Power posteriors do not reliably learn the number of components in a finite mixture

Diana Cai*
Dept. of Computer Science
Princeton University
Princeton, NJ 08544
dcai@cs.princeton.edu

Trevor Campbell*
Dept. of Statistics
University of British Columbia
Vancouver, BC V6T 1Z4
trevor@stat.ubc.ca

Tamara Broderick
CSAIL
MIT
Cambridge, MA 02139
tbroderick@csail.mit.edu

Abstract

Scientists and engineers are often interested in learning the number of subpopulations (or components) present in a data set. Data science folk wisdom tells us that a finite mixture model (FMM) with a prior on the number of components will fail to recover the true, data-generating number of components under model misspecification. But practitioners still widely use FMMs to learn the number of components, and statistical machine learning papers can be found recommending such an approach. Increasingly, though, data science papers suggest potential alternatives beyond vanilla FMMs, such as power posteriors, coarsening, and related methods. In this work we start by adding rigor to folk wisdom and proving that, under even the slightest model misspecification, the FMM component-count posterior diverges: the posterior probability of any particular finite number of latent components converges to 0 in the limit of infinite data. We use the same theoretical techniques to show that power posteriors with fixed power face the same undesirable divergence, and we provide a proof for the case where the power converges to a non-zero constant. We illustrate the practical consequences of our theory on simulated and real data. We conjecture how our methods may be applied to lend insight into other component-count robustification techniques.

1 Introduction

In any probabilistic model, simplifying assumptions are made out of necessity. Some models and conclusions nonetheless lead to reliable and useful inference. But some misspecification will have undesirable consequences. In this work we focus on the case of mixture models — which are widely used to discover latent groups, or components, within a population. Often the number of components is unknown in advance, and one of the principal inferential goals is estimating and interpreting this number. For example, practitioners might wish to find the number of latent genetic populations (Pritchard et al., 2000; Lorenzen et al., 2006; Huelsenbeck and Andolfatto, 2007), gene tissue profiles (Yeung et al., 2001; Medvedovic and Sivaganesan, 2002), cell types (Chan et al., 2008; Prabhakaran et al., 2016), haplotypes (Xing et al., 2006), switching Markov regimes in US dollar exchange rate data (Otranto and Gallo, 2002), gamma-ray burst types (Mukherjee et al., 1998), or segmentation regions in an image (e.g., tissue types in an MRI scan (Banfield and Raftery, 1993)). In all of these cases, it is typical to assume some standard parametric form for the mixture likelihoods, and we can expect that this form will be at least slightly misspecified.

A common approach to learning mixture components and their cardinality is to use a finite mixture model with a prior on the number of components (FMM) (Nobile, 1994; Stephens, 2000; Green and Richardson, 2001; Nobile, 2004, 2007; Nobile and Fearnside, 2007; Miller and Harrison, 2013, 2014,

*First authorship is shared jointly by D. Cai and T. Campbell.

1st I Can't Believe It's Not Better Workshop (ICBINB@NeurIPS 2020), Vancouver, Canada.

2018; Grazian et al., 2020). Typically the prior on component counts is supported on all possible strictly-positive integers, and we compute a posterior distribution over the number of components (Pritchard et al., 2000; Lorenzen et al., 2006; Huelsenbeck and Andolfatto, 2007). But empirical evidence (Miller and Dunson, 2019, e.g.) suggests that the posterior number of components is very sensitive to misspecification of the component likelihoods. Indeed, data science folk wisdom tells us that finite mixture models will tend to choose too many components in practice (Frühwirth-Schnatter, 2006, Chapter 7).

One proposal to address this mis-estimation is to use an α-posterior, or *power posterior*, where we replace the likelihood with the same likelihood but raised to a fixed power $\alpha > 0$, often between 0 and 1 (Grünwald, 2006; Grünwald and van Ommen, 2017; Royall and Tsou, 2003; Ghosh and Sudderth, 2012; Holmes and Walker, 2017). Power posteriors have much more general application than just to mixture models; we focus only on mixture models here and note that behavior of power posteriors for other models may be very different than for mixture models. In a separate line of work, Miller and Dunson (2019) propose a *coarsened posterior*, which they show can be closely approximated by a variant of the power posterior with exponent $\alpha_N \to 0$ as the number of data points $N \to \infty$; in fact, they use this approximation in all of their experiments. Note that we use the terminology "power posterior" or α-posterior throughout to refer to the *fixed* power case, and the terminology α_N-posterior to refer to the case where the power may depend on N.

In the present work, we start by adding rigor to data science folk wisdom and prove that, under even the slightest model misspecification, the posterior number of components in a classic finite mixture analysis diverges; that is, the posterior probability of any particular finite number of latent components converges to 0 in the limit of infinite data. We use a similar analysis to lend insight into the power posterior for finite mixtures; in particular, we find that the component-count power posterior, with power that is constant in N, diverges in the same way. We give a proof that an α_N-posterior, where α_N converges to a nonzero constant, also diverges. We support our theory with experiments on simulated and real data. We leave analysis of powers $\alpha_N \to 0$ to future work but provide some discussion here.

