{(n)}_{\nu * \Phi}(X_{1:n})\Pi(\mathrm{d}\nu)}. \tag{2.6}$$

We first show that our posterior distribution does not *overestimate* the number of components.

Theorem 2.1. *Assume that $k^\star \le \bar{k}_n \lesssim \log n / \log\log n$. Then with the prior distribution Π satisfying Assumption P, we have*

$$\inf_{\nu^\star \in \mathcal{M}_{k^\star}} \mathsf{P}^{(n)}_{\nu^\star * \Phi}\left[\Pi(\nu \in \mathcal{M}_{k^\star}|X_{1:n})\right] \to 1. \tag{2.7}$$

The following theorem shows the optimal concentration property of the posterior distribution of the mixing distribution.

Theorem 2.2. *Under the same assumptions of Theorem 2.1, we have*

$$\sup_{\nu^\star \in \mathcal{M}_{k^\star}} \mathsf{P}^{(n)}_{\nu^\star * \Phi}\left[\Pi\left(\mathsf{W}_1(\nu, \nu^\star) \ge M\bar{\epsilon}_{n,k^\star}\middle|X_{1:n}\right)\right] = o(1) \tag{2.8}$$

for some universal constant $M > 0$, where

$$\bar{\epsilon}_{n,k^\star} := (k^\star)^{\frac{3}{2}}\left(\frac{\log^2 n}{n}\right)^{\frac{1}{4k^\star - 2}}. \tag{2.9}$$

If the number of components $k^\star$ is fixed, the convergence rate in Theorem 2.2 is equivalent to the minimax optimal rate $n^{-1/(4k^\star-2)}$ [49, Proposition 7] up to at most a logarithmic factor. An additional logarithmic factor is common in the nonparametric Bayesian literature, which often arises due to the popular "prior mass and testing" proof technique which we also adopt in this paper. We refer to the papers [15,23] for discussions about this phenomenon.

To improve the convergence rate in Theorem 2.2, one may assume that atoms are well separated and the weights are bounded away from zero. We introduce the formal definition related to this notion.

Definition 1. An atomic distribution $\nu := \sum_{j=1}^{k} w_j \delta_{\theta_j}$ is said to be k_0 (γ,ω)-*separated* for $k_0 \in [k]$, $\gamma > 0$ and $\omega > 0$ if there exists a partition $S_1,\ldots,S_{k_0}$ of $[k]$ such that

- $|\theta_j - \theta_{j'}| \geq \gamma$ for any $j \in S_l$, $j' \in S_{l'}$ and any $l, l' \in [k_0]$ with $l \neq l'$;
- $\sum_{j\in S_l} w_j \geq \omega$ for any $l \in [k_0]$.

We let $\mathcal{M}_{k,k_0,\gamma,\omega} := \left\{\nu \in \mathcal{M}_k : \nu \text{ is } k_0\ (\gamma,\omega)\text{-separated}\right\}$.

In the next theorem, we derive the optimal posterior contraction rate of the mixing distribution under the separation assumption. We call this contraction rate an *adaptive rate* because the result is achieved without any knowledge of the number of well-separated components k_0 of the true mixing distribution.

Theorem 2.3. *Assume that* $k^\star \leq \bar{k}_n \lesssim \log n/\log\log n$ *and* $\gamma\omega > M'\bar{\epsilon}_{n,k^\star}$ *for a sufficiently large constant* $M' > 0$*, where* $\bar{\epsilon}_{n,k^\star}$ *is the rate defined in* (2.9)*. Then with the prior distribution* Π *satisfying Assumption P, we have*

$$\sup_{\nu^\star\in\mathcal{M}_{k^\star,k_0,\gamma,\omega}} \mathsf{P}^{(n)}_{\nu^\star * \Phi}\left[\Pi\left(\mathsf{W}_1(\nu,\nu^\star) \geq M\epsilon_{n,k^\star,k_0,\gamma}\middle|X_{1:n}\right)\right] = o(1), \tag{2.10}$$

for some universal constant $M > 0$*, where*

$$\epsilon_{n,k^\star,k_0,\gamma} := (k^\star)^{\frac{6k^\star-4k_0+3}{4(k^\star-k_0)+2}}\gamma^{-\frac{2k_0-2}{2(k^\star-k_0)+1}}\left(\frac{\log^2 n}{n}\right)^{\frac{1}{4(k^\star-k_0)+2}}. \tag{2.11}$$

Remark 2. A nice surprise from the result of Theorem 2.3 is that our Bayesian procedure can achieve a better convergence rate than the one in Theorem 2.2 without requiring any further condition on the prior distribution. This is because of fact that the condition $\gamma\omega > M'\bar{\epsilon}_n$ guarantees that the mixing distribution ν is k_0 $(a_0\gamma,0)$-separated asymptotically for some constant $a_0 \in (0,1)$ under the posterior distribution, provided that Theorem 2.2 holds.

In view of Proposition 2.4 presented below, the convergence rate in Theorem 2.3 is minimax optimal [20, Theorem 3.2] up to a logarithmic factor if the model parameters $k^\star, k_0$ and γ are fixed constants. Heinrich and Kahn [20] established the minimax optimal rate $n^{-1/(4(k^\star-k_0)+2)}$ of the estimation of the mixing distribution satisfying the *locally varying* condition. Namely, they showed that for fixed $k^\star \in \mathbb{N}$, $k_0 \in [k^\star]$ and $\nu_0 \in \mathcal{M}_{k_0} \setminus \mathcal{M}_{k_0-1}$, it follows that

$$\inf_{\{\hat{\nu}\}} \sup_{\nu^\star\in\mathcal{M}_{k^\star}:\mathsf{W}_1(\nu^\star,\nu_0)\leq\epsilon_n^\dagger} \mathsf{P}^{(n)}_{\nu^\star*\Phi}\left[\mathsf{W}_1(\hat{\nu},\nu^\star)\right] \gtrsim n^{-\frac{1}{4(k^\star-k_0)+2}}, \tag{2.12}$$

where the infimum ranges over all possible sequences of estimators and $\epsilon_n^\dagger := n^{-1/(4(k^\star-k_0)+2)+\iota}$ for some $\iota > 0$ (In fact, the above lower bound holds not only for the Gaussian location mixtures but

also general mixtures satisfying the strong identifiability condition provided in Definition 2). In other words, the above minimax argument is about the true mixing distribution which does not vary globally but *locally*. This locally varying condition is seemingly different from the separation condition given in Definition 1, but in fact the former is a sufficient condition of the latter. Intuitively, we can expect that the true distribution $\nu^\star \in \mathcal{M}_{k^\star}$ close to $\nu_0 \in \mathcal{M}_{k_0} \setminus \mathcal{M}_{k_0-1}$ has at least k_0 well-separated components, and therefore satisfies the separation condition. We formally state this argument in the next proposition.

Proposition 2.4. *Let $k_0 \in \mathbb{N}$ and $\nu_0 := \sum_{j=1}^{k_0} w_{0j}\delta_{\theta_{0j}} \in \mathcal{M}_{k_0} \setminus \mathcal{M}_{k_0-1}$. Define*

$$\gamma(\nu_0) := \min_{j,h\in[k_0]:j\neq h} |\theta_{0j} - \theta_{0h}|$$

$$\omega(\nu_0) := \min_{j\in[k_0]} w_{0j}.$$

Let $k \in \{k_0, k_0+1, \dots\}$ and $c \in (0, 1/4)$. Then we have

$$\left\{\nu \in \mathcal{M}_k : \mathsf{W}_1(\nu, \nu_0) < c\gamma(\nu_0)\omega(\nu_0)\right\} \subset \mathcal{M}_{k,k_0,(1-2c)\gamma(\nu_0),\frac{1-4c}{1-3c}\omega(\nu_0)}. \tag{2.13}$$

Due to Proposition 2.4, it is clear that our Bayesian procedure is also near-optimal for the estimation of the mixing distribution under the locally varying condition. We merely state the result.

