Bernoulli **27**(4), 2021, 2159–2188
https://doi.org/10.3150/20-BEJ1275

On posterior contraction of parameters and interpretability in Bayesian mixture modeling

ARITRA GUHA[1], NHAT HO[2] and XUANLONG NGUYEN[3]

[1]*Department of Statistical Science, Duke University, Durham, NC, USA. E-mail: aritra.guha@duke.edu*
[2]*Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX, USA. E-mail: minhnhat@utexas.edu*
[3]*Department of Statistics, University of Michigan, Ann Arbor, MI, USA. E-mail: xuanlong@umich.edu*

We study posterior contraction behaviors for parameters of interest in the context of Bayesian mixture modeling, where the number of mixing components is unknown while the model itself may or may not be correctly specified. Two representative types of prior specification will be considered: one requires explicitly a prior distribution on the number of mixture components, while the other places a nonparametric prior on the space of mixing distributions. The former is shown to yield an optimal rate of posterior contraction on the model parameters under minimal conditions, while the latter can be utilized to consistently recover the unknown number of mixture components, with the help of a fast probabilistic post-processing procedure. We then turn the study of these Bayesian procedures to the realistic settings of model misspecification. It will be shown that the modeling choice of kernel density functions plays perhaps the most impactful roles in determining the posterior contraction rates in the misspecified situations. Drawing on concrete posterior contraction rates established in this paper we wish to highlight some aspects about the interesting tradeoffs between model expressiveness and interpretability that a statistical modeler must negotiate in the rich world of mixture modeling.

Keywords: Mixture models; Wasserstein distance; Bayesian nonparametrics; Bayesian inference; misspecified models; post-processing algorithm

1. Introduction

Mixture models are one of the most useful tools in a statistician's toolbox for analyzing heterogeneous data populations. They can be a powerful black-box modeling device to approximate the most complex forms of density functions. Perhaps more importantly, they help the statistician express the data population's heterogeneous patterns and interpret them in a useful way [30,33,34]. The following are common, generic and meaningful questions a practitioner of mixture modeling may ask: (I) how many mixture components are needed to express the underlying latent subpopulations, (II) how efficiently can one estimate the parameters representing these components and, (III) what happens to a mixture model based statistical procedure when the model is actually misspecified?

How to determine the number of mixture components is a question that has long fascinated mixture modelers. Many proposed solutions approached this as a model selection problem. The number of model parameters, hence the number of mixture components, may be selected by optimizing with respect to some regularized loss function; see, for example, [6,26,30] and the references therein. A Bayesian approach to regularization is to place explicitly a prior distribution on the number of mixture components, for example, [36,39–41]. A convenient aspect of separating out the modeling and inference questions considered in (I) and (II) is that once the number of parameters is determined, the model parameters concerned by question (II) can be estimated and assessed via any standard parametric estimation methods.

In a number of modern applications of mixture modeling to heterogeneous data, such as in topic modeling, the number of mixture components (the topics) may be very large and not necessarily a

1350-7265

meaningful quantity [3,49]. In such situations, it may be appealing for the modeler to consider a nonparametric approach, where both (I) and (II) are considered concurrently. The object of inference is now the mixing measure which encapsulates all unknowns about the mixture density function. There were numerous works exemplifying this approach [11,24,29]. In particular, the field of Bayesian nonparametrics (BNP) has offered a wealth of prior distributions on the mixing measure based on which one can arrive at the posterior distribution of any quantity of interest related to the mixing measure [21].

A common choice of such priors is the Dirichlet process [2,10,45], resulting in the famous Dirichlet process mixture models [1,8,31]. Dirichlet process (DP) and its variants have also been adopted as a building block for more sophisticated hierarchical modeling, thanks to the ease with which computational procedures for posterior inference via Markov Chain Monte Carlo can be implemented [42,50]. Moreover, there is a well-established asymptotic theory on how such Bayesian nonparametric mixture models result in asymptotically optimal estimation procedures for the population density. See, for instance, [14,16,46] for theoretical results specifically on DP mixtures, and [15,47,52] for general BNP models. The rich development in both algorithms and theory in the past decades has contributed to the widespread adoption of these models in a vast array of application domains.

For some time there was a misconception among quite a few practitioners in various application domains, a misconception that may have initially contributed to their enthusiasm for Bayesian nonparametric modeling, that the use of such nonparametric models eliminates altogether the need for determining the number of mixture components, because the learning of such a quantity is “automatic” from the posterior samples of the mixing measure. The implicit presumption here is that a consistent estimate of the mixing measure may be equated with a consistent estimate of the number of mixture components. This is not correct, as has been noted, for instance, by [29] in the context of mixing measure estimation. More recently, [35] explicitly demonstrated that the common practice of drawing inference about the number of mixture components via the DP mixture, specifically by reading off the number of support points in the Dirichlet’s posterior sample, leads to an asymptotically inconsistent estimate.

Despite this inconsistency result, it will be shown in this paper that it is still possible to obtain a consistent estimate of the number of mixture components using samples from a Dirichlet process mixture, or any Bayesian nonparametric mixture, by applying a simple and fast post-processing procedure on samples drawn from the DP mixture’s posterior. On the other hand, the parametric approach of placing an explicit prior on the number of components yields both a consistent estimate of the number of mixture components, and more notably, an optimal posterior contraction rate for component parameters, under a minimal set of conditions. It is worth emphasizing that all these results are possible only under the assumption that the model is well-specified, that is, the true but unknown population density lies in the support of the induced prior distribution on the mixture densities.

As George Box has said, “all models are wrong”, but more relevant to us, all mixture models are misspecified in some way. The statistician has a number of modeling decisions to make when it comes to mixture models, including the selection of the class of kernel densities, and the support of the space of mixing measures. The significance of question (III) comes to the fore, because if the posterior contraction behavior of model parameters is very slow due to specific modeling choices, one has to be cautious about the interpretability of the parameters of interest. A very slow posterior contraction rate in theory implies that a given data set probably has relatively very slow influence on the movement of mass from the prior to the posterior distribution.

In this paper we study Bayesian estimation of model parameters with both well-specified and misspecified mixture models. There are two sets of results. The first results resolve several outstanding gaps that remain in the existing theory and current practice of Bayesian parameter estimation, given that the mixture model is well-specified. The second set of results describes posterior contraction properties of such procedures when the mixture model is misspecified. We proceed to describe these results, related works and implications to the mixture modeling practice.

1.1. Well-specified regimes

Consider discrete mixing measures $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$. Here, $\boldsymbol{p} = (p_1, \ldots, p_k)$ is a vector of mixing weights, while atoms $\{\theta_i\}_{i=1}^{k}$ are elements in a given compact space $\Theta \in \mathbb{R}^d$. Mixing measure G is combined with a likelihood function $f(\cdot|\theta)$ with respect to Lebesgue measure μ to yield a mixture density: $p_G(\cdot) = \int f(\cdot|\theta)\, dG(\theta) = \sum_{i=1}^{k} p_i f(\cdot|\theta_i)$. When $k < \infty$, we call this a *finite mixture model* with k components. We write $k = \infty$ to denote an *infinite mixture model*. The atoms θ_i's are representatives of the underlying subpopulations.

Assume that $X_1, \ldots, X_n$ are i.i.d. samples from a mixture density $p_{G_0}(x) = \int f(x|\theta)\, dG_0(\theta)$, where G_0 is a discrete mixing measure with *unknown* number of support points $k_0 < \infty$ residing in Θ. In the overfitted setting, that is, an upper bound $k_0 \leq \overline{k}$ is given so that one may work with an overfitted mixture with $\overline{k}$ mixture components, Chen [5] showed that the mixing measure G_0 can be estimated at a rate $n^{-1/4}$ under the L_1 metric, provided that the kernel f satisfies a second-order identifiability condition – this is a linear independence property on the collection of kernel function f and its first and second order derivatives with respect to θ.

Asymptotic analysis of Bayesian estimation of the mixing measure that arises in both finite and infinite mixtures, where the convergence is assessed under Wasserstein distance metrics, was first investigated by Nguyen [37]. Convergence rates of the mixing measure under a Wasserstein distance can be directly translated to the convergence rates of the parameters in the mixture model. Under the same (second-order) identifiability condition, it can be shown that either maximum likelihood estimation method or a Bayesian method with a non-informative (e.g., uniform) prior yields a $(\log n/n)^{1/4}$ rate of convergence [22,24,37]. Note, however, that $n^{-1/4}$ is not the optimal *pointwise* rate of convergence. Heinrich and Kahn [20] showed that a distance based estimation method can achieve $n^{-1/2}$ rate of convergence under W_1 metric, even though their method may not be easy to implement in practice. [23] described a minimum Hellinger distance estimator that achieves the same optimal rate of parameter estimation.

An important question in Bayesian analysis is whether there exists a suitable prior specification for mixture models according to which the posterior distribution on the mixing measure can be shown to contract toward the true mixing measure at the same fast rate $n^{-1/2}$. Rousseau and Mengersen [43] provided an interesting result in this regard, which states that for overfitted mixtures with a suitable Dirichlet prior on the mixing weights $\boldsymbol{p}$, assuming that an upper bound to the number of mixture component is given, in addition to a second-order type identifiability condition, then the posterior contraction to the true mixing measure can be established by the fact that the mixing weights associated with all redundant atoms of mixing measure G vanish at the rate close to the optimal $n^{-1/2}$.

In our first main result given in Theorem 3.1, we show that an alternative and relatively common choice of prior also yields optimal rates of convergence of the mixing measure (up to a logarithmic term), in addition to correctly recovering the number of mixture components, under considerably weaker conditions. In particular, we study the mixture of finite mixture (MFM) prior, which places an explicit prior distribution on the number of components k and a (conditional) Dirichlet prior on the weights $\boldsymbol{p}$, given each value of k. This prior has been investigated by Miller and Harrison [36]. Compared to the method of [43], no upper bound on the true number of mixture components is needed. In addition, only first-order identifiability condition is required for the kernel density f, allowing our results to apply to popular mixture models such as location-scale Gaussian mixtures. We also note that the MFM prior is one instance in a class of modeling proposals, eg, [39–41] for which the established convergence behavior continues to hold. In other words, from an asymptotic standpoint, all is good on the parametric Bayesian front.

Our second main result, given in Theorem 3.2, is concerned with a Bayesian nonparametric modeling practice. A Bayesian nonparametric prior on mixing measures places zero mass on measures with finite support points, so the BNP model is misspecified with respect to the number of mixture components. Indeed, when G_0 has only finite support the true density p_{G_0} lies at the boundary of the support of the class of densities produced by the BNP prior. Despite the inconsistency results mentioned earlier on the number of mixture components produced by Dirichlet process mixtures, we will show that this situation can be easily corrected by applying a post-processing procedure to the samples generated from the posterior distribution arising from the DP mixtures, or any sufficiently well-behaved Bayesian nonparametric mixture models. By "well-behaved" we mean any BNP mixtures under which the posterior contraction rate on the mixing measure can be guaranteed by an upper bound using a Wasserstein metric [37].

Our post-processing procedure is simple, and motivated by the observation that a posterior sample of the mixing measure tends to produce a large number of atoms with very small and vanishing weights [18,35]. Such atoms can be ignored by a suitable truncation procedure. In addition, similar atoms in the metric space Θ can also be merged in a systematic and probabilistic way. Our procedure, named Merge-Truncate-Merge algorithm, is guaranteed to not only produce a consistent estimate of the number of mixture components but also retain the posterior contraction rates of the original posterior samples for the mixing measure. Theorem 3.2 provides a theoretical basis for the heuristics employed in practice in dealing with mixtures with unknown number of components [18,40].

1.2. Misspecified regimes

There are several ways a mixture model can be misspecified: either in the kernel density function f, or the mixing measure G, or both. Thus, in the misspecified setting, we assume that the data samples $X_1, \dots, X_n$ are i.i.d. samples from a mixture density p_{G_0,f_0}, namely, $p_{G_0,f_0}(x) = \int f_0(x|\theta)G_0(d\theta)$, where both G_0 and f_0 are unknown. The statistician draws inference from a mixture model $p_{G,f}$, still denoted by p_G for short, where G is a mixing measure with support on compact Θ, and f is a chosen kernel density function. In particular, a Bayesian procedure proceeds by placing a prior on the mixing measure G and obtaining the posterior distribution on G given the n-data sample. In general, the true data generating density p_{G_0} lies outside the support of the induced prior on p_G. We study the posterior behavior of G as the sample size n tends to infinity.