2 Main results

We begin with a brief description of the finite mixture model we consider in this work and a statement of the main results. We defer precise details to Section 3.

Let g be a mixing measure $g := \sum_{j=1}^{k} w_j \delta_{\theta_j}$ on a parameter space Θ with $w_j \in [0, 1]$ and $\sum_{j=1}^{k} w_j = 1$, and let $\Psi = \{\psi_\theta : \theta \in \Theta\}$ be a family of component distributions dominated by a σ-finite measure μ. Then we can express a finite mixture f of the components as

$$f = \int_\Theta \psi_\theta dg(\theta) = \sum_{j=1}^{k} w_j \psi_{\theta_j}.$$

Consider a Bayesian model with a prior distribution Π on the set of all mixing measures $\mathbb{G}$ on Θ with finitely many atoms, i.e., $g \sim \Pi$, and likelihood corresponding to conditionally i.i.d. data from $f = \int \psi_\theta dg(\theta)$. The model assumes the likelihood is f, but the model is *misspecified*; i.e., the observations $X_{1:N} := (X_1, \ldots, X_N)$ are actually generated conditionally i.i.d. from f_0, which is not itself a finite mixture of distributions in Ψ. For example, the likelihood f might be a finite mixture of Gaussian distributions, but f_0 might represent a finite mixture of Laplace distributions.

Let $\Pi^{(\alpha)}(k \mid X_{1:N})$ denote the posterior marginal number of components induced by raising the likelihood to a fixed power $\alpha > 0$. Our main result is that under this misspecification of the likelihood, for any $\alpha \in (0, 1]$, the α-posterior on the number of components $\Pi^{(\alpha)}(k \mid X_{1:N})$ *diverges*; i.e., for any finite $k \in \mathbb{N}$, $\Pi^{(\alpha)}(k \mid X_{1:N}) \to 0$ as $N \to \infty$.

We make only two requirements of the mixture model to guarantee this result: (1) the true data-generating distribution f_0 must be arbitrarily well-approximated by finite mixtures of Ψ, and (2) the family Ψ must satisfy mild regularity conditions that hold for popular mixture models (e.g., the family Ψ of Gaussians parametrized by mean and variance). We provide precise definitions of the assumptions needed for Theorem 2.1 to hold in Section 3, and a proof in Appendix A.1.

Theorem 2.1 (Main result)**.** *Suppose observations $X_{1:N}$ are generated i.i.d. from a distribution f_0 that is not a finite mixture of Ψ. Assume that:*

Assumption 3.1: f_0 *is in the KL-support of the prior* Π*, and*

Assumption 3.6: Ψ *is continuous, is mixture-identifiable, and has degenerate limits.*

Then for any $\alpha \in (0, 1]$*, the* α*-posterior on the number of components diverges; i.e., for all* $k \in \mathbb{N}$*,*

$$\Pi^{(\alpha)}(k \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 0 \quad f_0\text{-a.s.} \tag{1}$$

Note that the conditions of the theorem—although technical—are satisfied by a wide class of models used in practice. Assumption 3.1 requires that the prior Π places enough mass on mixtures near the true generating distribution f_0. Assumption 3.6 enforces regularity of the component family and is satisfied by many popular models used in practice, such as the multivariate Gaussian family (Proposition B.2) and, more generally, mixture-identifiable location-scale families (Proposition B.3).

While the result of Theorem 2.1 assumes that the model uses a fixed prior Π, in many practical modeling scenarios it is common to specify a prior Π_N that depends on observed data $X_{1:N}$ (e.g., to establish a parameter range). If f_0 satisfies a modified KL-support condition with respect to the sequence of priors Π_N, it is straightforward to show that the α-posterior number of components also diverges in this setting by Corollary A.5.

2.1 Extension to sequences α_N

Much of the existing literature on power posteriors focuses on the case where the power is fixed between 0 and 1 (Grünwald, 2006; Grünwald and van Ommen, 2017; Walker and Hjort, 2001; Royall and Tsou, 2003). In Theorem 2.1, we considered the asymptotic behavior of the α-posterior number of components with a fixed $\alpha \in (0, 1)$.

We also provide a result about the α_N-posterior number of components, where the sequence α_N may depend on the sample size N. Corollary 2.2 states that if α_N converges to a value between 0 and 1, then the α_N-posterior number of components diverges; a proof is presented in Appendix A.2.