Corollary 2.5. *Assume $k^\star \leq \bar{k}_n \lesssim \log n/\log\log n$. Let $k_0 \in \mathbb{N}$ be a fixed constant such that $k_0 \leq k^\star$, and let $\nu_0 \in \mathcal{M}_{k_0} \setminus \mathcal{M}_{k_0-1}$ be a fixed distribution. Moreover, assume that the prior distribution Π satisfies Assumption P. Then there exist universal constants $\tau > 0$ and $M > 0$ such that*

$$\sup_{\nu^\star\in\mathcal{M}_{k^\star}:\mathsf{W}_1(\nu^\star,\nu_0)<\tau} \mathsf{P}^{(n)}_{\nu^\star * \Phi}\left[\Pi\left(\mathsf{W}_1(\nu,\nu^\star) \geq M\epsilon_{n,k^\star,k_0,1}\middle|X_{1:n}\right)\right] = o(1), \tag{2.14}$$

where $\epsilon_{n,k^\star,k_0,1}$ is the rate in (2.9) with $\gamma = 1$, i.e., $\epsilon_{n,k^\star,k_0,1} = (k^\star)^{\frac{6k^\star-4k_0+3}{4(k^\star-k_0)+2}}\left(\frac{\log^2 n}{n}\right)^{\frac{1}{4(k^\star-k_0)+2}}$.

As a byproduct, we can obtain the posterior consistency of the true number of components when the true mixing distribution $\nu^\star$ is perfectly separated, that is, $k^\star = k_0$. Note that in this case, $\nu^\star \in \mathcal{M}_{k^\star} \setminus \mathcal{M}_{k^\star-1}$. The following theorem states this formally.

Theorem 2.6. *Assume that $k^\star \leq \bar{k}_n \lesssim \log n/\log\log n$ and*

$$\gamma\omega > M' \max\{\bar{\epsilon}_{n,k^\star}, \epsilon_{n,k^\star,k^\star,\gamma}\} \tag{2.15}$$

for a sufficiently large constant $M' > 0$, where $\bar{\epsilon}_{n,k^\star}$ and $\epsilon_{n,k^\star,k^\star,\gamma}$ are the rates defined in (2.9) and (2.11) with $k_0 = k^\star$, respectively. Then with the prior distribution Π satisfying Assumption P, we have

$$\inf_{\nu^\star\in\mathcal{M}_{k^\star,k^\star,\gamma,\omega}} \mathsf{P}^{(n)}_{\nu^\star*\Phi}\left[\Pi\left(\nu \in \mathcal{M}_{k^\star} \setminus \mathcal{M}_{k^\star-1}|X_{1:n}\right)\right] \to 1. \tag{2.16}$$

The condition (2.15) provides a threshold for detection. This condition plays a similar role as the beta-min condition for variable selection in linear regression [5,28].

Guha, Ho and Nguyen [19] obtained the consistency result with a similar prior distribution to ours, but their analysis is restricted to the fixed truth cases.

2.5. Higher-order mixtures

In Section 2, we have assumed that $k^\star \lesssim \log n/\log\log n$. This assumption is justified by the minimax result for the estimation of the higher-order mixtures presented by [49]. In this section, we prove that there is a Bayesian procedure which is similar to the one considered in Section 2, but does not assume a known upper bound of the number of components, can attain this minimax optimality. In this case, we impose a milder condition than (P1) on the prior.

(P1′) There are constants $c_1 > 0$, $c_2 > 0$ and $b_0 > 0$ such that for any $n \in \mathbb{N}$ and $k^\circ \in \mathbb{N}$,

$$\Pi(k = k^\circ) \geq c_1 \mathrm{e}^{-(c_2 \log^{b_0} n)k^\circ}. \tag{2.17}$$

It is clear that any prior distribution satisfying (P1) satisfies (P1′) with $b_0 = 2$ since $\bar{k}_n \lesssim \log n$. Also, Assumption (P1′) can be met by the Poisson and geometric distribution with constant mean and success probability, respectively, which do not satisfy (P1).

The next theorem provides the convergence rate of mixing distribution estimation without any restriction on the true number of components.

Theorem 2.7. *Then with the prior distribution* Π *satisfying (P1′), (P2) and (P3), we have*

$$\sup_{\nu^\star \in \mathcal{M}} \mathsf{P}^{(n)}_{\nu^\star * \Phi}\left[\Pi\left(\mathsf{W}_1(\nu, \nu^\star) \geq M \frac{\log\log n}{\log n} \middle| X_{1:n}\right)\right] = o(1) \tag{2.18}$$

for some universal constant $M > 0$.

If the true mixing distribution $\nu^\star$ belongs to $\mathcal{M}_{k^\star}$ with $k^\star \asymp \log n/\log\log n$, the convergence rate in the above theorem is rate-exact optimal [49, Theorem 5].

Indeed, the above result holds even when the true generating process is given by $\mu^\star * \Phi$ with $\mu^\star \in \mathcal{P}([-L, L])$, which includes continuous or infinite mixtures.

2.6. Extension to general mixture models

In this section, we extend our analysis for the Gaussian location mixture model to general mixture models with potentially growing number of components. For a mixing distribution $\nu \in \mathcal{M}(\Theta)$ and a family $\{F(\cdot, \theta) : \theta \in \Theta\}$ of distribution functions on $\mathbb{R}$ for $\Theta \subset \mathbb{R}$, we let $\nu \bullet F$ denote the distribution having a density function

$$p_{\nu \bullet F}(\cdot) := \int f(\cdot, \theta)\nu(\mathrm{d}\theta), \tag{2.19}$$

where $f(\cdot, \theta)$ stands for the probability density function of $F(\cdot, \theta)$. We call $F(\cdot, \cdot)$ and $f(\cdot, \cdot)$ a *kernel distribution function* and a *kernel density function*, respectively.

We impose the following set of assumptions on the kernel distribution function.

Assumption F. The family $\{F(\cdot, \theta) : \theta \in \Theta\}$ of distribution functions on $\mathbb{R}$ satisfies the following conditions:

(F1) Θ is a compact subset of $\mathbb{R}$ with nonempty interior.

(F2) There is a constant $c_1 > 0$ such that

$$\left\| f(\cdot,\theta_1) - f(\cdot,\theta_2) \right\|_\infty \le c_1 |\theta_1 - \theta_2| \tag{2.20}$$

for any $\theta_1, \theta_2 \in \Theta$. Moreover, there are constants $c_2 > 0$ and $r \in (0,1]$ such that

$$\int p_{\nu_1 \bullet F}(x) \left(\frac{p_{\nu_1 \bullet F}(x)}{p_{\nu_2 \bullet F}(x)} \right)^r \lambda(\mathrm{d}x) \le c_2 \tag{2.21}$$

for any $\nu_1, \nu_2 \in \mathcal{M}(\Theta)$.

(F3) For any $k \in \mathbb{N}$, there exists an estimator $\hat{M}_k$ of the moment $m_k(\nu)$ based on the sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{P}_{\nu \bullet F}$ such that

$$\mathsf{P}^{(n)}_{\nu \bullet F} \hat{M}_k = m_k(\nu) \tag{2.22}$$

$$\mathsf{P}^{(n)}_{\nu \bullet F} \left(\hat{M}_k - m_k(\nu) \right)^2 \lesssim \frac{1}{n} (c_3 + \sqrt{k})^{2k} \tag{2.23}$$

for any $\nu \in \mathcal{M}(\Theta)$ for some constant $c_3 > 0$.

Assumption (F2) helps us establish a lower bound of a KL neighborhood of the true distribution $\nu^\star \bullet F$ by using the prior concentration conditions in Assumption P. This assumption is held for a wide range of choices of kernel density function, for example, the Gaussian location family [36] and more generally, location family of exponential power distributions [44]. Also, the Gaussian scale family satisfies Assumption (F2) as shown in the next example.