The behavior of Bayesian procedures under model misspecification has been investigated in the foundational work of [27,28]. These papers focus primarily on density estimation. In particular, assuming that the true data generating distribution's density lies outside the support of a Bayesian prior, then the posterior distribution on the model density can be shown to contract to an element of the prior's support, which is obtained by a Kullback–Leibler (KL) projection of the true density into the prior's support [27].

It can be established that the posterior of p_G contracts to a density p_{G_*}, where G_* is a probability measure on Θ such that p_{G_*} is the (unique) minimizer of the Kullback–Leilber divergence $K(p_{G_0,f_0}, p_G)$ among all probability measure G on Θ. This mere fact is readily deduced from the theory of [27], but the outstanding and relevant issue is whether the posterior contraction behavior carries over to that of G, and if so, at what rate. In general, G_* may not be unique, so posterior contraction of G cannot be established. Under identifiability, G_* is unique, but still $G_* \neq G_0$.

This leads to the question about interpretability when the model is misspecified. Specifically, when $f \neq f_0$, it may be unclear how one can interpret the parameters that represent mixing measure G, unless f can be assumed to be a reasonable approximation of f_0. Mixing measure G, too, may be misspecified, when the true support of G_0 may not lie entirely in Θ. In practice, it is a perennial challenge to explicate the relationship between G_* and the unknown G_0. In theory, it is mathematically an interesting question to characterize this relationship, if some assumption can be made on the true

G_0 and f_0, but this is beyond the scope of this paper. Regardless of the truth about this relationship, it is important for the statistician to know how impactful a particular modeling choice on f and G can affect the posterior contraction rates of the parameters of interest.

The main results that we shall present in Theorem 4.1 and Theorem 4.2 are on the posterior contraction rates of the mixing measure G toward the limit point G_*, under very mild conditions on the misspecification of f. In particular, we shall require that the tail behavior of function f is not much heavier than that of f_0 (cf. condition (P.5) or (P.5') in Section 4). Specific posterior contraction rates of contraction for G are derived when f is either Gaussian or Laplace density kernel, two representatives for supersmooth and ordinary smooth classes of kernel densities [9]. A key step in our proofs lies in several inequalities which provide upper bound of Wasserstein distances on mixing measures in terms of weighted Hellinger distances, a quantity that plays a fundamental role in the asymptotic characterization of misspecified Bayesian models [27].

It is interesting to highlight that the posterior contraction rate for the misspecified Gaussian location mixture is the same as that of well-specified setting, which is nonetheless extremely slow, in the order of $(1/\log n)^{1/2}$. On the other hand, using a misspecified Laplace location mixture results in some loss in the exponent γ of the polynomial rate $n^{-\gamma}$. Although the contrast in contraction rates for the two families of kernels is quite similar to what is obtained for well-specified deconvolution problems for both frequentist methods [9,55] and Bayesian methods [13,37], our results are given for misspecified models, which can be seen in a new light: since the model is misspecified anyway, the statistician should be "free" to choose the kernel that can yield the most favorable posterior contraction for the parameters of his/ her model. In that regard, Laplace location mixtures may be preferred to Gaussian location mixtures, provided that the limit G_* is not too far from the true G_0. When this is not the case, that is, when the bias of the misspecified model is too large due to the use of Laplace mixtures, it is more advisable to adopt Gaussian kernels instead, despite the latter's lagging posterior contraction behavior. Although it is quite clear that the ultimate model choice under misspecification will reside on resolving the tension between aforementioned bias and contracting variance, a satisfactory formulation and solution for such a model choice problem which accounts for parameter estimation and interpretability remains an interesting and important open question.

Additionally, we note that the relatively slow posterior contraction rate for G is due to the fact that the limiting measure G_* in general may have infinite support, regardless of whether the true G_0 has finite support or not. From a practical standpoint, it is difficult to interpret the estimate of G if G_* has infinite support. However, if G_* happens to have a finite number of support points, which is bounded by a known constant, say $\overline{k}$, then by placing a suitable prior on G to reflect this knowledge we show that the posterior of G contracts to G_* at a relatively fast rate $(\log n/n)^{1/4}$. This is the same rate obtained under the well-identified setting for overfitted mixtures.

1.3. Further remarks

The posterior contraction theorems in this paper provide an opportunity to re-examine several aspects of the fascinating picture about the tension between a model's expressiveness and its interpretability. They remind us once again about the tradeoffs a modeler must negotiate for a given inferential goal and the information available at hand. We enumerate a few such insights:

(1) "One size does not fit all": Even though the family of mixture models as a whole can be excellent at inferring about population heterogeneity and at density estimation as a black-box device, a specific mixture model specification cannot do a good job at both. For instance, a Dirichlet process mixture of Gaussian kernels may yield an asymptotically optimal density estimation machine but it performs poorly when it comes to learning of parameters.

(2) "Finite versus infinite": If the number of mixture components is known to be small and an object of interest, then employing an explicit prior on this quantity results in the optimal posterior contraction rate for the model parameters and thus is a preferred method. When this quantity is known to be high or not a meaningful object of inference, Bayesian nonparametric mixtures provide a more attractive alternative as it can flexibly adapt to complex forms of densities. Regardless, one can still consistently recover the true number of mixture components using a nonparametric approach.

(3) "Some forms of misspecification are more useful than others". When the mixture model is misspecified, careful design choices regarding the (mispecified) kernel density and the support of the mixing measure can significantly speed up the posterior contraction behavior of model parameters. For instance, a heavy-tailed and ordinary smooth kernel such as the Laplace, instead of the Gaussian kernel, is shown to be especially amenable to efficient parameter estimation.

The remainder of the paper is organized as follows. Section 2 provides necessary backgrounds about mixture models, Wasserstein distances and several key notions of strong identifiability. Section 3 presents posterior contraction theorems for well-misspecified mixture models for both parametric and nonparametric Bayesian models. Section 4 presents posterior contraction theorems when the mixture model is misspecified. In Section 5, we provide illustrations of the Merge-Truncate-Merge algorithm via a simulation study. Proofs of technical results are provided in the supplementary material [19].

Notation. Given two densities p, q (with respect to the Lebesgue measure μ), the total variation distance is given by $V(p,q) = (1/2)\int |p(x) - q(x)|\, d\mu(x)$. Additionally, the squared Hellinger distance is given by $h^2(p,q) = (1/2)\int (\sqrt{p(x)} - \sqrt{q(x)})^2\, d\mu(x)$. Furthermore, the Kullback–Leibler (KL) divergence is given by $K(p,q) = \int \log(p(x)/q(x)) p(x)\, d\mu(x)$ and the squared KL divergence is given by $K_2(p,q) = \int \log(p(x)/q(x))^2 p(x)\, d\mu(x)$. For a measurable function f, let Qf denote the integral $\int f\, dQ$. For any $\kappa = (\kappa_1, \dots, \kappa_d) \in \mathbb{N}^d$, we denote $\frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta) = \frac{\partial^{|\kappa|} f}{\partial \theta_1^{\kappa_1} \dots \partial \theta_d^{\kappa_d}}(x|\theta)$ where $\theta = (\theta_1, \dots, \theta_d)$. For any metric d on Θ, we define the open ball of d-radius ϵ around $\theta_0 \in \Theta$ as $B_d(\epsilon, \theta_0)$. We use $D(\epsilon, \Omega, \tilde{d})$ to denote the maximal ϵ-packing number for a general set Ω under a general metric $\tilde{d}$ on Ω. Additionally, the expression $a_n \gtrsim b_n$ will be used to denote the inequality up to a constant multiple where the value of the constant is independent of n. We also denote $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Furthermore, we denote A^c as the complement of set A for any set A while $B(x,r)$ denotes the ball, with respect to the l_2 norm, of radius $r > 0$ centered at $x \in \mathbb{R}^d$. Finally, we use $\text{Diam}(\Theta) = \sup\{\|\theta_1 - \theta_2\| : \theta_1, \theta_2 \in \Theta\}$ to denote the diameter of a given parameter space Θ relative to the l_2 norm, $\|\cdot\|$, for elements in $\mathbb{R}^d$.

2. Preliminaries

We recall the notion of Wasserstein distance for mixing measures, along with the notions of strong identifiability and uniform Lipschitz continuity conditions that prove useful in Section 3.

Mixture model. Throughout the paper, we assume that $X_1, \dots, X_n$ are i.i.d. samples from a true but unknown distribution P_{G_0} with given density function

$$p_{G_0} := \int f(x|\theta)\, dG_0(\theta) = \sum_{i=1}^{k_0} p_i^0 f\big(x|\theta_i^0\big),$$

where $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ is a true but unknown mixing distribution with exactly k_0 number of support points, for some unknown k_0. Also, $\{f(x|\theta), \theta \in \Theta \subset \mathbb{R}^d\}$ is a given family of probability densities (or equivalently kernels) with respect to a sigma-finite measure μ on $\mathcal{X}$ where $d \geq 1$. Furthermore, Θ is a chosen parameter space, where we empirically believe that the true parameters belong to. In a well-specified setting, all support points of G_0 reside in Θ, but this may not be the case in a misspecified setting.

Regarding the space of mixing measures, let $\mathcal{E}_k := \mathcal{E}_k(\Theta)$ and $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ respectively, denote the space of all mixing measures with exactly and at most k support points, all in Θ. Additionally, denote $\mathcal{G} := \mathcal{G}(\Theta) = \cup_{k \in \mathbb{N}_+} \mathcal{E}_k$ the set of all discrete measures with finite supports on Θ. Moreover, $\overline{\mathcal{G}}(\Theta)$ denotes the space of all discrete measures (including those with countably infinite supports) on Θ. Finally, $\mathcal{P}(\Theta)$ stands for the space of all probability measures on Θ.

Wasserstein distance. As in [22,37] it is useful to analyze the identifiability and convergence of parameter estimation in mixture models using the notion of Wasserstein distance, which can be defined as the optimal cost of moving masses transforming one probability measure to another [51]. Given two discrete measures $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^{k'} p_i' \delta_{\theta_i'}$, a coupling between $\boldsymbol{p}$ and $\boldsymbol{p'}$ is a joint distribution $\boldsymbol{q}$ on $[1\ldots,k] \times [1,\ldots,k']$, which is expressed as a matrix $\boldsymbol{q} = (q_{ij})_{1\leq i\leq k, 1\leq j\leq k'} \in [0,1]^{k\times k'}$ with marginal probabilities $\sum_{i=1}^{k} q_{ij} = p_j'$ and $\sum_{j=1}^{k'} q_{ij} = p_i$ for any $i = 1,2,\ldots,k$ and $j = 1,2,\ldots,k'$. We use $\mathcal{Q}(\boldsymbol{p},\boldsymbol{p'})$ to denote the space of all such couplings. For any $r \geq 1$, the r-th order Wasserstein distance between G and G' is given by

$$W_r\big(G,G'\big) = \inf_{\boldsymbol{q}\in\mathcal{Q}(\boldsymbol{p},\boldsymbol{p'})} \Big(\sum_{i,j} q_{ij} \|\theta_i - \theta_j'\|^r\Big)^{1/r},$$

where $\|\cdot\|$ denotes the l_2 norm for elements in $\mathbb{R}^d$. It is simple to see that if a sequence of probability measures $G_n \in \mathcal{O}_{k_0}$ converges to $G_0 \in \mathcal{E}_{k_0}$ under the W_r metric at a rate $\omega_n = o(1)$ for some $r \geq 1$ then there exists a subsequence of G_n such that the set of atoms of G_n converges to the k_0 atoms of G_0, up to a permutation of the atoms, at the same rate ω_n.

Strong identifiability and uniform Lipschitz continuity. The key assumptions that will be used to analyze the posterior contraction of mixing measures include uniform Lipschitz condition and strong identifiability condition. The uniform Lipschitz condition can be formulated as follows [22].