Corollary 2.2. *Suppose the conditions of Theorem 2.1 hold. Let* $\alpha_N \to \alpha$*, where* $\alpha \in (0, 1)$*. Then for any* $k \in \mathbb{N}$*,*

$$\Pi^{(\alpha_N)}(k \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 0, \qquad f_0\text{-a.s.}$$

Note that the above statement does not include sequences α_N that converge to 0. A notable instance of a sequence $\alpha_N \to 0$ is studied by Miller and Dunson (2019) as an approximation to a coarsened posterior. Specifically, Miller and Dunson (2019) consider a particular sequence $\alpha_N = \gamma/(\gamma + N)$, where γ is a coarsening parameter. The interpretation of this particular sequence is that the posterior will, for any number N of observations, behave as though there were roughly γ number of data points. Additionally, characterizing the behavior of the α_N-posterior number of components for other rates of convergence of $\alpha_N \to 0$ beyond the one proposed by Miller and Dunson (2019) remains an open question.

3 Setup and assumptions in Theorem 2.1

This section makes the details of the modeling setup and each of the conditions in Theorem 2.1 precise.

3.1 Notation and setup

Let $\mathbb{X}$ and Θ be Polish spaces for the observations and parameters, respectively, and endow both with their Borel σ-algebra. For a topological space $(\cdot)$, let $\mathscr{C}(\cdot)$ be the bounded continuous functions from $(\cdot)$ into $\mathbb{R}$, and $\mathscr{P}(\cdot)$ be the set of probability measures on $(\cdot)$ endowed with the weak topology metrized by the Prokhorov distance d (Ghosal and van der Vaart, 2017, Appendix A, p. 508), and Borel σ-algebra. We use $f_i \Rightarrow f$ and $f_i \iff f_i'$ to denote $\lim_{i\to\infty} d(f_i, f) = 0$ and $\lim_{i\to\infty} d(f_i, f_i') = 0$, respectively, for $f_i, f_i', f \in \mathscr{P}(\cdot)$.

We assume that the family of distributions $\Psi = \{\psi_\theta : \theta \in \Theta\}$ is absolutely continuous with respect to a σ-finite base measure μ, i.e., $\psi_\theta \ll \mu$ for all $\theta \in \Theta$, and that for measurable $A \subseteq \mathbb{X}$, $\psi_\theta(A)$ is a measurable function on Θ. Define the measurable mapping $F : \mathscr{P}(\Theta) \to \mathscr{P}(\mathbb{X})$ from mixing

measures to mixtures of Ψ, $F(g) = \int \psi_\theta dg(\theta)$. Let $\mathbb{G}$ be the set of atomic probability measures on Θ with finitely many atoms, and let $\mathbb{F}$ be the set of finite mixtures of Ψ.

In the Bayesian finite mixture model from Section 2, a mixing measure $g \sim \Pi$ is generated from a prior measure Π on $\mathbb{G}$, and $f = F(g)$ is a likelihood distribution. The α-posterior distribution on the mixing measure is

$$\forall \text{ measurable } A \subseteq \mathbb{G}, \quad \Pi^{(\alpha)}(A \mid X_{1:N}) = \frac{\int_A \prod_{n=1}^N (\frac{df}{d\mu})^\alpha(X_n)\, d\Pi(g)}{\int_{\mathbb{G}} \prod_{n=1}^N (\frac{df}{d\mu})^\alpha(X_n)\, d\Pi(g)}, \tag{2}$$

where $\frac{df}{d\mu}$ is the density of $f = F(g)$ with respect to μ and $\alpha > 0$. The α-posterior on the mixing measure $g \in \mathbb{G}$ induces an α-posterior on the number of components $k \in \mathbb{N}$ by counting the number of atoms in g, and it also induces a posterior on mixtures $f \in \mathbb{F}$ via the pushforward through the mapping F. We overload the notation $\Pi^{(\alpha)}(\cdot \mid X_{1:N})$ to refer to all of these α-posterior distributions and $\Pi(\cdot)$ to refer to prior distributions; the meaning should be clear from context.

3.2 Model assumptions

The first assumption of Theorem 2.1 is that while the true data-generating distribution f_0 is not contained in the model class $f_0 \notin \mathbb{F}$, it lies on the boundary of the model class. In particular, we assume f_0 is in the *KL-support* of the prior Π. Denote the Kullback-Leibler (KL) divergence between probability measures f_0 and f as

$$\mathrm{KL}(f_0, f) := \begin{cases} \int \log\left(\frac{df_0}{df}\right) df_0 & f_0 \ll f \\ \infty & \text{otherwise} \end{cases}.$$

Assumption 3.1. *For all $\epsilon > 0$, the prior distribution Π satisfies*

$$\Pi(f \in \mathbb{F} : \mathrm{KL}(f_0, f) < \epsilon) > 0.$$

We use Assumption 3.1 in the proof of Theorem 2.1 primarily to ensure that the Bayesian posterior is consistent for f_0. Note that Assumption 3.1 is fairly weak in practice. Intuitively, it just requires that the family Ψ is rich enough so that mixtures of Ψ can approximate f_0 arbitrarily well, and that the prior Π places sufficient mass on those mixtures close to f_0. For Bayesian mixture modeling, Ghosal et al. (1999, Theorem 3), Tokdar (2006, Theorem 3.2), Wu and Ghosal (2008, Theorem 2.3), and Petralia et al. (2012, Theorem 1) provide conditions needed to satisfy Assumption 3.1.