Example 4 (Gaussian scale family). Let F be the kernel distribution function such that $F(x,\sigma) = \Phi_\sigma(x)$, where Φ_σ denotes the distribution function of $\mathrm{N}(0,\sigma^2)$. Consider the Gaussian scale family $\left\{ F(\cdot,\sigma) : \sigma \in [1/L, L] \right\}$ for $L > 1$. Then this family satisfies (2.20) since $\left| \frac{\partial}{\partial \sigma} f(x,\sigma) \big|_{\sigma=\sigma_0} \right| < \infty$ for any $\sigma_0 \in [1/L, L]$ and $x \in \mathbb{R}$. For (2.21), let $r_0 := 1/(2L^4)$. Fix two mixing distributions $\nu_1 := \sum_{j=1}^{k_1} w_{1,j} \delta_{\sigma_{1,j}}$ and $\nu_2 := \sum_{j=1}^{k_2} w_{2,j} \delta_{\sigma_{2,j}}$. Without loss of generality, we assume $\sigma_{i,1} < \sigma_{i,2} < \cdots < \sigma_{i,k_i}$, for all $i = 1, 2$. Then

$$\begin{aligned}
\int p_{\nu_1 \bullet F}(x) \left(\frac{p_{\nu_1 \bullet F}(x)}{p_{\nu_2 \bullet F}(x)} \right)^{r_0} \lambda(\mathrm{d}x) &\lesssim \int \mathrm{e}^{-\frac{1}{2\sigma_{1,k_1}^2} x^2} \mathrm{e}^{-\frac{r_0}{2\sigma_{1,k_1}^2} x^2 + \frac{r_0}{2\sigma_{2,1}^2} x^2} \lambda(\mathrm{d}x) \\
&\le \int \mathrm{e}^{-\frac{1}{2L^2} x^2 + \frac{r_0 L^2}{2} x^2} \lambda(\mathrm{d}x) \\
&= \int \mathrm{e}^{-\frac{1}{4L^2} x^2} \lambda(\mathrm{d}x) < \infty,
\end{aligned}$$

which verifies (2.21).

Assumption (F3) requires the existence of the unbiased estimator of the moment of every order whose variance is bounded by certain quantity depending on the order. This condition enables us to use the theoretical tool developed in Wu and Yang [49], who studied an estimator of the mixing distribution based on the method of moments for the Gaussian location mixture model. Indeed, in the proof of our results for the Gaussian location mixture model, provided in Appendix A.2, we found the moment estimator (A.4) that satisfies Assumption (F3). We give some examples that satisfy Assumption (F3).

Example 5 (Gaussian location family). Our Lemma A.1 proves that the Gaussian location family satisfies (F3).

Example 6 (Gaussian scale family). Consider the Gaussian scale family $\{\Phi_\sigma : \sigma > 0\}$. It is easy to see that the moment estimator $\hat{M}_k = \sum_{i=1}^n X_i^k / \mathsf{E}(Z^k)$ with $Z \sim \mathrm{N}(0,1)$ satisfies (2.22) and (2.23).

Example 7 (Quadratic variance exponential family (QVEF)). An exponential family of which the variance of each distribution is at most a quadratic function of its mean is called a QVEF [33]. The class of QVEFs includes Poisson, gamma, binomial, and negative binomial distributions. If the family of distribution functions $\{F(\cdot,\theta) : \theta \in \Theta\}$ is a QVEF, then (F3) is satisfied. Indeed, Equations (8.8) and (8.6) of [33] verify (2.22) and (2.23), respectively.

Remark 3. As a reviewer pointed out, Assumption (F3) is somewhat strong and a number of mixture models do not satisfy this. For example, although the Cauchy location mixture model with $f(x,\theta) = (\pi(1+(x-\theta)^2)^{-1}$ is strongly identifiable (by [7, Theorem 3]) and so can be analyzed under a different theoretical framework given in Appendix B, it does not satisfy Assumption (F3) since the Cauchy distribution does not have finite moments of order greater than or equal to 1.

Since we consider a general set of atoms $\Theta \subset \mathbb{R}$ rather than the interval $[-L, L]$ to include, for example, scale mixtures and exponential family mixtures, Assumption (P3) is slightly modified to Equation (2.5) being met for any $k \in \mathbb{N}$ and $\theta^0 \in \Theta^k$. We also assume the kernel distribution function $F(\cdot,\cdot)$ is known, i.e., no misspecification of the kernel distribution function. That is, we consider the posterior distribution denoted by $\Pi_F(\cdot|X_{1:n})$, which is defined as

$$\Pi_F(\mathrm{d}\nu|X_{1:n}) := \frac{p^{(n)}_{\nu\bullet F}(X_{1:n})\Pi(\mathrm{d}\nu)}{\int p^{(n)}_{\nu\bullet F}(X_{1:n})\Pi(\mathrm{d}\nu)}. \tag{2.24}$$

Then all the results in Section 2 can be recovered by the posterior distribution $\Pi_F(\cdot|X_{1:n})$ on the mixture model that satisfies Assumption F.

Theorem 2.8. *Assume that $k^\star \le \bar{k}_n \lesssim \log n / \log\log n$. Moreover, assume that the family $\{F(\cdot,\theta) : \theta \in \Theta\}$ of distribution functions on $\mathbb{R}$ satisfies Assumption F and the prior distribution Π satisfies Assumption P. Then the followings are hold:*

(a) *It follows that*

$$\inf_{\nu^\star \in \mathcal{M}_{k^\star}} \mathsf{P}^{(n)}_{\nu^\star \bullet F}\left[\Pi_F(\nu \in \mathcal{M}_{k^\star}|X_{1:n})\right] \to 1; \tag{2.25}$$

(b) *There exists an universal constant $M_1 > 0$ such that*

$$\sup_{\nu^\star \in \mathcal{M}_{k^\star}} \mathsf{P}^{(n)}_{\nu^\star \bullet F}\left[\Pi_F\left(\mathsf{W}_1(\nu,\nu^\star) \ge M_1 \bar{\epsilon}_{n,k^\star}\middle|X_{1:n}\right)\right] = o(1), \tag{2.26}$$

where $\bar{\epsilon}_{n,k^\star}$ is the rate defined in (2.9);

(c) *There exist universal constants $M_2 > 0$ and $M_3 > 0$ such that if $\gamma\omega > M_2\bar{\epsilon}_{n,k^\star}$ then*

$$\sup_{\nu^\star \in \mathcal{M}_{k^\star,k_0,\gamma,\omega}} \mathsf{P}^{(n)}_{\nu^\star \bullet F}\left[\Pi_F\left(\mathsf{W}_1(\nu,\nu^\star) \ge M_3\epsilon_{n,k^\star,k_0,\gamma}\middle|X_{1:n}\right)\right] = o(1), \tag{2.27}$$

where $\epsilon_{n,k^\star,k_0,\gamma}$ is the rate defined in (2.11);

(d) *There exist universal constants* $\tau > 0$ *and* $M_4 > 0$ *such that*

$$\sup_{\nu^\star \in \mathcal{M}_{k^\star}: \mathsf{W}_1(\nu^\star, \nu_0) < \tau} \mathsf{P}^{(n)}_{\nu^\star \bullet F} \left[\Pi_F \left(\mathsf{W}_1(\nu, \nu^\star) \geq M_4 \epsilon_{n,k^\star,k_0,1} \middle| X_{1:n} \right) \right] = o(1) \tag{2.28}$$

for any fixed distribution $\nu_0 \in \mathcal{M}_{k_0} \setminus \mathcal{M}_{k_0-1}$*;*

(e) *There exists an universal constant* $M_5 > 0$ *such that if* $\gamma\omega > M_5 \max\{\bar{\epsilon}_{n,k^\star}, \epsilon_{n,k^\star,k^\star,\gamma}\}$ *then*

$$\inf_{\nu^\star \in \mathcal{M}_{k^\star,k^\star,\gamma,\omega}} \mathsf{P}^{(n)}_{\nu^\star \bullet F} \left[\Pi_F \left(\nu \in \mathcal{M}_{k^\star} \setminus \mathcal{M}_{k^\star-1} | X_{1:n} \right) \right] \to 1. \tag{2.29}$$

The proof of the theorem is straightforward, but for the sake of completeness, we provide it in Appendix A.7.

We have considered mixture models with the number of components $k^\star$ satisfying $k^\star \leq \bar{k}_n \lesssim \log n / \log\log n$ for general kernel functions. For higher-order mixture models with general kernel functions, we can obtain the same convergence rate as the one in Theorem 2.7, which proves convergence rates for higher-order Gaussian location mixtures.