Definition 2.1. We say the family of densities $\{f(x|\theta), \theta \in \Theta\}$ is uniformly Lipschitz up to the order r, for some $r \geq 1$, if f as a function of θ is differentiable up to the order r and its partial derivatives with respect to θ satisfy the following inequality

$$\sum_{|\kappa|=r} |\bigg(\frac{\partial^{|\kappa|} f}{\partial \theta^\kappa}(x|\theta_1) - \frac{\partial^{|\kappa|} f}{\partial \theta^\kappa}(x|\theta_2)\bigg)\gamma^\kappa| \leq C\|\theta_1 - \theta_2\|^\delta \|\gamma\|$$

for any $\gamma \in \mathbb{R}^d$ and for some positive constants δ and C independent of x and $\theta_1, \theta_2 \in \Theta$. Here, $\gamma^\kappa = \prod_{i=1}^d \gamma_i^{\kappa_i}$ where $\kappa = (\kappa_1,\ldots,\kappa_d)$.

The first order uniform Lipschitz condition is satisfied by many popular classes of density functions, including Gaussian, Student's t, and skew-normal family. Now, strong identifiability condition of the r^{th} order is formulated as follows.

Definition 2.2. For any $r \geq 1$, we say that the family $\{f(x|\theta), \theta \in \Theta\}$ (or in short, f) is *identifiable in the order* r, for some $r \geq 1$, if $f(x|\theta)$ is differentiable up to the order r in θ and the following holds

A1. For any $k \geq 1$, given k different elements $\theta_1, \ldots, \theta_k \in \Theta$. If we have $\alpha_\eta^{(i)}$ such that for almost all x

$$\sum_{l=0}^{r} \sum_{|\eta|=l} \sum_{i=1}^{k} \alpha_\eta^{(i)} \frac{\partial^{|\eta|} f}{\partial \theta^\eta}(x|\theta_i) = 0$$

then $\alpha_\eta^{(i)} = 0$ for all $1 \leq i \leq k$ and $|\eta| \leq r$.

Many commonly used families of density functions satisfy the first order identifiability condition, including location-scale Gaussian distributions and location-scale Student's t-distributions. Technically speaking, strong identifiability conditions are useful in providing the guarantee that we have some sort of lower bounds of Hellinger distance between mixing densities in terms of Wasserstein metric between mixing measures. For example, if f is identifiable in the first order, we have the following inequality [22]

$$h(p_G, p_{G_0}) \gtrsim W_1(G, G_0) \tag{1}$$

for any $G \in \mathcal{E}_{k_0}$. It implies that for any estimation method that yields the convergence rate $n^{-1/2}$ for density p_{G_0} under the Hellinger distance, the induced rate of convergence for the mixing measure G_0 is $n^{-1/2}$ under W_1 distance.

3. Posterior contraction under well-specified regimes

In this section, we assume that the mixture model is well-specified, that is, the data are i.i.d. samples from the mixture density p_{G_0}, where mixing measure G_0 has k_0 support points in compact parameter space $\Theta \subset \mathbb{R}^d$. Within this section, we assume further that the true but unknown number of components k_0 is finite. A Bayesian modeler places a prior distribution Π on a suitable subspace of $\overline{\mathcal{G}}(\Theta)$. Then the posterior distribution over G is given by:

$$\Pi(G \in B|X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p_G(X_i)\, d\Pi(G)}{\int_{\overline{\mathcal{G}}(\Theta)} \prod_{i=1}^n p_G(X_i)\, d\Pi(G)} \tag{2}$$

We are interested in the posterior contraction behavior of G toward G_0, in addition to recovering the true number of mixture components k_0.

3.1. Prior results

The customary prior specification for a finite mixture is to use a Dirichlet distribution on the mixing weights and another standard prior distribution on the atoms of the mixing measure. Let H be a distribution with full support on Θ. Thus, for a mixture of k components, the full Bayesian mixture model specification takes the form:

$$\boldsymbol{p} = (p_1, \ldots, p_k) \sim \text{Dirichlet}_k(\gamma/k, \ldots, \gamma/k),$$

$$\theta_1, \ldots, \theta_k \overset{i.i.d.}{\sim} H,$$

$$X_1, \dots, X_n | G = \sum_{i=1}^{k} p_i \delta_{\theta_i} \overset{i.i.d.}{\sim} p_G. \tag{3}$$

Suppose for a moment that k_0 is known, we can set $k = k_0$ in the above model specification. Thus, we would be in an *exact-fitted* setting. Provided that f satisfies both first-order identifiability condition and the uniform Lipschitz continuity condition, H is approximately uniform on Θ, then according to [22,37] it can be established that as n tends to infinity,

$$\Pi\big(G \in \mathcal{E}_{k_0}(\Theta) : W_1(G, G_0) \gtrsim (\log n/n)^{1/2} | X_1, \dots, X_n\big) \xrightarrow{p_{G_0}} 0. \tag{4}$$

The $(\log n/n)^{1/2}$ rate of posterior contraction is optimal up to a logarithmic term.

When k_0 is unknown, there may be a number of ways for the modeler to proceed. Suppose that an upper bound of k_0 is given, say $k_0 < \overline{k}$. Then by setting $k = \overline{k}$ in the above model specification, we have a Bayesian *overfitted* mixture model. Provided that f satisfies the second-order identifability condition and the uniform Lipschitz continuity condition, H is again approximately uniform distribution on Θ, then it can be established that [22,37]:

$$\Pi\big(G \in \mathcal{O}_{\overline{k}}(\Theta) : W_2(G, G_0) \gtrsim (\log n/n)^{1/4} | X_1, \dots, X_n\big) \xrightarrow{p_{G_0}} 0. \tag{5}$$

This result does not provide any guarantee about whether the true number of mixture components k_0 can be recovered. The rate (upper bound) $(\log n/n)^{1/4}$ under W_2 metric implies that under the posterior distribution the redundant mixing weights of G contracts toward zero at the rate $(\log n/n)^{1/2}$, but the posterior contraction to each of the k_0 atoms of G_0 occurs at the rate $(\log n/n)^{1/4}$ only.

Interestingly, it can be shown by Rousseau and Mengersen [43] that with a more judicious choice of prior distribution on the mixing weights, one can achieve a near-optimal posterior contraction behavior. Specifically, they continued to employ the Dirichlet prior, but they required the Dirichlet's hyperparameters set to be sufficiently small: $\gamma/k \leq d/2$ in (3) where $k = \overline{k}$, d is the dimension of the parameter space Θ. Then, under some conditions on kernel f approximately comparable to the second-order identifiability and the uniform Lipschitz continuity condition defined in the previous section, they showed that for any $\epsilon > 0$, as n tends to infinity

$$\Pi\left(\exists I \subset \{1, \dots, k\}, |I| = k - k_0 \text{ s.t. } \sum_{i \in I} p_i < n^{-1/2+\epsilon} | X_1, \dots, X_n\right) \xrightarrow{p_{G_0}} 1. \tag{6}$$

For a more precise statement along with the complete list of sufficient conditions leading to claim (6), we refer the reader to the original theorem of [43]. Although their theorem is concerned with only the behavior of the redundant mixing weights p_i, where $i \in I$, which vanish at a near-optimal rate $n^{-1/2+\epsilon}$, it can be deduced from their proof that the posterior contraction for the true atoms of G_0 occurs at this near-optimal rate as well. [43] also showed that this performance may not hold if the Dirichlet's hyperparameters are set to be sufficiently large. Along this line, concerning the recovery of the number of mixture components k_0, [4] demonstrated the convergence of the posterior mode of the number of components to the true number of components k_0 at a rate $n^{-\rho}$, where ρ depends on $\overline{k} - k_0$, the number of redundant components forced upon by our model specification. Finally, we note that in addition to Dirichlet-type prior specifications, other types of prior specifications have also been taken up by other researchers [12,53].

3.2. Optimal posterior contraction via a parametric Bayesian mixture

We will show that optimal posterior contraction rates for mixture model parameters can be achieved by a natural Bayesian extension on the prior specification, even when the upper bound on the number of mixture component k is unknown. The modeling idea is simple and truly Bayesian in spirit: since k_0 is unknown, let K be a natural-valued random variable representing the number of mixture components. We endow K with a suitable prior distribution q_K on the positive integers. Conditioning on $K = k$, for each k, the model is specified as before:

$$\begin{aligned} K &\sim q_K, \qquad (7) \\ \boldsymbol{p} = (p_1, \dots, p_k) | K = k &\sim \text{Dirichlet}_k(\gamma/k, \dots, \gamma/k), \\ \theta_1, \dots, \theta_k | K = k &\overset{i.i.d.}{\sim} H, \\ X_1, \dots, X_n | G = \sum_{i=1}^{k} p_i \delta_{\theta_i} &\overset{i.i.d.}{\sim} p_G. \qquad (8) \end{aligned}$$

This prior specification is called *mixture of finite mixtures* (MFM) model [36,41,48]. In the sequel, we show that the application of the MFM prior leads to the optimal posterior contraction rates for the model parameters. Interestingly, such guarantees can be established under very mild conditions on the kernel density f: only the uniform Lipschitz continuity and the first-order identifiability conditions will be required. The first-order identifiability condition is the minimal condition for which the optimal posterior contraction rate can be established by the proof technique employed, since this condition is also necessary for exact-fitted mixture models to receive the $n^{-1/2}$ posterior contraction rate. We proceed to state such conditions.

(P.1) The parameter space Θ is compact, while kernel density f is first-order identifiable and admits the uniform Lipschitz property up to the first order.

(P.2) The base distribution H is absolutely continuous with respect to the Lebesgue measure μ on $\mathbb{R}^d$ and admits a density function $g(\cdot)$. Additionally, H is approximately uniform, that is, $\min_{\theta \in \Theta} g(\theta) > c_0 > 0$.

(P.3) There exists $\epsilon_0 > 0$ such that $\int (p_{G_0}(x))^2 / p_G(x) \, d\mu(x) \leq M(\epsilon_0)$ as long as $W_1(G, G_0) \leq \epsilon_0$ for any $G \in \mathcal{O}_{k_0}$ where $M(\epsilon_0)$ depends only on ϵ_0, G_0, and Θ.

(P.4) The prior q_K places positive mass on the set of natural numbers, that is, $q_K(k) > 0$ for all $k \in \mathbb{N}$.

Theorem 3.1. *Under assumptions* (P.1), (P.2), (P.3), *and* (P.4) *on MFM*, *we have that*

(a) $\Pi(K = k_0 | X_1, \dots, X_n) \to 1$ *a.s. under* P_{G_0}.

(b) *Moreover*,

$$\Pi\big(G \in \overline{\mathcal{G}}(\Theta) : W_1(G, G_0) \lesssim (\log n/n)^{1/2} | X_1, \dots, X_n\big) \to 1$$

in P_{G_0}*-probability.*

The proof of Theorem 3.1 is deferred to Appendix A.1. We now make several remarks regarding the conditions required in the theorem.

(i) It is worth stating up front that these conditions are almost minimal in order for the optimal posterior contraction to be guaranteed, and are substantially weaker than previous works (as discussed above). In particular, assumption (P.1) is crucial in establishing that the Hellinger distance $h(p_G, p_{G_0}) \geq C_0 W_1(G, G_0)$ where C_0 is some positive constant depending only on G_0 and Θ. Assumption (P.2) and (P.4) are standard conditions on the support of the prior so that posterior consistency can be guaranteed for any unknown G_0 with unknown number of support atoms residing on Θ. The role of (P.3) is to help control the growing rate of KL neighborhood, which is central in the analysis of posterior convergence rate of mixing measures. This assumption is held for various choices of kernel f, including location families and location-scale families. Therefore, the assumptions (P.1), (P.2), (P.3) and (P.4) are fairly general and satisfied by most common choice of kernel densities.

(ii) Condition (P.2) may be replaced by the following weaker condition:

(P.2') The base distribution H is absolutely continuous with respect to the Lebesgue measure μ on $\mathbb{R}^d$ and admits a density function $g(\cdot)$. Additionally, H must contain sufficient mass near the atoms of G_0, that is, $\min_{\theta:\|\theta-\theta_i^0\|\leq\epsilon} g(\theta) \geq c_0 > 0$ for some $\epsilon > 0$.

We prefer (P.2) which is required for unknown G_0 and is a reasonable assumption in practice.

(iii) The contraction rate with respect to the W_1 norm for strongly identifiable family of densities is $O_P((\log(n)/n)^{1/2})$. The contraction rates relative to the L_q norms for $q \geq 1$ can be obtained by Lemma D.4 and it is easy to show that the corresponding contraction rates are $O_P((\log(n)/n)^{1/2q})$ for $1 \leq q \leq 2$ and $O_P((\log(n)/n)^{1/q})$ for $q \geq 2$.