The second assumption of Theorem 2.1 is that the family of component distributions Ψ is well-behaved. This assumption has three stipulations. First, the mapping $\theta \mapsto \psi_\theta$ must be continuous; this condition essentially asserts that similar parameter values θ must result in similar component distributions ψ_θ.

Definition 3.2. *The family Ψ is* continuous *if the map $\theta \mapsto \psi_\theta$ is continuous.*

Second, the family Ψ must be *mixture-identifiable*, which guarantees that each mixture $f \in \mathbb{F}$ is associated with a unique mixing measure $G \in \mathbb{G}$.

Definition 3.3 (Teicher (1961, 1963))**.** *The family Ψ is* mixture-identifiable *if the mapping $F(g) = \int \psi_\theta dg(\theta)$ restricted to finite mixtures $F : \mathbb{G} \to \mathbb{F}$ is a bijection.*

In practice, one should always use an identifiable mixture model for clustering; without identifiability, the task of learning the number of components is ill posed. And many models satisfy mixture-identifiability, such as finite mixtures of the multivariate Gaussian family (Yakowitz and Spragins, 1968), the Cauchy family (Yakowitz and Spragins, 1968), the gamma family (Teicher, 1963), the generalized logistic family, the generalized Gumbel family, the Weibull family, and von Mises family (Ho and Nguyen, 2016, Theorem 3.3). A number of authors (e.g., Chen (1995); Ishwaran et al. (2001); Nguyen (2013); Ho and Nguyen (2016); Guha et al. (2019); Heinrich and Kahn (2018)) appeal to stronger notions of identifiability for mixtures than Definition 3.3. But, to show posterior divergence in the present work, we do not require conditions stronger than Definition 3.3.

The third stipulation—that the family Ψ has *degenerate limits*—guarantees that a "poorly behaved" sequence of parameters $(\theta_i)_{i\in\mathbb{N}}$ creates a likewise "poorly behaved" sequence of distributions $(\psi_{\theta_i})_{i\in\mathbb{N}}$.

This condition allows us to rule out such sequences in the proof of Theorem 2.1, and is the essential regularity condition to guarantee that a sequence of finite mixtures of at most k components cannot approximate f_0 arbitrarily closely.

Definition 3.4. *A sequence of distributions $(\psi_i)_{i=1}^{\infty}$ is* μ-wide *if for any closed set C such that $\mu(C) = 0$ and any sequence of distributions $(\phi_i)_{i=1}^{\infty}$ such that $\psi_i \iff \phi_i$,*

$$\limsup_{i\to\infty} \phi_i(C) = 0.$$

Definition 3.5. *The family Ψ has* degenerate limits *if for any tight, μ-wide sequence $(\psi_{\theta_i})_{i\in\mathbb{N}}$, we have that $(\theta_i)_{i\in\mathbb{N}}$ is relatively compact.*

The contrapositive of Definition 3.5 provides an intuitive explanation of the condition: as $i \to \infty$, for any sequence of parameters θ_i that eventually leaves every compact set $K \subseteq \Theta$, either the ψ_{θ_i} become "arbitrarily flat" (not tight) or "arbitrarily peaky" (not μ-wide). For example, consider the family Ψ of Gaussians on $\mathbb{R}$ with Lebesgue measure μ. If the variance of ψ_{θ_i} shrinks as i grows, the sequence of distributions converges weakly to a sequence of point masses (not dominated by the Lebesgue measure). If either the variance or the mean diverges, the distributions flatten out and the sequence is not tight. We use the fact that these are the only two possibilities when a sequence of parameters is poorly behaved (not relatively compact) in the proof of Theorem 2.1.

These three stipulations together yield Assumption 3.6.

Assumption 3.6. *The mixture component family Ψ is continuous, is mixture-identifiable, and has degenerate limits.*

4 Proof of Theorem 2.1 and extension to sequences α_N

4.1 Proof of Theorem 2.1

The proof has two essential steps. The first is to show that for any $\alpha \in (0, 1]$, the Bayesian posterior is weakly consistent for the mixture f_0; i.e., for any weak neighborhood U of f_0 the sequence of posterior distributions satisfies

$$\Pi^{(\alpha)}(U \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1, \quad f_0\text{-a.s.} \tag{3}$$

Weak consistency for f_0 is guaranteed directly by Assumption 3.1 and the fact that Ψ is dominated by a σ-finite measure μ. For $\alpha = 1$, Assumption 3.1 implies weak consistency for f_0, i.e., Equation (3) (Ghosh and Ramamoorthi, 2003, Theorem 4.4.2). For $\alpha \in (0, 1)$, if Assumption 3.1 holds, by Ghosal and van der Vaart (2017, Theorem 6.54), with f_0-probability 1, for any Hellinger neighborhood V of f_0 and any $\alpha \in (0, 1)$, $\Pi^{(\alpha)}(V|X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1$. Thus, since $V \subseteq U$,

$$\Pi^{(\alpha)}(U \mid X_{1:N}) \geq \Pi^{(\alpha)}(V \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1, \qquad f_0\text{-a.s.}$$

The second step is to show that for any finite $k \in \mathbb{N}$, there exists a weak neighborhood U of f_0 containing no mixtures of the family Ψ with at most k components. Together, these steps show that the posterior probability of the set of all k-component mixtures converges to 0 f_0-a.s. as the amount of observed data grows.