Theorem 2.9. *Assume that the family* $\{F(\cdot, \theta) : \theta \in \Theta\}$ *of distribution functions on* $\mathbb{R}$ *satisfies Assumption F and the prior distribution* Π *satisfies (P1′), (P2) and (P3). Then*

$$\sup_{\nu^\star \in \mathcal{M}} \mathsf{P}^{(n)}_{\nu^\star \bullet F} \left[\Pi_F \left(\mathsf{W}_1(\nu, \nu^\star) \geq M \frac{\log\log n}{\log n} \middle| X_{1:n} \right) \right] = o(1) \tag{2.30}$$

for some universal constant $M > 0$*.*

3. Dirichlet process mixtures for inference of finite mixtures

In this section, we consider Dirichlet process (DP) prior [11] on the mixing distribution which results in an infinite mixture model– the popular Dirichlet process (DP) mixture model. Although a DP mixture model is minimax optimal in density estimation [16,17], it suffers from a very slow convergence rate of $(\log n)^{-1/2}$ in estimating the mixing distribution of the Gaussian location mixtures as shown by [36]. Their result assumes that the number of component $k^\star$ is fixed. We consider the DP prior for the mixture distribution estimation and derive the posterior contraction rates in the most general set up by allowing the number of the components of the true mixing distribution to grow. Further more, we adopt a natural strategy of using the *number of the clusters T of the data* to estimate the number of components and we establish posterior consistency of such a procedure.

Note that the DP prior does not satisfy Assumption (P1), and thus the theorems in Section 2.4 do not cover the case of DP prior. This section aims to separately analyze concentration properties of the posterior of the DP mixture models.

In our Gaussian location mixture setup, the DP is a distribution on *infinite*-atomic distributions of the form

$$\tilde{\nu} := \sum_{j=1}^{\infty} w_j \delta_{\theta_j} \tag{3.1}$$

where $w_1, w_2, \cdots \in [0,1]$ are mixing weights such that $\sum_{j=1}^{\infty} w_j = 1$ and $\theta_1, \theta_2, \cdots \in [-L, L]$. We let $\mathcal{M}_\infty$ be the set of distributions of the form (3.1). The DP with a concentration parameter $\kappa > 0$ and

base distribution H, denoted by $\mathrm{DP}(\kappa, H)$, can be expressed by the following stick-breaking generation process [46]: $w_j = E_j \prod_{h=1}^{j-1}(1-E_h)$ where $E_j \overset{\text{iid}}{\sim} \mathrm{BETA}(1,\kappa)$ and $\theta_j \overset{\text{iid}}{\sim} H$. Since every w_j generated from the above procedure is positive with probability 1, one can say that $\Pi_{\mathrm{DP}}(\tilde{\nu} \in \mathcal{M}_\infty \setminus \mathcal{M}) = 1$. This implies that every mixing distribution generated from the posterior of the DP mixture model has infinite number of components, therefore the posterior distribution of the number of components k cannot provide any reasonable estimate of the true number of components.

One possible solution is to use an additional post-processing procedure for the posterior distribution. For example, Guha, Ho and Nguyen [19] proposed the operator $\mathcal{T}$ to infinite mixing distributions which removes weak components (in a sense that the corresponding weights are very small) and merges similar components (whose atoms are very close) of an infinite mixing distribution so that $\mathcal{T}(\tilde{\nu})$ is a finite mixing distribution. They proved that for a *fixed* truth $\nu^\star \in \mathcal{M}_{k^\star} \setminus \mathcal{M}_{k^\star-1}$, the posterior distribution of the finite mixing distribution $\mathcal{T}(\tilde{\nu})$ obtained after post processing concentrates to the model $\mathcal{M}_{k^\star} \setminus \mathcal{M}_{k^\star-1}$ under the DP prior distribution with a fixed concentration parameter.

We propose another way to infer the number of components with the DP prior. Our idea is to use the posterior distribution of the number of clusters, say T_n, of the data $X_{1:n}$ as an estimate of the number of components. Note that for $i \in [n]$, $X_i \overset{\text{iid}}{\sim} \tilde{\nu} * \Phi$ can be written equivalently with the latent *assignment variable* $Z_i \in \mathbb{N}$ as

$$Z_i \overset{\text{iid}}{\sim} w[\tilde{\nu}] := \sum_{j=1}^{\infty} w_j \delta_j,$$

$$X_i | Z_i \overset{\text{ind}}{\sim} \mathrm{N}(\theta_{Z_i}, 1).$$

where $w[\tilde{\nu}] \in \mathcal{P}(\mathbb{N})$ can be viewed as the distribution on $\mathbb{N}$ such that $w[\tilde{\nu}](J) = \tilde{\nu}(\{\theta_j : j \in J\})$ for any $J \subset \mathbb{N}$. *The number of clusters* T_n is defined by

$$T_n := T_n(Z_{1:n}) := \left| \left\{ j \in \mathbb{N} : \exists i \in [n] \text{ s.t. } Z_i = j \right\} \right|.$$

Here we consider the joint posterior distribution of the mixing distribution $\tilde{\nu}$ and the latent assignment variable $Z_{1:n}$ conditioned on the data $X_{1:n}$, which is given as

$$\Pi_{\mathrm{DP}}(\mathrm{d}\tilde{\nu}, Z_{1:n} | X_{1:n}) := \frac{\left[\prod_{i=1}^n \phi(X_i - \theta_{Z_i}) p_{w[\tilde{\nu}]}(Z_i)\right] \Pi_{\mathrm{DP}}(\mathrm{d}\tilde{\nu})}{\int \sum_{Z_{1:n} \in \mathbb{N}^n} \left[\prod_{i=1}^n \phi(X_i - \theta_{Z_i}) p_{w[\tilde{\nu}]}(Z_i)\right] \Pi_{\mathrm{DP}}(\mathrm{d}\tilde{\nu})}, \tag{3.2}$$

where $\phi(\cdot)$ denotes the probability density function of the standard normal distribution and Π_{DP} denotes the DP prior.

Note that the data are still assumed to be generated from the finite Gaussian mixture model $\nu^\star * \Phi$ where $\nu^\star \in \mathcal{M}_{k^\star}$ for $k^\star \in \mathbb{N}$ but we allow the number of components to grow at an arbitrary fast speed. Even in such general situations, we show in the following theorem that the DP prior with a suitably chosen concentration parameter can provide a nearly tight upper bound of *the true number of components*.

Theorem 3.1. *With the DP prior* $\mathrm{DP}(\kappa_n, H)$*, where* $\kappa_n \asymp n^{-a_0}$ *for* $a_0 > 0$ *and* H *is the uniform distribution on* $[-L, L]$*, we have*

$$\sup_{\nu^\star \in \mathcal{M}_{k^\star}} \mathsf{P}^{(n)}_{\nu^\star * \Phi} \left[\Pi_{\mathrm{DP}}(T_n > C k^\star | X_{1:n}) \right] = o(1) \tag{3.3}$$

for some constant $C > 1$ *depending only on the prior distribution.*

Miller and Harrison [30,31] showed that the posterior distribution of the number of clusters does not concentrate at the true number of components if one uses the DP prior with a *constant concentration parameter*. In particular, if the true data generating process is $\mathrm{N}(0,1) = \delta_0 * \Phi$, the posterior probability that the number of components is equal to the true number of components (i.e., 1) goes to zero [30, Theorem 5.1]. Our proposed sample size dependent concentration parameter resolves this inconsistency. See our simulation study in Section 4 for numerical confirmations.

Remark 4. Under the MFM prior of [32], which is an example of prior distributions considered in Section 2, the posterior distribution of T_n is asymptotically the same as the one of k. Miller and Harrison [32] proved that $|\Pi(k = k^\circ|X_{1:n}) - \Pi(T_n = k^\circ|X_{1:n})| \to 0$ almost surely for $k^\circ \in \mathbb{N}$ as long as $\Pi(k = k') > 0$ for any $k' \in [k^\circ]$. In view of this fact, the number of clusters T_n can be used to infer the true number of clusters $k^\star$ even if we use the MFM prior distribution.