Theorem 3.1 provides a positive endorsement for employing the MFM prior when the number of mixture components is unknown, but is otherwise believed to be finite and an important quantity of inferential interest. The papers of [36,41] discuss additional favorable properties of this class of models. However, when the true number of mixture components is large, posterior inference with the MFM may still be inefficient in practice. This is because much of the computational effort needs to be expended for the model selection phase, so that the number of mixture components can be reliably ascertained. Only then does the fast asymptotic rate of parameter estimation come meaningfully into effect.

3.3. A posteriori processing for BNP mixtures

Instead of placing a prior distribution explicitly on the number of mixture components when this quantity is unknown, another predominant approach is to place a Bayesian nonparametric prior on the mixing measure G, resulting in infinite mixture models. Bayesian nonparametric models such as Dirichlet process mixtures and the variants have remarkably extended the reach of mixture modeling into a vast array of applications, especially those areas where the number of mixture components in the modeling is very large and difficult to fathom, or when it is a quantity of only tangential interest. For instance, in topic modeling applications of web-based text corpora, one may be interested in the most "popular" topics, the number of topics is less meaningful [3,38,50,54]. DP mixtures and variants can also serve as an asymptotically optimal device for estimating the population density, under standard conditions on the true density's smoothness, see, for example, [16,17,44,46].

Since a nonparametric Bayesian prior such as the Dirichlet process places zero probability on mixing measures with finite number of supporting atoms, the Dirichlet process mixture's posterior is inconsistent on the number of mixture components, provided the true number of mixture components is finite [35]. It is well known in practice that Dirichlet process mixtures tend to produce many small extraneous components around the "true" clusters, making them challenging to use to draw conclusion about the true number of mixture components when this becomes a quantity of interest [18,32]. In this section, we describe a simple posteriori processing algorithm that consistently estimates the number of components for any general Bayesian prior, even without the exact knowledge of its structure as long as the posterior for that prior contracts at some known rate to the true G_0.

Our starting point is the availability of a mixing measure sample G that is drawn from the posterior distribution $\Pi(G|X_1, \dots, X_n)$, where $X_1, \dots, X_n$ are i.i.d. samples of the mixing density p_{G_0}. Under certain conditions on the kernel density f, it can be established that for some Wasserstein metric W_r, as $n \to \infty$

$$\Pi\big(G \in \overline{\mathcal{G}}(\Theta) : W_r(G, G_0) \leq \delta\omega_n | X_1, \dots, X_n\big) \xrightarrow{p_{G_0}} 1 \tag{9}$$

for *all* constant $\delta > 0$, while $\omega_n = o(1)$ is a vanishing rate. Thus, ω_n can be taken to be (slightly) slower than actual rate of posterior contraction of the mixing measure. Concrete examples of the posterior contraction rates in infinite and (overfitted) finite mixtures are given in [13,22,37].

The posterior processing algorithm operates on an instance of mixing measure G, by suitably merging and truncating atoms that provide the support for G. The only inputs to the algorithm, which we call *Merge-Truncate-Merge* (MTM) algorithm is G, in addition to the upper bound of posterior contraction rate ω_n, and a tuning parameter $c > 0$. The tuning parameter c is useful in practice, as we shall explain, but in theory the algorithm "works" for any constant $c > 0$. Thus, the method is almost "automatic" as it does not require any additional knowledge about the kernel density f or the space of support Θ for the atoms. It is also simple and fast. We shall show that the outcome of the algorithm is a consistent estimate of both the number of mixing components and the mixing measure. The latter admits a posterior contraction rate's upper bound ω_n as well.

The detailed pseudocode of MTM algorithm is summarized in Algorithm 1. At a high level, it consists of two main stages. The first stage involves a probabilistic procedure for merging atoms that may be clustered near one another. The second stage involves a deterministic procedure for truncating extraneous atoms and merging them suitably with the remaining ones in a systematic way. The driving force of the algorithm lies in the asymptotic bound on the Wasserstein distance, that is, $W_r(G, G_0) \leq c\omega_n$ with high probability. When $c\omega_n$ is sufficiently small, there may be many atoms that concentrate around each of the supporting atoms of G_0. Although G_0 is not known, such clustering atoms may be merged into one, by our first stage of probabilistic merging scheme. The second stage (truncate-merge) is also necessary in order to obtain a consistent estimate of k_0, because there remain distant atoms which carry a relatively small amount of mass. They will need to be suitably truncated and merged with the other more heavily supported atoms. In other words, our method can be viewed as a formal procedure of the common practices employed by numerous practitioners.

We proceed to present the theoretical guarantee for the outcome of Algorithm 1.

Theorem 3.2. *Let G be a posterior sample from posterior distribution of any Bayesian procedure, namely, $\Pi(\cdot|X_1, \dots, X_n)$ according to which the upper bound (9) holds for all $\delta > 0$. Let $\tilde{G}$ and $\tilde{k}$ be the outcome of Algorithm 1 applied to G, for an arbitrary constant $c > 0$. Then the following hold as $n \to \infty$.*

(a) $\Pi(\tilde{k} = k_0 | X_1, \dots, X_n) \to 1$ *in P_{G_0}-probability.*
(b) *For all $\delta > 0$, $\Pi(G \in \overline{\mathcal{G}}(\Theta) : W_r(\tilde{G}, G_0) \leq \delta\omega_n | X_1, \dots, X_n) \longrightarrow 1$ in P_{G_0}-probability.*

We add several comments concerning this theorem.

(i) The proof of this theorem is deferred to Appendix A.2, where we clarify carefully the roles played by each step of the MTM algorithm.
(ii) Although it is beyond the scope of this paper to study the practical viability of the MTM algorithm, for interested readers, we present a brief illustration of the algorithm via simulations in Section 5.
(iii) In practice, one may not have a mixing measure G sampled from the posterior $\Pi(\cdot|X_1, \dots, X_n)$ but a sample from G itself, say F_n, the empirical distribution function. Then one can apply

Algorithm 1 Merge-Truncate-Merge Algorithm

```
Input: Posterior sample $G = \sum_i p_i \delta_{\theta_i}$ from (9), rate $\omega_n$, constant $c$.
Output: Discrete measure $\widetilde{G}$ and its number of supporting atoms $\tilde{k}$.
  {Stage 1: Merge procedure:}
  1: Reorder atoms $\{\theta_1, \theta_2, \dots\}$ by simple random sampling without replacement with corresponding weights $\{p_1, p_2, \dots\}$.
         let $\tau_1, \tau_2, \dots$ denote the new indices, and set $\mathcal{E} = \{\tau_j\}_j$ as the existing set of atoms.
  2: Sequentially for each index $\tau_j \in \mathcal{E}$, if there exists an index $\tau_i < \tau_j$ such that $\|\theta_{\tau_i} - \theta_{\tau_j}\| \le \omega_n$, then:
         update $p_{\tau_i} = p_{\tau_i} + p_{\tau_j}$, and remove $\tau_j$ from $\mathcal{E}$.
  3: Collect $G' = \sum_{j:\tau_j \in \mathcal{E}} p_{\tau_j} \delta_{\theta_{\tau_j}}$.
         write $G'$ as $\sum_{i=1}^{k} q_i \delta_{\phi_i}$ so that $q_1 \ge q_2 \ge \dots$.
  {Stage 2: Truncate-Merge procedure:}
  4: Set $\mathcal{A} = \{i : q_i > (c\omega_n)^r\}$, $\mathcal{N} = \{i : q_i \le (c\omega_n)^r\}$.
  5: For each index $i \in \mathcal{A}$, if there is $j \in \mathcal{A}$ such that $j < i$ and $q_i \|\phi_i - \phi_j\|^r \le (c\omega_n)^r$, then
         remove $i$ from $\mathcal{A}$ and add it to $\mathcal{N}$.
  6: For each $i \in \mathcal{N}$, find atom $\phi_j$ among $j \in \mathcal{A}$ that is nearest to $\phi_i$
         update $q_j = q_j + q_i$.
  7: Return $\widetilde{G} = \sum_{j \in \mathcal{A}} q_j \delta_{\phi_j}$ and $\tilde{k} = |\mathcal{A}|$.
```

> the MTM algorithm to F_n instead. Assume that F_n is sufficiently close to G, in the sense that $W_r(F_n, G) \lesssim W_r(G, G_0)$, it is straightforward to extend the above theorem to cover this scenario.

Practical implications. At this point, one may look forward to some guidance regarding the modeling choices of parametrics versus nonparametrics. Even in the tight arena of Bayesian mixture modeling, the jury may still be out. The results in this section seems to provide a stronger theoretical support for the former, when it comes to the efficiency of parameter estimation and the corresponding model interpretation. However, as we will see in the next section, when the mixture model is misspecified, the fast posterior contraction rate offered by the use of the MFM prior is no longer valid. On the other hand, Bayesian nonparametric models are more versatile in adapting to complex forms of population densities. In many modern applications it is not meaningful to estimate the number of mixing components, only the most "significant" ones in a sense suitably defined. Perhaps a more meaningful question concerning a Bayesian nonparametric mixture model is whether it is capable of learning selected mixture components in an efficient way.

4. Posterior contraction under model misspecification

In this section, we study the posterior contraction behavior of the mixing measure under the realistic scenarios of model misspecification. There are several ways a mixture model can be misspecified, due to the misspecification of the kernel density function f, or the support of the mixing measure G, or both. From here on, we shall assume that the data population follows a mixture distribution composed of unknown kernel density f_0 and unknown mixing measure G_0 — thus, in this section the true density shall be denoted by p_{G_0,f_0} to highlight the possibility of misspecification.

To avoid heavy subscripting, we continue to use p_G instead of $p_{G,f}$ to represent the density function of the mixture model that we operate on. The kernel density f is selected by the modeler. Additionally, G is endowed with a suitable prior Π on the space of mixing measures with support belonging to compact parameter space Θ. By Bayes rule (Eq. (2)) one obtains the posterior distribution $\Pi(G|X_1,\dots,X_n)$, where the n-i.i.d. sample $X_1,\dots,X_n$ are generated by p_{G_0,f_0}. It is possible that $f \neq f_0$. It is also possible that the support of G_0 does not reside within Θ. In practice, the statistical modeler would hope that the kernel choice of f is not too different from the true but unknown f_0. Otherwise, it would be unclear how one can interpret the parameters that represent the mixing measure G. Our goal is to investigate the posterior contraction of $\Pi(G|X_1,\dots,X_n)$ in such situations, as sample size n tends to infinity. The theory is applicable for a broad class of prior specification on the mixing measures on Θ, including the MFM prior and a nonparametric Bayesian prior such as the Dirichlet process.

A fundamental quantity that arises in the theory of Bayesian misspecification for density estimation is the minimizer of the Kullback–Leibler (KL) divergence from the true population density to a density function residing in the support of the induced prior on the space of densities p_G, which we shall assume to exist (cf. [27]). Moreover, assume that the KL minimizer can be expressed as a mixture density p_{G_*}, where G_* is a probability measure on Θ. We may write

$$G_* \in \operatorname*{arg\,min}_{G\in\mathcal{P}(\Theta)} K(p_{G_0,f_0}, p_G). \tag{10}$$

We will see in the sequel that the existence of the KL minimizer p_{G^*} entails its uniqueness. In general, however, G_* may be non-unique. Thus, define

$$\mathcal{M}^* := \bigl\{G_* \in \mathcal{P}(\Theta) : G_* \in \operatorname*{arg\,min}_{G\in\mathcal{P}(\Theta)} K(p_{G_0,f_0}, p_G)\bigr\}.$$

It is challenging to characterize the set $\mathcal{M}^*$ in general. However, a very useful technical property can be shown as follows:

Lemma 4.1. *For any $G \in \mathcal{P}(\Theta)$ and $G_* \in \mathcal{M}^*$, it holds that $\int \frac{p_G(x)}{p_{G_*}(x)} p_{G_0,f_0}(x)\,dx \leq 1$.*

By exploiting the fact that the class of mixture densities is a convex set, the proof of this lemma is similar to that of Lemma 2.3 of [27], so it is omitted. This leads quickly to the following fact.

Lemma 4.2. *For any two elements $G_{1,*}, G_{2,*} \in \mathcal{M}^*$, $p_{G_{1,*}}(x) = p_{G_{2,*}}(x)$ for almost all $x \in \mathcal{X}$.*

In other words, the mixture density p_{G_*} is uniquely identifiable. Under a standard identifiability condition of the kernel f, which is satisfied by the examples considered in this section, it follows that G_* is unique. Due to the model misspecification, in general $G_* \neq G_0$. The best we can hope for is that the posterior distribution of the mixing measure G contracts toward G_* as n tends to infinity.