We provide a proof of the second step. To begin, note that Assumption 3.1 has two additional implications about f_0 beyond Equation (3). First, f_0 must be absolutely continuous with respect to the dominating measure μ; if it were not, then there exists a measurable set A such that $f_0(A) > 0$ and $\mu(A) = 0$. Since μ dominates Ψ, any $f \in \mathbb{F}$ satisfies $f(A) = 0$. Therefore $\mathrm{KL}(f_0, f) = \infty$, and the prior support condition cannot hold. Second, it implies that f_0 can be arbitrarily well-approximated by finite mixtures under the weak metric, i.e., there exists a sequence of finite measures $f_i \in \mathbb{F}, i \in \mathbb{N}$ such that $f_i \Rightarrow f_0$ as $i \to \infty$. This holds because $\sqrt{\frac{1}{2}\mathrm{KL}(f_0, f)} \geq \mathrm{TV}(f_0, f) \geq d(f_0, f)$.

Now suppose the contrary of the claim for the second step, i.e., that there exists a sequence $(f_i)_{i=1}^{\infty}$ of mixtures of at most k components from Ψ such that $f_i \Rightarrow f_0$. By mixture-identifiability, we have a sequence of mixing measures g_i with at most k atoms such that $F(g_i) = f_i$. Suppose first that the

atoms of the sequence $(g_i)_{i\in\mathbb{N}}$ either stay in a compact set or have weights converging to 0. More precisely, suppose there exists a compact set $K \subseteq \Theta$ such that

$$g_i\left(\Theta \setminus K\right) \to 0. \tag{4}$$

Decompose each $g_i = g_{i,K} + g_{i,\Theta\setminus K}$ such that $g_{i,K}$ is supported on K and $g_{i,\Theta\setminus K}$ is supported on $\Theta \setminus K$. Define the sequence of probability measures $\hat{g}_{i,K} = \frac{g_{i,K}}{g_{i,K}(\Theta)}$ for sufficiently large i such that the denominator is nonzero. Then Equation (4) implies

$$F\left(\hat{g}_{i,K}\right) \Rightarrow f_0.$$

Since Ψ is continuous and mixture-identifiable, the restriction of F to the domain $\mathbb{G}$ is continuous and invertible; and since K is compact, the elements of $(\hat{g}_{i,K})_{i\in\mathbb{N}}$ are contained in a compact set $\mathbb{G}_K \subseteq \mathbb{G}$ by Prokhorov's theorem (Ghosal and van der Vaart, 2017, Theorem A.4). Therefore $F(\mathbb{G}_K) = \mathbb{F}_K$ is also compact, and the map F restricted to the domain $\mathbb{G}_K$ is uniformly continuous with a uniformly continuous inverse by Rudin (1976, Theorems 4.14, 4.17, 4.19). Next since $F(\hat{g}_{i,K}) \Rightarrow f_0$, the sequence $F(\hat{g}_{i,K})$ is Cauchy in $\mathbb{F}_K$; and since F^{-1} is uniformly continuous on $\mathbb{F}_K$, the sequence $\hat{g}_{i,K}$ must also be Cauchy in $\mathbb{G}_K$. Since $\mathbb{G}_K$ is compact, $\hat{g}_{i,K}$ converges in $\mathbb{G}_K$. Lemma A.1 below guarantees that the convergent limit g_K is also a mixing measure with at most k atoms; continuity of F implies that $F(g_K) = f_0$, which is a contradiction, since by assumption f_0 is not representable as a finite mixture of Ψ.