Remark 5. One may wonder whether the choice of the concentration parameter $\kappa_n \asymp n^{-a_0}$ would lead to slower posterior contraction rate when the DP mixture model is used for *density estimation* as a DP mixture model is commonly adopted for. It turns out that it would not. In fact, even for $\kappa_n \asymp n^{-a_0}$, one can show that there is a universal constant $M > 0$ such that

$$\mathsf{P}^{(n)}_{\nu^\star * \Phi}\left[\Pi_{\mathrm{DP}}\left(\mathfrak{H}(p_{\tilde{\nu}*\Phi}, p_{\nu^\star*\Phi}) \geq M\frac{\log^c n}{\sqrt{n}}|X_{1:n}\right)\right] = o(1)$$

for any $\nu^\star \in \mathcal{P}([-L,L])$, for some $c > 0$. One can easily check the above result. Following the proof of Theorem 5.1 of [17] and applying Lemma A.6, we can see that the prior concentration near the true mixing distribution is lower bounded by $(n^{-a_0})^{c_1 \log n} \gtrsim \exp(-a_0 c_1 \log^2 n)$ for some $c_1 > 0$. Thus usual prior mass and testing approach leads to the conclusion in the preceding display for estimating the density.

For the estimation of the mixing distribution (of general order), we obtain the following convergence rate for the DP mixture model.

Theorem 3.2. *With the DP prior* $\mathrm{DP}(\kappa_n, H)$*, where* $\exp(-c_0 \log^{b_0} n) \lesssim \kappa_n \lesssim 1$ *for some* $b_0 > 0$ *and* $c_0 > 0$ *and* H *is the uniform distribution on* $[-L,L]$*, we have*

$$\sup_{\nu^\star \in \mathcal{M}} \mathsf{P}^{(n)}_{\nu^\star*\Phi}\left[\Pi_{\mathrm{DP}}\left(\mathsf{W}_1(\tilde{\nu}, \nu^\star) \geq M\frac{\log\log n}{\log n}\Big|X_{1:n}\right)\right] = o(1) \tag{3.4}$$

for some universal constant $M > 0$*.*

As one can see from our theorem above, if the true mixing distribution is of high order such that $k^\star \asymp \log n/\log\log n$, the posterior of the DP mixture model attains the minimax optimality [49, Theorem 5]. However, unlike the Bayesian procedure proposed in Section 2, we conjecture that posterior of the DP mixture model cannot obtain an improved convergence rate for estimating a mixing distribution when the true number of components grows slowly, say $k^\star \ll \log n/\log\log n$, because it tends to produce many redundant components. Nguyen [36] analyzed the posterior of Dirichlet process mixture endowed with a fixed concentration parameter for estimating mixing distribution with a *fixed* number of components and obtain a slow convergence rate $(\log n)^{-1/2}$ with respect to the second-order Wasserstein distance.

4. Numerical examples

4.1. Simulation study

We conduct numerical experiments to validate our theoretical findings. For the prior distribution, we use a MFM prior consisting of a Poisson distribution with mean λ on the number of components, the Dirichlet distribution on the weights and the uniform distribution on the atoms. For the Dirichlet distribution prior on the mixing weights, we fix its concentration parameter as a k-dimensional vector of 1's. For the mean parameter of the Poisson distribution, we consider the following two choices: the constant one and the one decaying with an appropriate order depending on the sample size. We call the former `MFM_const` and the latter `MFM_vary`. Note that `MFM_vary` is motivated by our theory. Python codes for reproducing the results in this section are available at https://github.com/ilsangohn/bayes_mixture

4.1.1. Inference for the mixing distribution

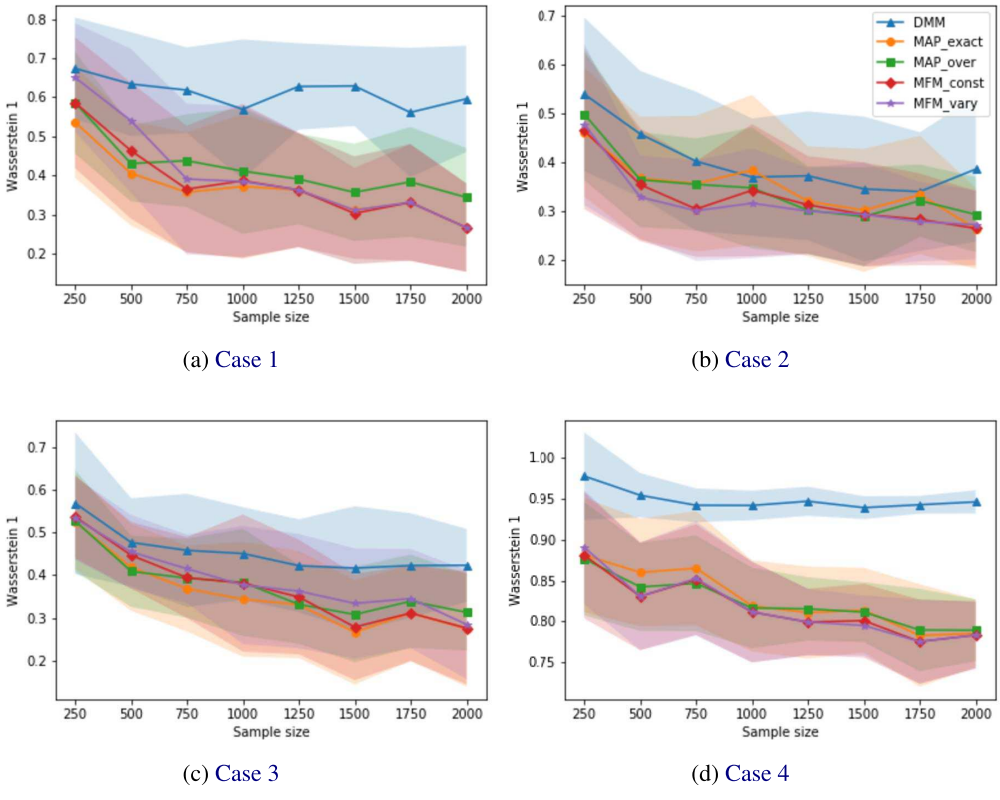
We compare the performance of the proposed Bayesian method with other competitors. We consider the denoised method of moment (`DMM`) estimator proposed by [49] and the maximum a posteriori (MAP) estimator with the Dirichlet distribution prior on the weights and the uniform distribution prior on the atoms. In the implementation of the `DMM` algorithm, we use the authors' Python codes which are available on their github repository (https://github.com/albuso0/mixture). We consider the MAP estimators of two types of mixture models: exact-fitted and over-fitted mixtures. The number of components of the exact-fitted mixture is exactly equal to the true number of components and the one of the over-fitted mixture is some upper bound $\bar{k}$ of the true number of components, in this simulation, we set $\bar{k} = 2k^\star$. We call the MAP estimator of the exact-fitted mixture `MAP_exact` and the one of the over-fitted mixture `MAP_over`. We use the standard expectation-maximization (EM) algorithm to obtain MAP estimators. For the proposed Bayesian method, we use the posterior mode of the mixing distribution as an estimator. We obtain such a mode by applying the EM algorithm to mixture models with the different numbers of components and selecting the best number of components $\hat{k}$ which maximizes the posterior density of the mode. We consider the two choices of the mean parameter of the Poisson prior, $\lambda_n = \exp(-0.05 \log^2 n / \log\log n))$ (`MFM_vary`) and $\lambda_n = 1$ (`MFM_const`). For all the four Bayesian methods, we set the support of the uniform distribution prior the interval $[-6, 6]$ and the concentration parameter of the Dirichlet distribution prior the vector of 1's.

We generated synthetic data sets from a Gaussian mixture model $\nu^\star * \Phi$ with $\nu^\star := \sum_{j=1}^{k^\star} w_j^\star \delta_{\theta_j^\star}$. We consider the following four different cases of the true mixing distribution.