The goal of the remaining of this section is to study the posterior contraction behavior of the (misspecified) mixing measure G towards the unique G_*.

Following the theoretical framework of [27] and [37], the posterior contraction behavior of the mixing measure G can be obtained by studying the relationship of a weighted version of Hellinger distance and corresponding Wasserstein distances between G and the limiting point G_*. In particular, for a fixed pair of mixture densities p_{G_0,f_0} and p_{G_*}, the weighted Hellinger $\overline{h}$ between two mixture densities is defined as follows [27].

Definition 4.1. For $G_1, G_2 \in \mathcal{P}(\Theta)$,

$$\overline{h}^2(p_{G_1}, p_{G_2}) := \frac{1}{2}\int \big(\sqrt{p_{G_1}(x)} - \sqrt{p_{G_2}(x)}\big)^2 \frac{p_{G_0,f_0}(x)}{p_{G_*}(x)}\, dx.$$

It is clear that when $G_* = G_0$ and $f = f_0$, the weighted Hellinger distance reduces to the standard Hellinger distance. In general they are different due to misspecification. According to Lemma 4.1, we have $\overline{h}(p_{G_1}, p_{G_2}) \leq 1$ for all $G_1, G_2 \in \mathcal{P}(\Theta)$.

Choices of prior on mixing measures. As in the previous section, we work with two representative priors on the mixing measure: the MFM prior and the Dirichlet process prior. Both prior choices may contribute to the model misspecification, if the true mixing measure G_0 lies outside of the support of the prior distribution.

Recall the MFM prior specification given in Eq. (7). We also need a stronger condition on q_K:

(P.4') The prior distribution q_K on the number of components satisfies $q_k \gtrsim k^{-\alpha_0}$ for some $\alpha_0 > 1$.

The $\alpha_0 > 1$ condition is placed in order to ensure that q_K is a proper distribution on natural numbers. Note that the assumption with prior on the number of components q_K is mild and satisfied by many distributions, such as Poisson distribution. In order to obtain posterior contraction rates, one needs to make sure the prior places sufficient mass on the (unknown) limiting point of interest. For the MFM prior, such a condition is guaranteed by the following lemma.

Lemma 4.3. *Let Π denote the prior for generating G based on MFM (7), where H admits condition (P.2) and q_K admits (P.4'). Fix $r \geq 1$. Then the following holds, for any $G_* \in \mathcal{P}(\Theta)$*

$$\Pi\big(G : W_r^r(G, G_*) \leq (2^r + 1)\epsilon^r\big) \gtrsim \frac{\gamma\Gamma(\gamma)D!q_D}{D}\left(c_0\left(\frac{\epsilon}{\mathrm{Diam}(\Theta)}\right)^d\right)^D \left(\frac{1}{D}\left(\frac{\epsilon}{\mathrm{Diam}(\Theta)}\right)^r\right)^{\gamma(D-1)/D} \tag{11}$$

for all ϵ sufficiently small so that $D(\epsilon, \Theta, \|.\|) > \gamma$. Here, $D = D(\epsilon, \Theta, \|.\|)$ and q_D stand for the maximal ϵ-packing number for Θ under $\|.\|$ norm and the prior weight $\Pi(K = D)$, respectively.

The proof of Lemma 4.3 is provided in Appendix A.3. Alternatively, for a Dirichlet process prior, G is distributed a priori according to a Dirichlet measure with concentration parameter $\gamma > 0$ and base measure H satisfying condition (P.2). An analogous concentration bound for such a prior is given in Lemma 5 of [37].

It is somewhat interesting to note that the difference in the choices of prior under misspecification does not affect the posterior contraction bounds that we can establish. In particular, as we have seen for the definition, G_* does not depend on a specific choice of prior distribution (only its support). Due to misspecification, G_* may have infinite support, even if the true G_0 has a finite number of support points. When G_* has infinite support, the posterior contraction toward G_* becomes considerably slower compared to the well-specified setting. In addition to the structure of G_*, we will see in the sequel that the modeler's specific choice of kernel density f proves to be especially impactful on the rate of posterior contraction.

4.1. Gaussian location mixtures

Consider a class of kernel densities that belong to the supersmooth location family of density functions. A particular example that we focus on in this section is a class of Gaussian distributions with some fixed covariance matrix Σ. More precisely, f has the following form:

$$\left\{f(\cdot|\theta), \theta \in \Theta \subset \mathbb{R}^d : f(x|\theta) := \frac{\exp(-(x-\theta)^\top \Sigma^{-1}(x-\theta)/2)}{|2\pi\Sigma|^{-1/2}}\right\}, \tag{12}$$

where $|\cdot|$ stands for matrix determinant. Note that, Gaussian kernel is perhaps the most popular choice in mixture modeling.

With the Gaussian location kernel, it is possible to obtain a lower bound on the Hellinger distance between the mixture densities in terms of the Wasserstein distance between corresponding mixing measures [37]. More useful in the misspecified setting is a key lower bound for the weighted Hellinger distance in terms of the Wasserstein metric, which is given below in Prop. 4.1. In order to establish this bound, we shall require a technical condition relating f to the true f_0 and G_0. This condition is stated by assumption (P.5) or a weaker version (P.5').

(P.5) The support of G_0, namely, $\text{supp}(G_0)$ is a bounded subset of $\mathbb{R}^d$. Moreover, there are some constants $C_0, C_1, \alpha > 0$ such that for any $R > 0$,

$$\sup_{x\in\mathbb{R}^d, \theta\in\Theta, \theta_0\in\text{supp}(G_0)} \frac{f(x|\theta)}{f_0(x|\theta_0)} \mathbb{1}_{\|x\|_2 \le R} \le C_1 \exp\bigl(C_0 R^\alpha\bigr).$$

The condition in (P.5) that the support of G_0 has the same dimension d is purely for the sake of interpretability, if the quantity of inferential interest is the mixing measure G. This is also related to the condition in (P.5) on the density ratio $f(x/\theta)/f_0(x|\theta_0)$. In fact, both conditions on the support of G_0 and on f_0 are not strictly necessary from a technical standpoint; only a "black-box" condition directly placed on the true density p_{G_0,f_0} will be sufficient. Accordingly, (P.5) may be replaced by the following weaker condition.

(P.5') Assume that there are some constants $C_0, C_1, \alpha > 0$ such that for any $R > 0$,

$$\sup_{x\in\mathbb{R}^d, \theta\in\Theta} \frac{f(x|\theta)}{p_{G_0,f_0}(x)} \mathbb{1}_{\|x\|_2 \le R} \le C_1 \exp\bigl(C_0 R^\alpha\bigr).$$

It is simple to verify that (P.5) implies (P.5').

Examples. In the following examples, the statistician decides to fit the data with a Gaussian location mixture model $p_{G,f}$, where the kernel $f(x|\theta)$ corresponds to a Gaussian kernel with mean parameter $\theta \in \Theta \subset \mathbb{R}^d$, a fixed non-degenerate covariance Σ as given by Eq. (12). In addition, the mixing measure $G \in \mathcal{P}(\Theta)$.

1. If f_0 is a Gaussian kernel with mean parameter in a bounded set $\Theta_0 \subset \mathbb{R}^d$ and fixed non-degenerate covariance Σ_0, and $G_0 \in \mathcal{P}(\mathbb{R}^d)$, then the true density p_{G_0,f_0} corresponds to a Gaussian location mixture. The model may be misspecified due to either $\Sigma_0 \neq \Sigma$, or $\text{supp}(G_0) \not\subset \Theta$, or both. In this case, (P.5) is satisfied for all $\alpha \geq 2$. The constant C_0 depends on Σ, Σ_0 as well as $\text{supp}(G_0)$ and Θ. On the other hand, C_1 depends on the eigenvalues of Σ, Σ_0 as well as the value of α.

2. If f_0 is a Gaussian kernel with both mean and covariance parameter varying in some compact subsets of $\mathbb{R}^d$ and positive definite $d \times d$ matrices, respectively, so that the true density p_{G_0, f_0} corresponds to a Gaussian location-scale mixture. In this case, (P.5) is not applicable, but (P.5') holds with all $\alpha \geq 2$. The constant C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters. On the other hand, C_1 depends on the value of α chosen, Σ as well as the compact subset corresponding to the covariance parameter.
3. If f_0 is a Student's t kernel, with both mean and covariance parameter varying in some compact subsets of $\mathbb{R}^d$ and positive definite $d \times d$ matrices, respectively, then p_{G_0, f_0} corresponds to a location-scale mixture of t distributions. In this scenario too, (P.5) may not be applicable, but (P.5') is, for all $\alpha > 0$. Both C_0 and C_1 depend on the choice of α. In addition, C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters, while C_1 depends on Σ as well as the compact subset corresponding to the covariance parameter.
4. If f_0 is a Laplace kernel with mean parameter in a bounded set $\Theta_0 \subset \mathbb{R}^d$, fixed covariance Σ_0, fixed scale parameter λ_0, and $G_0 \in \mathcal{P}(\mathbb{R}^d)$, (P.5) is satisfied for all $\alpha > 0$. Both C_0 and C_1 depend on the choice of α. In addition C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters, while C_1 depends on Σ as well as the compact subset corresponding to the covariance parameter.

Proposition 4.1. *Let f be a Gaussian kernel given by* (12), *Θ a bounded subset of $\mathbb{R}^d$. Moreover, assume that f, Θ and the true data generating distribution P_{G_0, f_0} satisfy either condition* (*P*.5) *or* (*P*.5') *for some $\alpha \leq 2$. Then, there exists $\epsilon_0 > 0$ depending on Θ and Σ, such that for any $G, G' \in \mathcal{P}(\Theta)$, whenever $\overline{h}(p_G, p_{G'}) \leq \epsilon_0$, the following inequality holds*

$$\overline{h}(p_G, p_{G'}) \geq C \exp\bigl(-\bigl(1 + 8\lambda_{\max}\bigl(\lambda_{\min}^{-1} + C_0\bigr)\bigr)/W_2^2(G, G')\bigr).$$

Here, $\lambda_{\max}$ and $\lambda_{\min}$ are respectively the maximum and minimum eigenvalue of Σ. C is a constant depending on the parameter space Θ, the dimension d, the covariance matrix Σ, G_0 and C_1 in condition (*P*.5) *or* (*P*.5').

The proof of Proposition 4.1 is provided in Appendix A.4. We are ready to prove the first main result of this section.

Theorem 4.1. *Assume that f satisfies condition specified in Prop.* 4.1, *and Π is an MFM prior on $\mathcal{P}(\Theta)$ specified in Lemma* 4.3. *Then, as n tends to infinity, the following holds*

$$\Pi\biggl(G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_*) \lesssim \left(\frac{\log\log n}{\log n}\right)^{1/2} \bigg| X_1, \ldots, X_n\biggr) \to 1$$

in p_{G_0, f_0}-probability.

The proof of Theorem 4.1 is given in Appendix D.1. The same posterior contraction behaviors hold if we replace MFM prior by the Dirichlet process prior with no change in the proof, except that Lemma 5 of [37] is used in place of Lemma 4.3.

The above theorem provides a result on parameter estimation in the misspecified context. We also provide a result on density estimation as follows, even though it is not the primary focus of the paper.

In order to facilitate the presentation, we denote

$$\|p_{G_1} - p_{G_2}\|_{\tilde{L}_q} := \left(\int \frac{p_{G_0, f_0}(x)}{p_{G_*}(x)} \bigl|p_{G_1}(x) - p_{G_2}(x)\bigr|^q \, dx\right)^{1/q}.$$

It is clear that when $G_* = G_0$ and $f = f_0$, the $\tilde{L}_q$ norm becomes the standard L_q norm. In general, we define $\tilde{L}_q$ norm between mixing densities to account for the model misspecification.

Proposition 4.2. *Assume that Π is either the Dirichlet process prior or the MFM prior on $\mathcal{P}(\Theta)$ specified in Lemma 4.3. If f is a multivariate Gaussian kernel with covariance Σ, then we have*

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : \|p_G - p_{G_*}\|_{\tilde{L}_q} \lesssim \left(\frac{(\log(n))^2}{n}\right)^{1/q(d+2)} \middle| X_1, \dots, X_n\right) \to 1$$

in p_{G_0,f_0}-probability for $1 \le q \le 2$.