Now we consider the remaining case: for all compact sets $K \subseteq \Theta$, $g_i(\Theta\setminus K) \not\to 0$. Therefore there exists a sequence of parameters $(\theta_i)_{i=1}^{\infty}$ that is not relatively compact such that $\limsup_{i\to\infty} g_i(\{\theta_i\}) > 0$. By Assumption 3.6, the sequence $(\psi_{\theta_i})_{i\in\mathbb{N}}$ is either not tight or not μ-wide. If $(\psi_{\theta_i})_{i\in\mathbb{N}}$ is not tight then $f_i = F(g_i)$ is not tight, and by Prokhorov's theorem f_i cannot converge to a probability measure, which contradicts $f_i \Rightarrow f_0$. If $(\psi_{\theta_i})_{i\in\mathbb{N}}$ is not μ-wide then $f_i = F(g_i)$ is not μ-wide. Denote $(\phi_i)_{i\in\mathbb{N}}$ to be the singular sequence associated with $(f_i)_{i\in\mathbb{N}}$ and C to be the closed set such that $\limsup_{i\to\infty} \phi_i(C) > 0$, $\mu(C) = 0$, and $\phi_i \iff f_i$ per Definition 3.4. Since $f_0 \ll \mu$, $f_0(C) = 0$. But $f_i \Rightarrow f_0$ implies that $\phi_i \Rightarrow f_0$, so $\limsup_{i\to\infty} \phi_i(C) = f_0(C) = 0$ by the Portmanteau theorem (Ghosal and van der Vaart, 2017, Theorem A.2). This is a contradiction.

5 Experiments

For all experiments below, we use a finite mixture model with a Gaussian component family and a conjugate prior. In particular, consider number of components k, mixture weights $p \in \mathbb{R}^k$, Gaussian component precisions $\tau \in \mathbb{R}_+^k$ and means $\theta \in \mathbb{R}^k$, labels $Z \in \{1,\dots,k\}^N$, and data $X \in \mathbb{R}^N$. Then the probabilistic generative model is

$$\begin{aligned} k &\sim \text{Geom}(r) & w &\sim \text{Dirichlet}_k(\gamma,\dots,\gamma) \\ \tau_j &\overset{\text{i.i.d.}}{\sim} \text{Gam}(\alpha,\beta) & \theta_j &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(m,\kappa^{-1}) \\ Z_n &\overset{\text{i.i.d.}}{\sim} \text{Categorical}(w) & X_n &\overset{\text{ind}}{\sim} \mathcal{N}(\theta_{z_n},\tau_{z_n}^{-1}), \end{aligned}$$

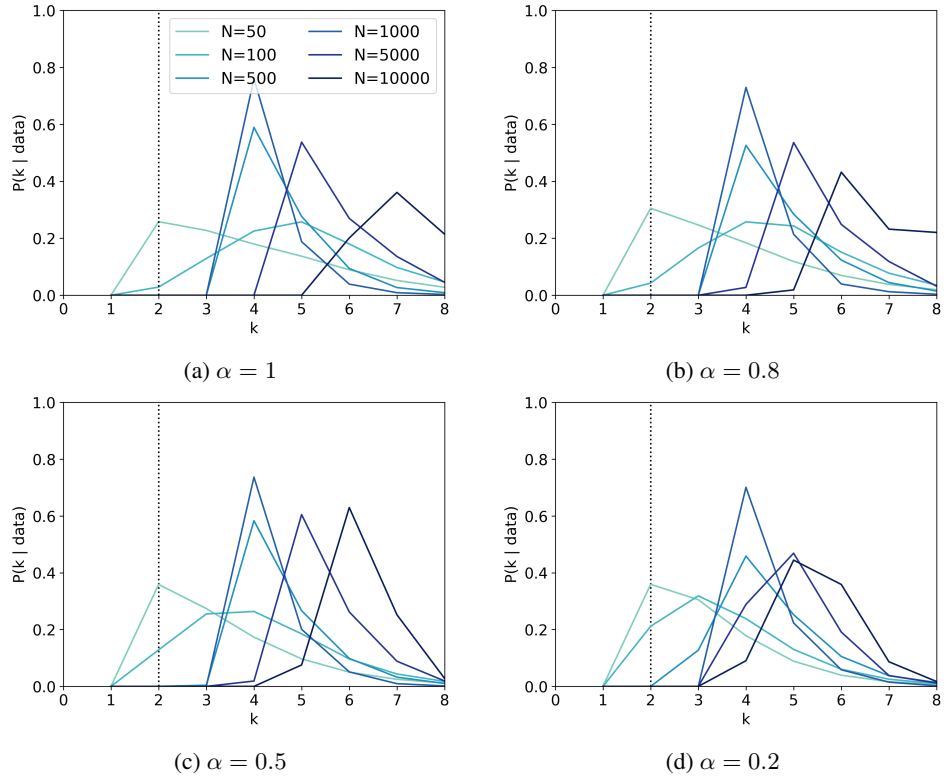
where j ranges from $1,\dots,k$, and n ranges from $1,\dots,N$.

For posterior inference, we used a Gibbs sampler (Miller and Harrison, 2018, Sec. 7.2.2, Algorithm 1), coupled with the approximation described in Miller and Dunson (2019, Section 5). Note that we use this model primarily to illustrate the problem of α-posterior divergence under model misspecification; it should not be interpreted as a carefully-specified model for the data examples that we study. Also note that while the empirical examples below involve Gaussian FMMs, our theory applies to a more general class of component distributions.