Case 1 (Well-separated) $\theta^\star = (-3, -1, 1, 3)$, $w^\star = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
Case 2 (Overlapped components) $\theta^\star = (-1.5, -1, 1, 3)$, $w^\star = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
Case 3 (Weak component) $\theta^\star = (-3, -1, 1, 3)$, $w^\star = (\frac{2}{5}, \frac{1}{10}, \frac{1}{4}, \frac{1}{4})$
Case 4 (Higher-order) $\theta^\star = (-6, -4, -2, 0, 2, 4, 6)$, $w^\star = (\frac{1}{7}, \ldots, \frac{1}{7})$

For Case 1, all the true components are well-separated. The true components from Case 2 and Case 3 are not well-separated. In Case 2, there are two close atoms and in Case 3, there is a weak component. Case 4 is a higher-order mixture setup. For each setup, we let the sample size n range over $\{250, 500, \ldots, 2000\}$. We repeat this data generation 50 times for each experiment and report the average of the first order Wasserstein distance between each estimator and the true mixing distribution.

Figure 1 displays the average of the the first order Wasserstein errors of the five estimators for the four cases of the data generating process. Contrary to its theoretical optimality, `DMM` performs the worst among the five estimators for all the scenarios. The performance gap of DMM to the Bayesian



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

Figure 1. The average (curve) and the standard deviation (band) of the first-order Wasserstein errors of five estimators by sample size.

methods are the largest for Case 4. We observed that there is numerical instability of the `DMM` implementation when the number of components is larger than 5, which results in failure of computation. Thus we fixed the number of components as 4 instead of 7. This leads to the poor performance of the method. The overall result seems to inconsistent to the simulation result of the `DMM` paper [49], which showed better or at least competitive performance of `DMM` compared to the MAP estimators. But the simulation setup is different. The authors of [49] considered the two simulation scenarios, the first case where the true five number of components are very close to each other (indeed, $\theta^\star = (-0.236, -0.168, -0.987, 0.299, 0.150)$ and $w^\star = (0.123, 0.552, 0.010, 0.080, 0.235)$) so that the corresponding mixture density seems to be unimodal, and the second case where there are only two components of the true mixing distribution. Thus, we conjecture that `DMM` performs worse for mixture densities with many modes. However, as one of the reviewers pointed out, this conjecture could be potentially ungrounded. Another possible reason behind this is numerical instability. Since moments typically span several orders of magnitude, semidefinite programming in DMM procedures can suffer from numerical instabilities (see [1]). This could be the reason behind the relatively poor performance of DMM estimator.

Generally, all the four Bayesian methods performs almost similar. For Case 1, the over-fitted mixture model `MAP_over` performs worse than the other Bayesian methods, but does similar for the other three cases. For Case 2 and Case 3, `MFM_vary` tends to select the smaller mixture than the true mixture, in general, its posterior distribution is maximized at $k = 3$ which is less than the true one $k^\star = 4$. Note that

this does not contradict our theoretical results where we establish the consistent estimation of the number of well-separated components, which might be equal to 3 in these two cases. This leads to slightly better performance for Case 2 where overlapped components exist and slightly worse performance for Case 3 where weak components exist. Overall, knowing the true number of components does not give substantial improvement of empirical performance. Our theory for optimal estimation of the mixing distribution is built on the result of vanishing posterior probability of overestimation of the number of components (Theorem 2.1). It would be interesting to investigate the optimality of Bayesian posteriors that do not enjoy such asymptotic property, for instance, the overfitted Bayesian mixture model.

4.1.2. Inference for the number of components

In this experiment, we assess the performance of the proposed Bayesian procedure and the DP mixture model with sample size dependent hyperparameters. We generated the Gaussian mixture with atoms $(-2, 0, 2)$ and equal weights $(1/3, 1/3, 1/3)$. Five independent data sets are generated from this Gaussian mixture model for each sample size $n \in \{50, 100, 250, 1000, 2500\}$. We compare four Bayesian methods: the two MFM models with Poisson mean parameter $\lambda_n = 10 \exp\left(-\frac{1}{5}\frac{\log^2 n}{\log\log n}\right)$ (`MFM_vary`) and $\lambda_n = 1$ (`MFM_const`) and the two DP mixtures models with concentration parameter $\kappa_n = 20/n$ (`DP_vary`) and $\kappa_n = 0.4$ (`DP_const`). We use the uniform distribution on $[-6, 6]$ for both the prior on the atoms for the MFM and the base distribution for the DP mixture.

For posterior computation for the MFM models, we employ the reversible jump MCMC algorithm of [41]. For each posterior computation, we ran a single Markov chain with length 105,000. We saved every 100-th sample after a burn-in period of 5,000 samples. On the other hand, we use Neal's Algorithm 8 [34] for non-conjugate priors to compute the posterior distributions of the DP mixtures.

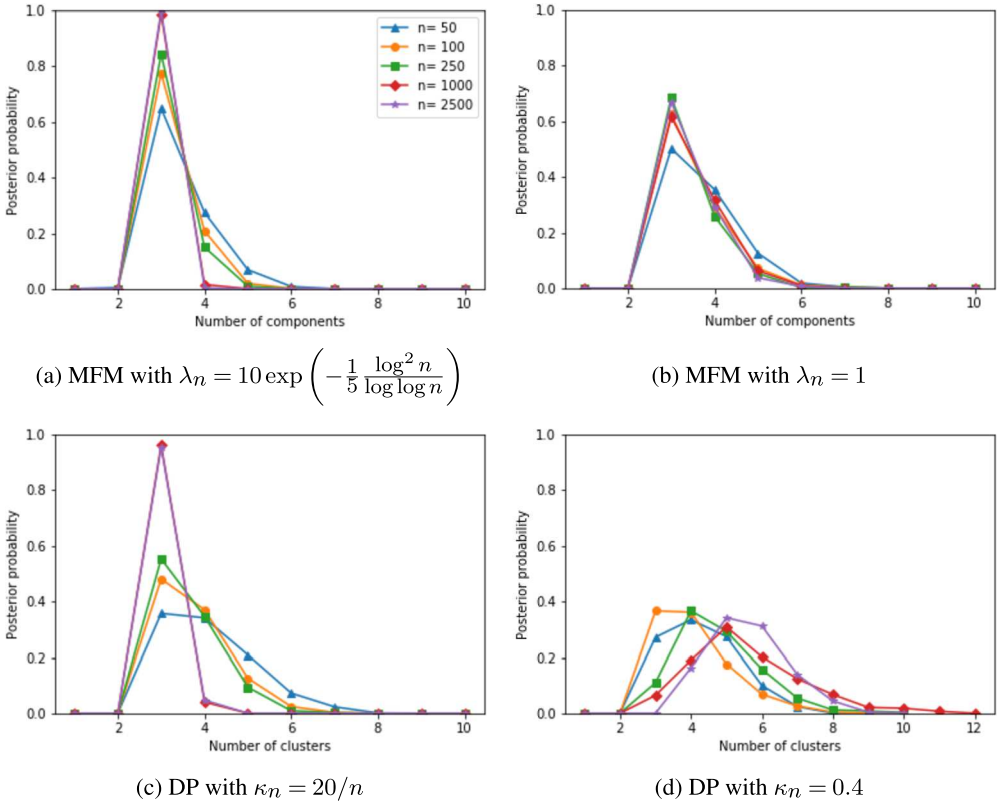
Figure 2 presents the posterior distributions of the number of components for the two MFMs and of the number of cluster for the two DP mixtures, respectively. It clearly shows that the diminishing choices of hyperparameter advocated by our theory outperforms the constant counterparts. It is worth to notice that the posterior distribution of `DP_vary` captures the true number of components well for large samples. It is a widely observed that the DP mixture tends to produce redundant clusters, in particular, Miller and Harrison [32] and Guha, Ho and Nguyen [19] observed this phenomenon in their simulation studies, however our simulation shows that a sample size dependent concentration parameter inversely related to the sample size can circumvent this issue.

4.2. Real data analysis

4.2.1. Galaxy data

In this section, we consider an application to the galaxy data of Roeder [42] which record velocity measurements (1,000 Km/sec) of 82 galaxies from the Corona Borealis region. This data set has been widely used as a benchmark for mixture modelling methods, e.g., [8,10,37,47]

To gain flexibility, we used the Gaussian location-scale mixture model instead of the Gaussian location mixture model that we have focused on. We considered the MFM and DP prior distributions. The MFM prior consists of $\text{POISSON}(\lambda)$ on the number of components, $\text{DIR}(1, \ldots, 1)$ on the mixing weights, $\text{UNIF}([0, 40])$ on the location parameter and $\text{GAMMA}(1, 1)$ on the scale parameter. We fitted the model with the MFM prior with five different values of the mean parameter λ of the Poisson distribution: $3, 1, 0.5, 0.1$ and 0.01. The base distribution of the DP for the location parameter is chosen to be the same as the MFM, but we choose the inverse gamma distribution in order to employ conjugacy. We fitted the DP mixture model with five different values of the concentration parameter κ, which are the same as λ.