Furthermore, for $q \ge 2$ we obtain

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : \|p_G - p_{G_*}\|_{\tilde{L}_q} \lesssim \left(\frac{(\log(n))^2}{n}\right)^{2/q(d+2)} \middle| X_1, \dots, X_n\right) \to 1$$

in p_{G_0,f_0}-probability.

The proof of Proposition 4.2 is provided in Section D.2 in the Appendix.

4.2. Laplace location mixtures

Next, we consider a class of multivariate Laplace kernel, a representative in the family of ordinary smooth density functions. It was shown by [37] that under a Dirichlet process location mixture with a Laplace kernel, assuming the model is well-specified, the posterior contraction rate of mixing measures to G_0 is of order $n^{-\gamma}$ for some constant $\gamma > 0$. Under the current misspecification setting, we will be able to derive contraction rates toward G_* in the order of $n^{-\gamma'}$ for some constant γ' dependent on γ. The density of location Laplace distributions is given by:

$$f(x|\theta) = \frac{2}{\lambda(2\pi)^{d/2}} \frac{K_{(d/2)-1}(\sqrt{2/\lambda}\sqrt{(x-\theta)^\top \Sigma^{-1}(x-\theta)})}{(\sqrt{\lambda/2}\sqrt{(x-\theta)^\top \Sigma^{-1}(x-\theta)})^{(d/2)-1}}, \tag{13}$$

where Σ and $\lambda > 0$ are respectively fixed covariance matrix and scale parameter such that $|\Sigma| = 1$. Here, K_v is a Bessel function of the second kind of order v. As discussed in [7], $K_m(x) \sim \sqrt{\frac{\pi}{2x}} \exp(-x)$ as $|x| \to \infty$. Therefore, there exists $\tilde{R}$ such that as long as $\|x - \theta\| > \tilde{R}$, we have

$$f(x|\theta) \asymp \frac{\exp(-\sqrt{\frac{2}{\lambda}}\|x-\theta\|_{\Sigma^{-1}})}{(\|x-\theta\|_{\Sigma^{-1}})^{(d-1)/2}},$$

where we use the shorthand notation $\|y\|_{\Sigma^{-1}} = \sqrt{y^\top \Sigma^{-1} y}$. To ease the ensuing presentation, we denote

$$\tau(\alpha) := \frac{\sqrt{2/(\lambda\lambda_{\max})}}{(\sqrt{2/(\lambda\lambda_{\min})} + \sqrt{2/(\lambda\lambda_{\max})} + C_0)^{1/\alpha}}.$$

The following proposition provides a key lower bound of weighted Hellinger distance in terms of the Wasserstein metric.

Proposition 4.3. *Let f be a Laplace kernel given by* (13) *for fixed Σ and λ such that $|\Sigma| = 1$. Moreover, f, Θ and G_0 satisfy either condition (P.5) or (P.5') for some $\alpha \geq 1$. Then, there exists $\epsilon_0 > 0$ depending on Θ, λ and Σ, such that for any $G, G' \in \mathcal{P}(\Theta)$, whenever $\overline{h}(p_G, p_{G'}) \leq \epsilon_0$, the following inequality holds*

$$\left(\log \frac{1}{\overline{h}(p_G, p_{G'})}\right)^{d/(2\alpha)} \exp\left(-\tau(\alpha)\left(\log \frac{1}{\overline{h}(p_G, p_{G'})}\right)^{1/\alpha}\right) \geq C W_2^{2/m}(G, G').$$

for any positive constant $m < 4/(4+5d)$. Here, $\lambda_{\max}$ and $\lambda_{\min}$ are respectively, the maximum and minimum eigenvalue of Σ. The constant C depends on the parameter space Θ, the dimension d, the covariance matrix Σ, the scale parameter λ, G_0 and C_1 in (P.5) or (P.5').

The proof of Proposition 4.3 is provided in Appendix A.5. Given the above result, the posterior contraction rate for mixing measures G in the location family of Laplace mixture distributions can be obtained from the following result.

Theorem 4.2. *Assume that f is given by equation* (13) *for fixed Σ and λ such that $|\Sigma| = 1$. Additionally, assume that f satisfies condition specified in Prop. 4.3, and Π an MFM prior on $\mathcal{P}(\Theta)$ specified in Lemma 4.3. Then, as n tends to infinity, the following holds*

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_*) \lesssim \exp\left(-\frac{m\tau(\alpha)}{2}\left(\frac{\log n - \log\log n}{2(d+2)}\right)^{1/\alpha}\right)\middle| X_1, \ldots, X_n\right) \to 1$$

in p_{G_0, f_0}-probability for any positive constant $m < 4/(4+5d)$.

The proof of Theorem 4.2 is straightforward using the result in Proposition 4.3 and analogous to the proof argument of Theorem 4.1; therefore, it is omitted. Note that, identical to the Gaussian kernel case, a similar contraction behavior also holds for the Laplace kernel with the Dirichlet process prior. The proof can be obtained similar to the MFM prior by invoking Lemma 5 of [37] instead of Lemma 4.3.

We see from Theorems 4.1 and 4.2 that for parameter estimation, the Laplace kernel provide a faster contraction as compared to Gaussian kernels. However, the following result suggests that it may be the opposite when it comes to density estimation for misspecified scenarios.

Proposition 4.4. *Assume that Π is either the Dirichlet process prior or the MFM prior on $\mathcal{P}(\Theta)$ specified in Lemma 4.3. If f is a multivariate Laplace distribution given by equation* (13) *for fixed Σ and λ such that $|\Sigma| = 1$, then we arrive at*

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : \|p_G - p_{G_*}\|_{\tilde{L}_q} \lesssim \left(\frac{(\log(n))^2}{n}\right)^{1/q(2d+1)}\middle| X_1, \ldots, X_n\right) \to 1$$

in p_{G_0, f_0}-probability for $1 \leq q \leq 2$.

When $q \geq 2$, we find that

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : \|p_G - p_{G_*}\|_{\tilde{L}_q} \lesssim \left(\frac{(\log(n))^2}{n}\right)^{2/q(2d+1)}\middle| X_1, \ldots, X_n\right) \to 1$$

in p_{G_0, f_0}-probability.

The proof of Proposition 4.4 is analogous to that of Proposition 4.2 and is therefore omitted. The rates obtained for Propositions 4.2 and 4.4 are probably not sharp and may be improved upon, perhaps under additional assumptions.

Note that we only need condition (P.5') for the proofs of Theorem 4.2 and Proposition 4.3 to hold. Condition (P.5) is a stricter condition which ensures (P.5'). Relevant examples of condition (P.5) or (P.5') are provided as follows.

Examples. In the examples that follow, the statistician decides to fit the data with a Laplace location mixture model $p_{G,f}$, where the kernel $f(x|\theta)$ corresponds to a Laplace kernel with mean parameter $\theta \in \Theta \subset \mathbb{R}^d$, a fixed non-degenerate covariance Σ with $|\Sigma| = 1$ and a scale parameter $\lambda > 0$, as given by Eq. (13). In addition, the mixing measure $G \in \mathcal{P}(\Theta)$.

1. If f_0 is a Gaussian kernel with mean parameter in a bounded set $\Theta_0 \subset \mathbb{R}^d$ and fixed non-degenerate covariance Σ_0, and $G_0 \in \mathcal{P}(\mathbb{R}^d)$, then the true density p_{G_0,f_0} corresponds to a Gaussian location mixture. In this case, (P.5) is satisfied for all $\alpha \geq 2$. The constant C_0 depends on λ, Σ, Σ_0 as well as $\text{supp}(G_0)$ and Θ. On the other hand, C_1 depends on the eigenvalues of λ, Σ, Σ_0 as well as the value of α.
2. If f_0 is a Gaussian kernel with both mean and covariance parameter varying in some compact subsets of $\mathbb{R}^d$ and positive definite $d \times d$ matrices, respectively, so that the true density p_{G_0,f_0} corresponds to a Gaussian location-scale mixture. In this case, (P.5) is not applicable, but (P.5') holds with all $\alpha \geq 2$. The constant C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters. On the other hand, C_1 depends on the value of α chosen, Σ as well as the compact subset corresponding to the covariance parameter.
3. If f_0 is a Student's t kernel, with both mean and covariance parameter varying in some compact subsets of $\mathbb{R}^d$ and positive definite $d \times d$ matrices, respectively, then p_{G_0,f_0} corresponds to a location-scale mixture of t distributions. In this scenario too, (P.5) may not be applicable, but (P.5') is, for all $\alpha > 0$. Both C_0 and C_1 depend on the choice of α. In addition C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters, while C_1 depends on Σ as well as the compact subset corresponding to the covariance parameter.
4. If f_0 is a Laplace kernel with mean parameter in a bounded set $\Theta_0 \subset \mathbb{R}^d$, fixed covariance Σ_0, fixed scale parameter λ_0, and $G_0 \in \mathcal{P}(\mathbb{R}^d)$, (P.5) is satisfied for all $\alpha \geq 1$. Both C_0 and C_1 depend on the choice of α. In addition C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters, while C_1 depends on Σ as well as the compact subset corresponding to the covariance parameter.
5. If f_0 is a Laplace kernel with mean, scale and covariance parameters varying in some compact subsets of $\mathbb{R}_+$, $\mathbb{R}^d$ and positive definite $d \times d$ matrices with determinant 1, respectively, so that the true density p_{G_0,f_0} corresponds to a Laplace location-scale mixture. In this case, (P.5') holds with all $\alpha \geq 1$. The constant C_0 depends on Σ, λ, Θ as well as the compact subsets corresponding to the location, scale and covariance parameters. On the other hand, C_1 depends on the value of α chosen, Σ as well as the compact subsets corresponding to the scale and covariance parameters.

Remarks. (i) It is worth noting that compared to the well-specified setting, the posterior contraction upper bound obtained for Gaussian location mixtures remains the same slow logarithmic rate $(\log\log n/\log n)^{1/2}$. For Laplace mixtures, when the truth f_0 satisfies condition (P.5) with $\alpha \leq 1$, the posterior contraction upper bound obtained under misspecification remains a polynomial rate of the form $n^{-\gamma'}$ modulo a logarithmic term. Due to misspecification there is a loss of a constant factor in the exponent γ', which is dependent on the shape of the kernel density as it is captured by the term $\tau(\alpha)$.

(ii) Although Gaussian mixtures have proved to be an asymptotically optimal density estimation device under suitable and mild conditions (cf. [16]), the results obtained in this section raise some cautions for Gaussian kernels as a choice for mixture modeling under model misspecification, even if the true G_0 has finite number of support points, when the primary interest is in the quality of model parameter estimates. Mixtures of heavy-tailed and ordinary smooth kernel densities such as the Laplace prove to be more amenable to efficient parameter estimation. Thus, the modeler may be tempted to select for f, say, a Laplace kernel over a supersmooth kernel such as Gaussian kernel, provided that either condition (P.5) or (P.5') is valid.

(iii) It is interesting to consider the scenario where the true kernel f_0 happens to be a Gaussian kernel: if we use either a well-specified or a misspecified Gaussian kernel to fit the data, the posterior contraction bound is the extremely slow $(\log\log n/\log n)^{1/2}$ accordingly to Theorem 4.1. This rate may be too slow to be practical interpretation of parameters. If the statistician is too impatient to get to the truth G_0, because sample size n is not sufficiently large, he may well decide to select a Laplace kernel f instead. Despite the intentional misspecification, he might be comforted by the fact that the posterior distribution of G contracts at an exponentially faster rate to a G_* given by Theorem 4.2 for $\alpha = 2$. It is of interest, in theory at least, in this scenario to study the relation between G_* and true G_0, given certain assumptions on the true density p_{G_0,f_0}.