In this section, we consider the behavior of the α-posterior only for fixed $\alpha \in (0,1)$. In Appendix C.2, we show the α_N-posterior number of components for $\alpha_N \to \alpha \in (0,1)$ exhibits similar divergent behavior.

5.1 Synthetic mixture data

Here we study the effects of varying data set sizes under a misspecified model. We generated data sets of increasing size $N \in \{50, 100, 500, 1000, 5000, 10000\}$ from a 2-component Laplace mixture models, where the 2-component distributions have means $(-5, 5)$, scales $(1.5, 1)$, and mixing weights



(a) $\alpha = 1$ (b) $\alpha = 0.8$

(c) $\alpha = 0.5$ (d) $\alpha = 0.2$

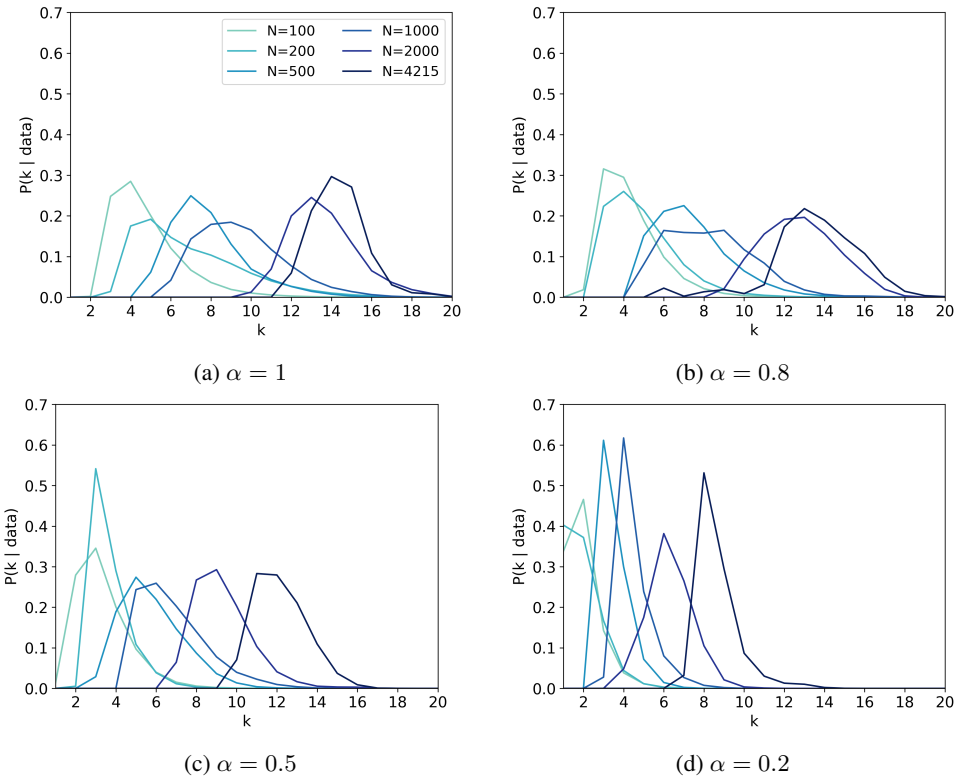
Figure 1: Synthetic data generated from a 2-component Laplace mixture model. Curves are α-posteriors on number of components (with fixed α) as N varies. The vertical black dotted line denotes the generating number of components.

$(0.4, 0.6)$. We generated the sequence of data sets such that each was a subset of the next, larger data set in the sequence. Following Miller and Harrison (2018, Section 7.2.1), we set the hyperparameters of the Bayesian finite mixture model as follows: $m = \frac{1}{2}(\max_{n\in[\tilde{N}]} X_n + \min_{n\in[\tilde{N}]} X_n)$ where $\tilde{N} = 10{,}000$, $\kappa = (\max_{n\in[\tilde{N}]} X_n - \min_{n\in[\tilde{N}]} X_n)^{-2}$, $\alpha = 2$, $r = 0.1$, $\gamma = 1$, and $\beta = 1$. We ran a total of 150,000 Markov chain Monte Carlo iterations per data set; we discarded the first 50,000 iterations as burn-in.

In Figure 1, we show the α-posterior number of components resulting from fixed $\alpha = 1, 0.8, 0.5, 0.2$. Note that $\alpha = 1$ is the usual posterior distribution on the number of components. The figures show that as α decreases, the posterior mass tends to shift to smaller numbers of components and also becomes less concentrated. However, as in the usual posterior ($\alpha = 1$), the α-posterior still diverges, though more slowly with lower values of α.