(a) MFM with $\lambda_n = 10 \exp\left(-\frac{1}{5}\frac{\log^2 n}{\log\log n}\right)$

(b) MFM with $\lambda_n = 1$

(c) DP with $\kappa_n = 20/n$

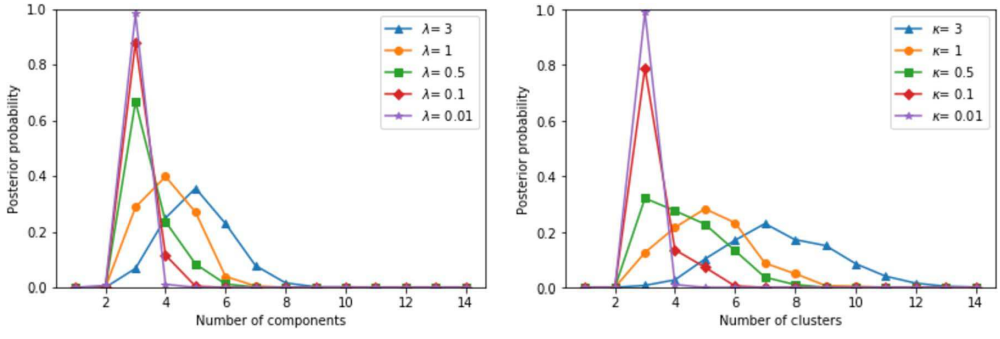
(d) DP with $\kappa_n = 0.4$

Figure 2. Posterior distribution of the number of components for the MFM and of the number of clusters for the DP mixture. The true number of components is 3.

Figure 3 presents a posterior distribution of the number of components for the MFM and the one of the number of clusters for the DP mixture. The shape of the posterior distribution is substantially different by the choices of the hyperparameters λ for the MFM and κ for the DP. The posterior distribution of the number of clusters for the DP mixture model with a small concentration parameter concentrates near the value of 3, as the posterior distribution of the number of components for the MFM does. This result implies that the DP mixture can be used as a proxy of the MFM for estimating the number of components when the small concentration parameter is used, as our theory suggests. Figure 4 displays posterior predictive densities for the MFM and DP mixture as well as histogram of the galaxy data. For density estimation, it also seems that the choice of the hyperparameters is more critical than the choice among the MFM and DP mixture.

4.2.2. Old faithful geyser eruption data

In this section, we consider the Old faithful geyser eruption data which consists of two measurements, duration time and waiting time to the next eruption, taken on $n = 272$ eruptions for the Old Faithful geyser in Yellowstone National Park. We analyzed the data using the multivariate Gaussian location-scale mixture model $\sum_{j=1}^{k} w_j \mathrm{N}(\theta_j, \Sigma_j)$ where $(w_1, \ldots, w_k) \in \Delta_k$ are weights, $\theta_1, \ldots, \theta_k$ are 2-dimensional vectors and $\Sigma_1, \ldots, \Sigma_k$ are 2×2 symmetric positive definite matrices. We first standardized the data and imposed the following MFM prior distribution: $k \sim \mathrm{POISSON}(\lambda)$, $(w_1, \ldots, w_k) \sim$



(a) MFM (b) DP

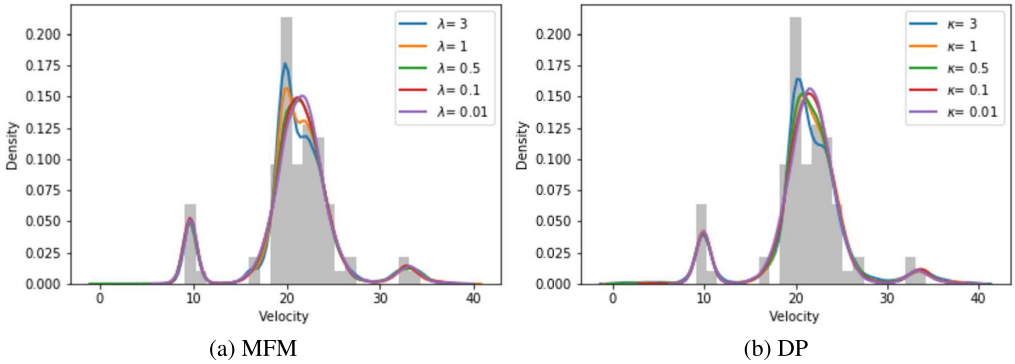
Figure 3. Posterior distribution of the number of components for the MFM and of the number of clusters for the DP mixture for the galaxy data

$\text{DIR}(1,\ldots,1)$, $\theta_j \overset{\text{iid}}{\sim} \text{UNIF}([-2,2]^2)$ and $\Sigma_j \overset{\text{iid}}{\sim} \text{WISHART}(5, \frac{1}{10}I)$, where $I \in \mathbb{R}^{2\times 2}$ denotes the identity matrix. We considered four different values of the mean parameter λ of the Poisson distribution: $1, 0.1, 0.01$ and 0.001.

Figure 5 displays the posterior distribution of the number of components by the value of λ. We see that the smaller the parameter λ, the more the posterior distribution concentrates on the value of 2, which seems to be enough to explain the data, see the scatter plot of the data in Figure 6. This result implies that our sample size dependent prior distribution may work even for the multivariate Gaussian location-scale mixture model which is much more complex than the univariate Gaussian location mixture model. Figure 6 presents the posterior predictive density by the value of λ. There is not much difference in four posterior predictive density estimates.

Acknowledgements

We thank the Editor, the Associate Editor and two reviewers for their valuable comments. We are in particular grateful to the two reviewers for their careful going over the technical proof of our theorems. We would like to thank Minwoo Chae for very useful comments and discussions.



(a) MFM (b) DP

Figure 4. Histogram of the galaxy data with posterior predictive density estimates

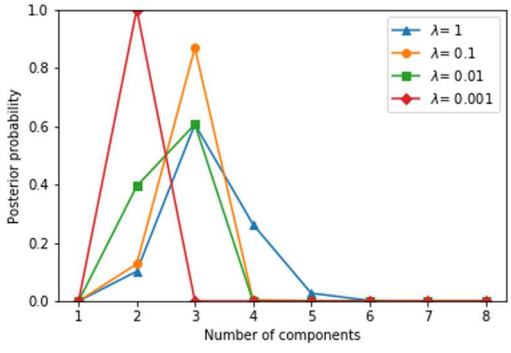
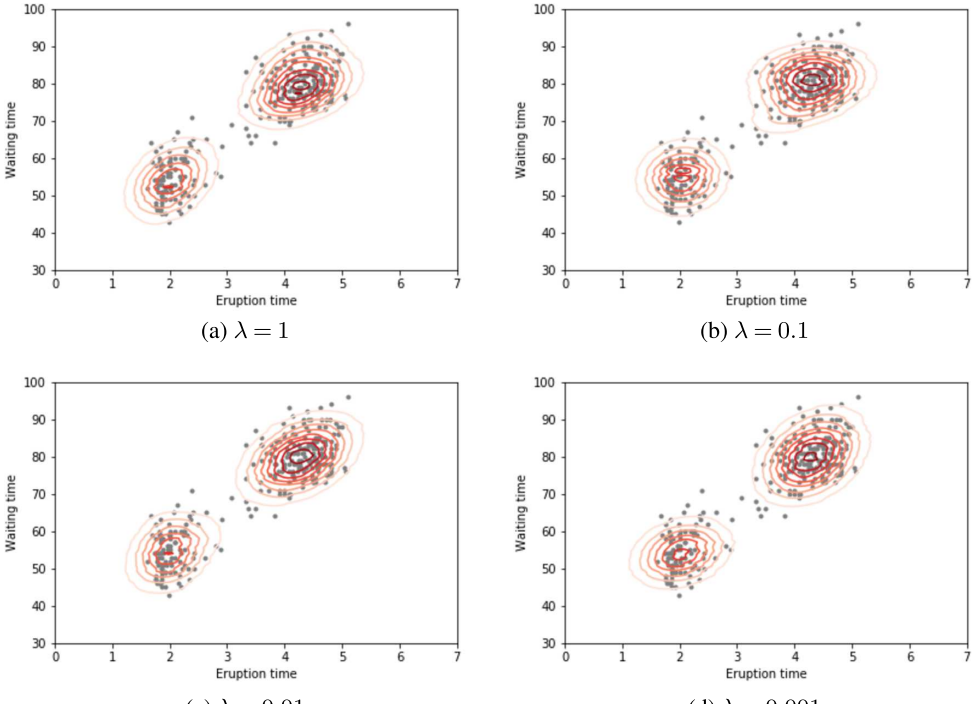