Practical implications. All models are misspecified in practice. The question of model choice in general, and kernel selection in particular is a challenging one, especially when one seeks the mixing measure G as a device for representing the heterogeneity of the data population. When the kernel family is misspecified, in general positions the limiting mixing measure G_* almost always has infinite support. This means in practice when we employ (Bayesian) nonparametric models, the more data we have the more heterogeneous patterns will show up via posterior estimates. As such, Theorems 4.1 and 4.2 inform us how the choice of the (likely misspecified) kernel affects the quality of the estimates for G. In the language of Bayesian inference, the theorems quantify in an asymptotic sense the role of data sample in transforming the prior distribution to the posterior distribution on the quantity of interest, whereas the matter of consistency toward the truth G_0 is left unknown (and in fact, unknowable in practice). At the same time, these theorems are not viewed by the authors as an endorsement of one kernel choice over another. For the purpose of interpreting latent subpopulations, it does not make sense to use G as a device for heterogeneity of the data population unless the kernel choice f is believed to be meaningful for subpopulations of interest, that is, f is sufficiently close to the true f_0. This is how a practitioner typically assumes. Once such a kernel choice f has been made, we have shown that some (misspecified) kernels result in more efficient estimates, and hence more amenable to interpretation, than others.

4.3. When G_* has finite support

The source of the deterioration in the statistical efficiency of parameter estimation under model misspecification is ultimately due to the increased complexity of the limiting point G_*. Even if the true G_0 has a finite number of support points, this is not the case for G_* in general. Unfortunately, it is very difficult to gain concrete information about G_* both in practice and in theory, due to the lack of knowledge about the true p_{G_0,f_0}. When some precious information about G_* is available, specifically, suppose that we happen to know G_* has a bounded number of support points k_* such that $k_* < \overline{k}$ for some known $\overline{k}$. Then it is possible to devise a new prior specification on the mixing measure G so that one can gain a considerably improved posterior contraction rate toward G_*. We will show that it is possible to obtain the contraction rate of the order $(\log n/n)^{1/4}$ under W_2 metric — this is the same rate of posterior contraction one would get with overfitted mixtures in the well-specified regime.

In order to analyze the convergence rate of mixing measure under that setting of k_*, we introduce a relevant notion of integral Lipschitz property, which is a generalized form of the uniform Lipschitz property for the misspecification scenarios.

Definition 4.2. For any given $r \geq 1$, we say that the family of densities f admits the *integral Lipschitz* property up to the order r with respect to two mixing measures G_0 and G_*, if f as a function of θ is differentiable up to the order r and its partial derivatives with respect to θ satisfy the following inequality

$$\sum_{|\kappa|=r} \left| \left(\frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta_1) - \frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta_2) \right) \gamma^{\kappa} \right| \leq C(x) \|\theta_1 - \theta_2\|^{\delta} \|\gamma\|^{r}$$

for any $\gamma \in \mathbb{R}^d$ and for some positive constants δ independent of x and $\theta_1, \theta_2 \in \Theta$. Here, $C(x)$ is some function such that $\int C(x) \frac{p_{G_0, f_0}(x)}{p_{G_*}(x)} \, dx < \infty$.

It is clear that when f has integral Lipschitz property up to the order r, for some $r \geq 1$, with respect to G_0 and G_*, then it will admit uniform Lipschitz property up to the order r. We can verify that the first order intergral Lipschitz property is satisfied by many popular kernels, including location-scale Gaussian distribution and location-scale Cauchy distribution.

In the following, we shall work with the MFM prior (7). Moreover,

(M.0) q_K places positive masses on $K \in \{1, \ldots, \overline{k}\}$ and 0 mass elsewhere, where $\overline{k} \gg k_*$ is a fixed number.

Given that k_* is finite, we obtain a key lower bound of weighted Hellinger distance in terms of the Wasserstein metric under strong identiability of f:

Proposition 4.5. *Assume that f is second order identifiable and admits uniform integral Lipschitz property up to the second order. Then, for any $G \in \mathcal{O}_{\overline{k}}$, the following inequality holds*

$$\overline{h}(p_G, p_{G_*}) \gtrsim W_2^2(G, G_*).$$

The proof of Proposition 4.5 is in Appendix B.2. Before stating the final theorem of this section, we will need following assumptions:

(M.1) The assumptions of Proposition 4.5 hold, that is, f is second order identifiable and admits uniform integral Lipschitz property up to the second order.

(M.2) There exists $\epsilon_0 > 0$ such that $\int (p_{G_0, f_0}(x)) p_{G_*}(x)/p_G(x) \, d\mu(x) \leq M^*(\epsilon_0)$ whenever we have $W_1(G, G_*) \leq \epsilon_0$ for any $G \in \mathcal{O}_{k_*}$ where $M^*(\epsilon_0)$ depends only on ϵ_0, G_*, G_0, and Θ.

(M.3) The parameter γ in Dirichlet distribution in MFM satisfies $\gamma < \overline{k}$. Additionally, the base distribution H satisfies Assumption (P.2).

Theorem 4.3. *Assume $k_0 < \infty$, and assumptions* (*M*.0), (*M*.1), (*M*.2) *and* (*M*.3) *hold. Then we have that,*

$$\Pi\big(G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_*) \lesssim (\log n/n)^{1/4} | X_1, \ldots, X_n\big) \to 1$$

in p_{G_0, f_0}-probability.

The proof of Theorem 4.3 is deferred to Section D.6.

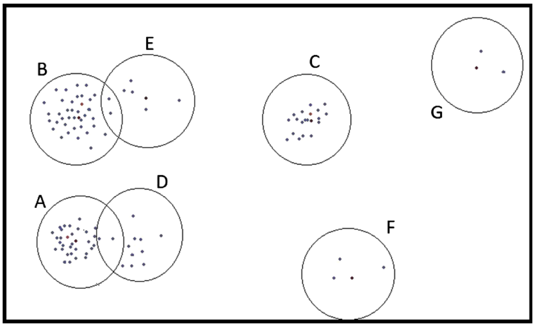

Figure 1. Initial distribution G.

Further remarks. The above theorem raises a promising prospect for combating model misspecification, by having the modeler fit the data to an *underfitted* mixture model p_G. Unfortunately, this theorem does not address this scenario, under which the limiting mixing measure would correspond to the KL minimizer

$$G_{**} = \underset{G \in \mathcal{O}_{\overline{k}}(\Theta)}{\arg\min}\, K(p_{G_0, f_0}, p_G).$$

for some $\overline{k} < \infty$, provided that this quantity exists (compare this with G_* given in (10)). Due to the lack of convexity of the class of mixture densities with bounded number of mixture components, the theory developed in this section (tracing back to the work of [27]) is not applicable. Thus, posterior contraction behaviors in an underfitted mixture models remain an interesting open question.

5. Simulation studies

In this section, we provide an illustration of the MTM algorithm's behavior via a simple simulation study. Figures 1, 2, 3 and 4 illustrate the different stages in the application of MTM algorithm 1. In each figure, green dots denote the atoms in the set of "remaining atoms" at each stage, with weights proportional to their sizes. Red dots denote the supporting atoms of the true mixing measure G_0. Black

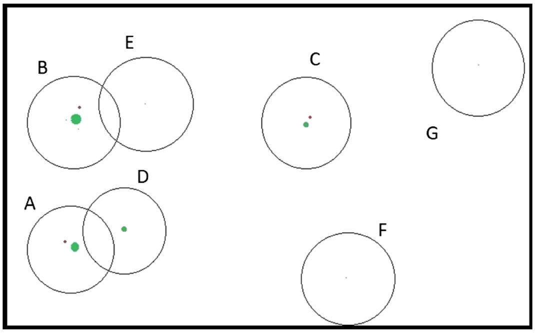

Figure 2. After first stage-"merge".

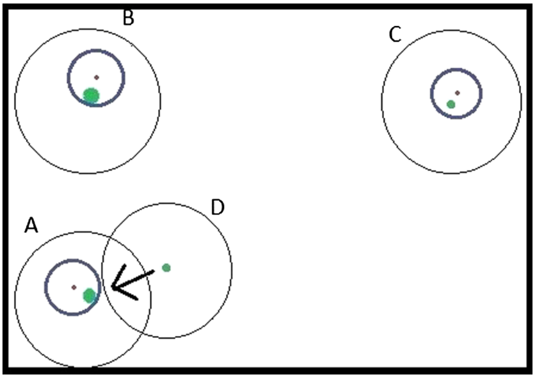

Figure 3. After second stage-“truncation”.

circles denote balls of radius ω_n around each of the “remaining atoms”. Blue circles denote balls of radius $\frac{\omega_n}{4k_0}$ around the atoms of G_0.

Starting with an input measure G represented in Figure 1, the first stage of the algorithm (merge procedure, from line 1 to line 4) merges nearby atoms to produce G', which is represented by Figure 2. There remains some atoms that carry very small mass, they are suitably truncated (via line 5 in the algorithm), and then merged accordingly (via line 6). Figure 3 and Figure 4 represent the outcome after these two steps of the algorithm. Observe how the atoms in each of the blue circles are merged to produced a reasonably accurate estimate of the corresponding atom of G_0. The number of such circles gives the correct number of the supporting atoms of G_0.

Next, we illustrate the performance of the MTM algorithm as it is applied to the samples from a Dirichlet process mixture, given the data generated by mixtures of three location Gaussian distributions:

$$p_{G_0}(\cdot) = \sum_{i=1}^{3} p_i^0 \mathcal{N}\big(\cdot|\mu_i^0, \Sigma^0\big),$$

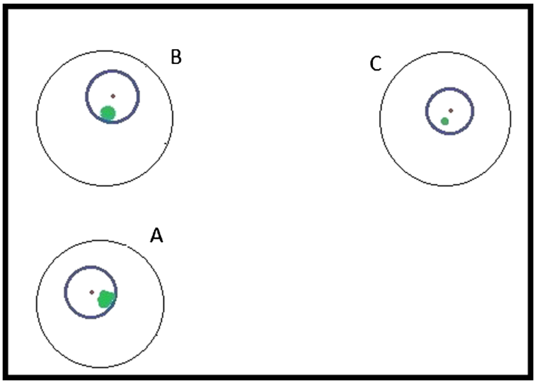

Figure 4. After second stage-“merge”.

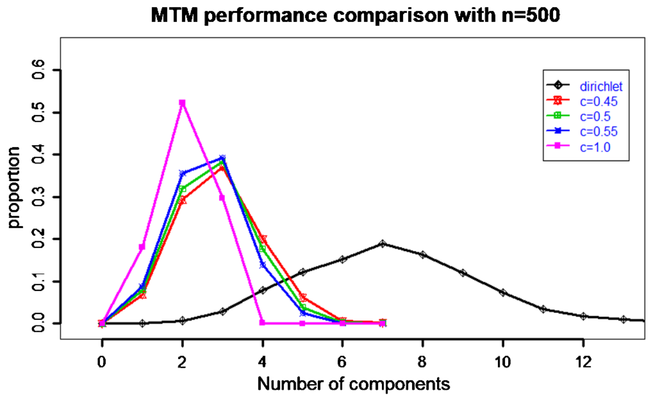

Figure 5. Case A.

where $\mathcal{N}(\cdot|\mu, \Sigma)$ is the Gaussian distribution with mean vector μ and covariance matrix Σ. For simulation purposes, we consider the following four different settings (n is the sample size):

1. Case A: $\mu_1^0 = (0.8, 0.8)$, $\mu_2^0 = (0.8, -0.8)$, $\mu_3^0 = (-0.8, 0.8)$, $\Sigma^0 = 0.05I_3$, $n = 500$.
2. Case B: $\mu_1^0 = (0.8, 0.8)$, $\mu_2^0 = (0.8, -0.8)$, $\mu_3^0 = (-0.8, 0.8)$, $\Sigma^0 = 0.05I_3$, $n = 1500$.
3. Case C: $\mu_1^0 = (1.8, 1.8)$, $\mu_2^0 = (1.8, -1.8)$, $\mu_3^0 = (-1.8, 1.8)$, $\Sigma^0 = 0.05I_3$, $n = 500$.
4. Case D: $\mu_1^0 = (0.8, 0.8)$, $\mu_2^0 = (0.8, -0.8)$, $\mu_3^0 = (-0.8, 0.8)$, $\Sigma^0 = 0.01I_3$, $n = 1500$.

Here, I_3 is the identity matrix of dimension 3. Additionally, the weight vector for all these cases is chosen as $p^0 = (p_1^0, p_2^0, p_3^0) = (0.4, 0.3, 0.3)$.

As mentioned above, a Dirichlet process prior with an uniform prior base measure H in the region $[-6, 6] \times [-6, 6]$, along with concentration parameter $\alpha = 1$. This choice of prior enables us to sample significantly larger numbers of components of the mixing measure than the true number of three components.