Figure 5. Posterior distribution of the number of components for the old Faithful geyser eruption data



(a) $\lambda = 1$ (b) $\lambda = 0.1$

(c) $\lambda = 0.01$ (d) $\lambda = 0.001$

Figure 6. Scatterplot of the old Faithful geyser eruption data with contour plot of the posterior predictive density

Funding

We acknowledge the generous support of NSF grants DMS CAREER 1654579 and DMS 2113642.

Supplementary Material

Supplement to "Optimal Bayesian estimation of Gaussian mixtures with growing number of components" (DOI: 10.3150/22-BEJ1495SUPP; .pdf). The proofs of all the Theorems are contained in Appendix A of Ohn and Lin [39]. Analysis of general mixture models in the framework of Heinrich and Kahn [20] is provided in Appendix B of Ohn and Lin [39].

References

[1] Backenköhler, M., Bortolussi, L. and Wolf, V. (2020). Bounding mean first passage times in population continuous-time Markov chains. In *International Conference on Quantitative Evaluation of Systems* 155–174. Springer.

[2] Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.

[3] Bing, X., Bunea, F. and Wegkamp, M. (2020). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli* **26** 1765–1796. MR4091091 https://doi.org/10.3150/19-BEJ1166

[4] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

[5] Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077 https://doi.org/10.1214/12-AOS1029

[6] Chambaz, A. and Rousseau, J. (2008). Bounds for Bayesian order identification with application to mixtures. *Ann. Statist.* **36** 938–962. MR2396820 https://doi.org/10.1214/009053607000000857

[7] Chen, J.H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. MR1331665 https://doi.org/10.1214/aos/1176324464

[8] Drton, M. and Plummer, M. (2017). A Bayesian information criterion for singular models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 323–380. MR3611750 https://doi.org/10.1111/rssb.12187

[9] Eghbal-zadeh, H., Zellinger, W. and Widmer, G. (2019). Mixture density generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 5820–5829.

[10] Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

[11] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

[12] Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635 https://doi.org/10.1198/016214502760047131

[13] Frühwirth-Schnatter, S., Celeux, G. and Robert, C.P., eds. (2019). *Handbook of Mixture Analysis*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. Boca Raton, FL: CRC Press. MR3889980

[14] Gao, C., van der Vaart, A.W. and Zhou, H.H. (2020). A general framework for Bayes structured linear models. *Ann. Statist.* **48** 2848–2878. MR4152123 https://doi.org/10.1214/19-AOS1909

[15] Gao, C. and Zhou, H.H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. MR3449770 https://doi.org/10.1214/15-AOS1368

[16] Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864 https://doi.org/10.1214/009053606000001271

[17] Ghosal, S. and van der Vaart, A.W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329 https://doi.org/10.1214/aos/1013203453

[18] Greggio, N., Bernardino, A., Laschi, C., Dario, P. and Santos-Victor, J. (2012). Fast estimation of Gaussian mixture models for image segmentation. *Machine Vision and Applications* **23** 773–789.

[19] Guha, A., Ho, N. and Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli* **27** 2159–2188. MR4303879 https://doi.org/10.3150/20-BEJ1275

[20] Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Ann. Statist.* **46** 2844–2870. MR3851757 https://doi.org/10.1214/17-AOS1641

[21] Ho, N. and Nguyen, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electron. J. Stat.* **10** 271–307. MR3466183 https://doi.org/10.1214/16-EJS1105
[22] Ho, N., Nguyen, X. and Ritov, Y. (2020). Robust estimation of mixing measures in finite mixture models. *Bernoulli* **26** 828–857. MR4058353 https://doi.org/10.3150/18-BEJ1087
[23] Hoffmann, M., Rousseau, J. and Schmidt-Hieber, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985 https://doi.org/10.1214/15-AOS1341
[24] Jiang, S. and Tokdar, S.T. (2021). Variable selection consistency of Gaussian process regression. *Ann. Statist.* **49** 2491–2505. MR4338372 https://doi.org/10.1214/20-aos2043
[25] Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62** 49–66. MR1769735
[26] Kruijer, W., Rousseau, J. and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885 https://doi.org/10.1214/10-EJS584
[27] Martin, R. (2012). Convergence rate for predictive recursion estimation of finite mixtures. *Statist. Probab. Lett.* **82** 378–384. MR2875226 https://doi.org/10.1016/j.spl.2011.10.023
[28] Martin, R., Mess, R. and Walker, S.G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23** 1822–1847. MR3624879 https://doi.org/10.3150/15-BEJ797
[29] McLachlan, G.J., Lee, S.X. and Rathnayake, S.I. (2019). Finite mixture models. *Annu. Rev. Stat. Appl.* **6** 355–378. MR3939525 https://doi.org/10.1146/annurev-statistics-031017-100325
[30] Miller, J.W. and Harrison, M.T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems* 199–206.
[31] Miller, J.W. and Harrison, M.T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.* **15** 3333–3370. MR3277163
[32] Miller, J.W. and Harrison, M.T. (2018). Mixture models with a prior on the number of components. *J. Amer. Statist. Assoc.* **113** 340–356. MR3803469 https://doi.org/10.1080/01621459.2016.1255636
[33] Morris, C.N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10** 65–80. MR0642719
[34] Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653
[35] Newton, M.A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhya, Ser. A* **64** 306–322. MR1981761
[36] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422 https://doi.org/10.1214/12-AOS1065
[37] Nobile, A. and Fearnside, A.T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Stat. Comput.* **17** 147–162. MR2380643 https://doi.org/10.1007/s11222-006-9014-7
[38] Ohn, I. and Kim, Y. (2021). Posterior consistency of factor dimensionality in high-dimensional sparse factor models. *Bayesian Anal.* **1** 1–24.
[39] Ohn, I., Lin, L. (2023). Supplement to "Optimal Bayesian estimation of Gaussian mixtures with growing number of components." https://doi.org/10.3150/22-BEJ1495SUPP
[40] Richardson, E. and Weiss, Y. (2018). On GANs and GMMs. In *Advances in Neural Information Processing Systems* 5847–5858.
[41] Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095
[42] Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* **85** 617–624.
[43] Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. MR2867454 https://doi.org/10.1111/j.1467-9868.2011.00781.x
[44] Scricciolo, C. (2011). Posterior rates of convergence for Dirichlet mixtures of exponential power densities. *Electron. J. Stat.* **5** 270–308. MR2802044 https://doi.org/10.1214/11-EJS604
[45] Scricciolo, C. (2019). Bayesian Kantorovich deconvolution in finite mixture models. In *New Statistical Developments in Data Science*. *Springer Proc. Math. Stat.* **288** 119–134. Cham: Springer. MR4008311 https://doi.org/10.1007/978-3-030-21158-5_1

[46] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433
[47] Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. MR1762903 https://doi.org/10.1214/aos/1016120364
[48] Tokdar, S.T., Martin, R. and Ghosh, J.K. (2009). Consistency of a recursive estimate of mixing distributions. *Ann. Statist.* **37** 2502–2522. MR2543700 https://doi.org/10.1214/08-AOS639
[49] Wu, Y. and Yang, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. *Ann. Statist.* **48** 1981–2007. MR4134783 https://doi.org/10.1214/19-AOS1873

Received January 2021 and revised March 2022