It is known that the contraction rate of mixing measures under location Gaussian DPMM is $\tilde{C}(\log(n)^{-1/2})$ with respect to the Wasserstein-2 norm, for some constant $\tilde{C}$ which depends on Σ^0

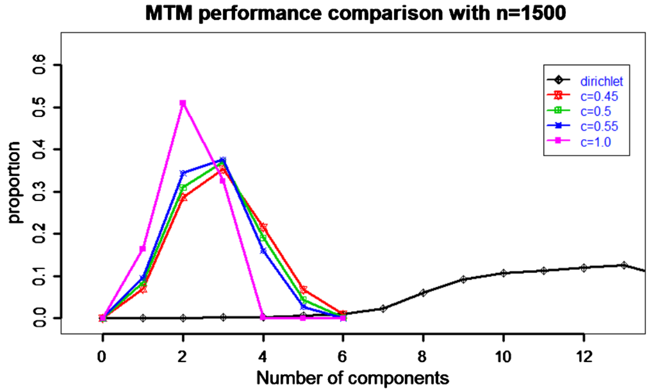

Figure 6. Case B.

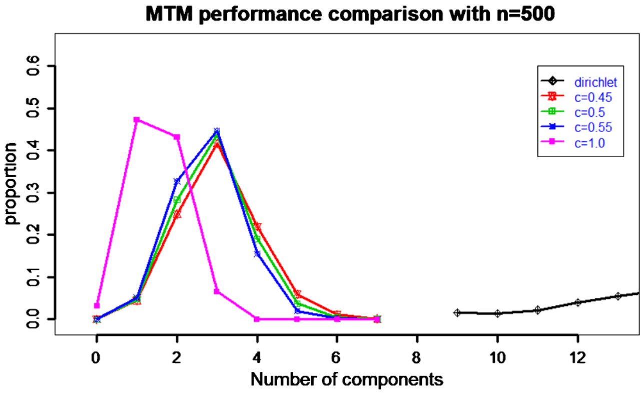

Figure 7. Case C.

(the covariance matrix), the location parameters μ_i^0 and the weights p_i^0 [37]. For our purpose, in order for ω_n to satisfy Equation (9), we may choose any ω_n as long as $\frac{\omega_n}{\log(n)^{-1/2}} \to \infty$. We selected $\omega_n = (\frac{\log(\log(n))}{(\log(n))})^{1/2}$ for all our applications of the MTM algorithm.

The MTM algorithm is provably consistent (in the asymptotic sense) for all chosen constants $c > 0$. In practice for n being fixed, the input c to Algorithm 1 should be chosen so that $\frac{\tilde{C}}{(\log(\log(n)))^{1/2}} \leq c$. Moreover, for finite n it is not expected that the posterior probability for $k = k_0$ is close to 1. However, for identifying the number of components the posterior mode provides a reasonable estimate. In particular, $(1 - \sum_{i=1}^{3} \frac{c}{p_i^0})$ forms a useful lower bound on the posterior mass at the true parameter as identified in Equation (13) in the supplement. To identify $k = k_0$ consistently using the posterior mode safely, one needs to choose $c < c_0$, with c_0 satisfying $(1 - \sum_{i=1}^{3} \frac{c_0}{p_i^0}) > 1/2$. The exact computation of the upper bound c_0 and the lower bound $\frac{\tilde{C}}{(\log(\log(n)))}$ for c may be unrealistic but a reasonable estimate may be possible. Nonetheless, we simply considered a large range of c and show there is a range where we can robustly identify the true number of components via the posterior mode.

For the DP mixture's posterior computation, we make use of the non-conjugate split-merge sampler of Jain and Neal [25] with (5, 1, 1, 5) scheme, that is, 5 scans to reach the split launch state, 1 split-merge move per iteration, 1 Gibbs scan per iteration, and 5 moves to reach the merge launch state. We run our experiments for two settings corresponding to sample sizes 500 and 1500. The sampler had 2000 burn-in iterations followed by 18,000 sample iterations (a total 20,000), with each 10th iteration being counted.

The experiments run for DP mixture-based sampler, followed by application of the MTM procedure for 4 different values of the tuning parameter c in Algorithm 1, namely, for $c = 0.45, 0.5, 0.55, 1.0$. The proportional frequencies are plotted in Figure 5 and Figure 6 respectively, along with the proportional frequencies for DP mixture. The uniform base measure for the Dirichlet Process prior is chosen so as to enable easier creation of newer components in the split-merge scheme. As a consequence the DP mixture's posterior yields quite bad results as far as the number of mixture components is concerned. However, even under that case, we can recover the true number of components by considering the mode of the frequency distribution after an application of the MTM algorithm on the posterior samples, with appropriate constant c. It is expected, however, that a large choice of c would underestimate the number of components. This is also what is observed from the simulations, where the procedure breaks down when $c = 1.0$.

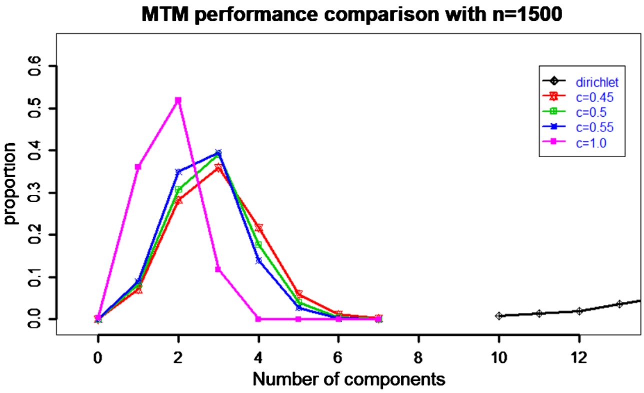

Figure 8. Case D.

We perform the experiments under four different settings of data populations. In particular, Figure 7 consists of data generated from mixture of Gaussians with more widely spread location parameter values. In this case, it is expected that the convergence to the true number of components via Algorithm 1 will be faster for the posterior mode, in comparison to the situation where the location parameters are closer together. This is indeed what is observed in our simulations. The value of the covariance matrix Σ^0, on the other hand does not seem to noticeably affect the results. This is again expected, since the prior support $[-6, 6] \times [-6, 6]$ is quite large in comparison to the eigenvalues of the covariance matrix chosen.

Funding

This research is supported in part by grants NSF CAREER DMS-1351362, NSF CNS-1409303, a research gift from Adobe Research and a Margaret and Herman Sokol Faculty Award.

Supplementary Material

Supplement "On posterior contraction of parameters and interpretability in Bayesian mixture modeling" (DOI: 10.3150/20-BEJ1275SUPP; .pdf). In this supplemental material, we provide self-contained proofs of several key results in the paper.

References

[1] Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

[2] Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

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

[4] 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

[5] 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

[6] Dacunha-Castelle, D. and Gassiat, E. (1997). The estimation of the order of a mixture model. *Bernoulli* **3** 279–299. MR1468306 https://doi.org/10.2307/3318593

[7] Eltoft, T., Kim, T. and Lee, T. (2006). On the multivariate Laplace distribution. *IEEE Signal Process*. *Lett*. **13** 300–303.

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

[9] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann*. *Statist*. **19** 1257–1272. MR1126324 https://doi.org/10.1214/aos/1176348248

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

[11] Figueiredo, M. and Jain, A.K. (1993). Unsupervised learning of finite mixture models. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **24**.

[12] Fúquene, J., Steel, M. and Rossell, D. (2019). On choosing mixture components via non-local priors. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **81** 809–837. MR4025398 https://doi.org/10.1111/rssb.12333

[13] Gao, F. and van der Vaart, A. (2016). Posterior contraction rates for deconvolution of Dirichlet–Laplace mixtures. *Electron*. *J*. *Stat*. **10** 608–627. MR3471990 https://doi.org/10.1214/16-EJS1119

[14] Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann*. *Statist*. **27** 143–158. MR1701105 https://doi.org/10.1214/aos/1018031105

[15] Ghosal, S., Ghosh, J.K. and van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann*. *Statist*. **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

[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] Green, P.J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scand*. *J*. *Stat*. **28** 355–375. MR1842255 https://doi.org/10.1111/1467-9469.00242

[19] Guha, A., Ho, N. and Nguyen, X. (2021). Supplement to "On posterior contraction of parameters and interpretability in Bayesian mixture modeling." https://doi.org/10.3150/20-BEJ1275SUPP

[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] Hjort, N., Holmes, C., Mueller, P. and Walker, S. (2010). *Bayesian Nonparametrics*: *Principles and Practice*. Cambridge University Press.

[22] 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

[23] 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

[24] Ishwaran, H., James, L.F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J*. *Amer*. *Statist*. *Assoc*. **96** 1316–1332. MR1946579 https://doi.org/10.1198/016214501753382255

[25] Jain, S. and Neal, R.M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Anal*. **2** 445–472. MR2342168 https://doi.org/10.1214/07-BA219

[26] Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *J*. *Amer*. *Statist*. *Assoc*. **90** 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572

[27] Kleijn, B.J.K. and van der Vaart, A.W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann*. *Statist*. **34** 837–877. MR2283395 https://doi.org/10.1214/009053606000000029

[28] Kleijn, B.J.K. and van der Vaart, A.W. (2012). The Bernstein–Von-Mises theorem under misspecification. *Electron*. *J*. *Stat*. **6** 354–381. MR2988412 https://doi.org/10.1214/12-EJS675

[29] Leroux, B.G. (1992). Consistent estimation of a mixing distribution. *Ann*. *Statist*. **20** 1350–1360. MR1186253 https://doi.org/10.1214/aos/1176348772

[30] Lindsay, B. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, CA: IMS.

[31] Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann*. *Statist*. **12** 351–357. MR0733519 https://doi.org/10.1214/aos/1176346412

[32] MacEachern, S. and Mueller, P. (1998). Estimating mixture of Dirichlet process models. *J*. *Comput*. *Graph*. *Statist*. **7** 223–238.

[33] McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models*: *Inference and Applications to Clustering*. *Statistics*: *Textbooks and Monographs* **84**. New York: Dekker. MR0926484

[34] Mengersen, K.L., Robert, C.P. and Titterington, D.M., eds. (2011). *Mixtures*: *Estimation and Applications*. *Wiley Series in Probability and Statistics*. Chichester: Wiley. MR2867716 https://doi.org/10.1002/9781119995678

[35] 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

[36] 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

[37] 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

[38] Nguyen, X. (2015). Posterior contraction of the population polytope in finite admixture models. *Bernoulli* **21** 618–646. MR3322333 https://doi.org/10.3150/13-BEJ582

[39] Nobile, A. (1994). Bayesian analysis of finite mixture distributions. Ph.D. thesis, Dept. Statistics, Carnegie Mellon University, Pittsburgh, PA.

[40] 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

[41] Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

[42] Rodríguez, A., Dunson, D.B. and Gelfand, A.E. (2008). The nested Dirichlet process. *J*. *Amer*. *Statist*. *Assoc*. **103** 1131–1144. MR2528831 https://doi.org/10.1198/016214508000000553

[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. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal*. **9** 475–520. MR3217004 https://doi.org/10.1214/14-BA863

[45] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist*. *Sinica* **4** 639–650. MR1309433

[46] Shen, W., Tokdar, S.T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441 https://doi.org/10.1093/biomet/ast015

[47] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann*. *Statist*. **29** 687–714. MR1865337 https://doi.org/10.1214/aos/1009210686

[48] 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

[49] Tang, J., Meng, Z., Nguyen, X., Mei, Q. and Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the International Conference on Machine Learning*.

[50] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006). Hierarchical Dirichlet processes. *J*. *Amer*. *Statist*. *Assoc*. **101** 1566–1581. MR2279480 https://doi.org/10.1198/016214506000000302

[51] Villani, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Berlin: Springer. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[52] Walker, S.G., Lijoi, A. and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann*. *Statist*. **35** 738–746. MR2336866 https://doi.org/10.1214/009053606000001361

[53] Xie, F. and Xu, Y. (2020). Bayesian repulsive Gaussian mixture model. *J*. *Amer*. *Statist*. *Assoc*. **115** 187–203. MR4078456 https://doi.org/10.1080/01621459.2018.1537918

[54] Yurochkin, M., Guha, A. and Nguyen, X. (2017). Conic scan and cover algorithms for nonparametric topic modeling. In *NIPS* **31**.

[55] Zhang, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann*. *Statist*. **18** 806–831. MR1056338 https://doi.org/10.1214/aos/1176347627

Received October 2019 and revised May 2020