5.2 Galaxy mixture data

Mixture models are used in astronomy to characterize stellar populations (Nemec and Nemec, 1991), including analysis of star and galaxy clusters. We study the Shapley galaxy data set (Drinkwater et al., 2004), which contains measurements of redshifts (i.e., velocities in km/s) for 4215 galaxies in the Shapley Concentration regions. To examine the effect of increasing data set size on inferential results, we randomly sampled subsets of increasing size without replacement with $N \in \{100, 200, 500, 1000, 2000, 4215\}$; each smaller subset was contained in the next larger data set. We set the hyperparameters of the Bayesian finite mixture model as follows: $m = \frac{1}{2}(\max_{n\in[\tilde{N}]} X_n + \min_{n\in[\tilde{N}]} X_n)$ where $\tilde{N} = 4215$, $\kappa = (\max_{n\in[\tilde{N}]} X_n - \min_{n\in[\tilde{N}]} X_n)^{-2}$, $\alpha = 2$, $r = 0.1$, $\gamma = 1$, and $\beta = 1$. We ran a total of 100,000 Markov chain Monte Carlo iterations per data set; we discarded the first 50,000 iterations as burn-in.



(a) $\alpha = 1$ (b) $\alpha = 0.8$

(c) $\alpha = 0.5$ (d) $\alpha = 0.2$

Figure 2: Shapley galaxy data. Curves are α-posteriors on the number of components (with fixed α) as N varies.

The α-posterior number of components for this model is displayed in Figure 2. For $\alpha = 1$, we find that as we examine more and more data, the posterior diverges. Similar behavior occurs with fractional α-posteriors. With smaller values of α, the α-posterior diverges more slowly.

6 Discussion and future work

We have shown that the posterior distribution for the number of components in finite mixtures diverges when the mixture component family is misspecified. Our results show that this divergence holds even when using a power posterior, with fixed power between 0 and 1, for robustness. It follows that, with either a standard or power posterior, inferential results on the number of components will change substantively with more data, no matter how much data has been analyzed so far. This divergence calls into question the usefulness of FMMs, and power posterior FMMs, in applications.

A number of open questions remain. We here point to similar power-posterior behavior even when the power changes in N but converges to a constant in $(0, 1)$. However, Miller and Dunson (2019) consider powers that converge to 0 in the limit of $N \to \infty$. It remains to investigate this case, especially for more general sequences of powers converging to zero, beyond the particular sequence suggested by Miller and Dunson (2019).

Because our analysis is inherently asymptotic, it is possible that the α-posterior on the number of components may still provide useful inferences for a finite sample — even when the power is fixed or converging to a non-zero constant. But our results suggest care would need to be taken to account for the inferential dependence on data size.

A Proofs

A.1 Additional details for the proof of Theorem 2.1

Lemma A.1. *Suppose $\phi, (\phi_i)_{i\in\mathbb{N}}$ are Borel probability measures on a Polish space such that $\phi_i \Rightarrow \phi$ and $\sup_i |\operatorname{supp}\phi_i| \le k \in \mathbb{N}$. Then $|\operatorname{supp}\phi| \le k$.*

Proof. Suppose $|\operatorname{supp}\phi| > k$. Then we can find $k+1$ distinct points $x_1, \dots, x_{k+1} \in \operatorname{supp}\phi$. Pick any metric ρ on the Polish space, and denote the minimum pairwise distance between the points 2ϵ. Then for each point $j = 1, \dots, k+1$ define the bounded, continuous function $h_j(x) = 0 \vee \big(1 - \epsilon^{-1}\rho(x, x_j)\big)$. Since $x_j \in \operatorname{supp}\phi$, we have that $\int h_j d\phi > 0$. Weak convergence $\phi_i \Rightarrow \phi$ therefore implies $\min_{j=1,\dots,k+1} \liminf_{i\to\infty} \int h_j d\phi_i > 0$. But the h_j are nonzero on disjoint sets, and each ϕ_i only has k atoms; the pigeonhole principle yields a contradiction. □

A.2 Proof of Corollary 2.2

The only difference of the proof of Corollary 2.2 with the proof of Theorem 2.1 is the first step, i.e., weak consistency of the α_N-posterior for f_0. Here we provide a proof of the first step, which is a straightforward generalization of Ghosal and van der Vaart (2017, Theorem 6.54).

Let $p := \frac{df}{d\mu}$ and $p_0 := \frac{df_0}{d\mu}$ denote the density of f and f_0 with respect to μ, respectively.

For any weak neighborhood U of f_0, we can express the α_N-posterior as

$$\Pi^{(\alpha_N)}(U^c|X_{1:N}) = \frac{\int_{U^c}\prod_{n=1}^N p^{\alpha_N}(X_n)d\Pi(p)}{\int\prod_{n=1}^N p^{\alpha_N}(X_n)d\Pi(p)} = \frac{\int_{U^c}\prod_{n=1}^N (\frac{p}{p_0})^{\alpha_N}(X_n)d\Pi(p)}{\int\prod_{n=1}^N (\frac{p}{p_0})^{\alpha_N}(X_n)d\Pi(p)}. \tag{5}$$

Since f_0 is in the KL support of the prior Π, by Lemma A.2, for any $\beta > 0$ and $\alpha_N \to \alpha \in (0,1)$, the denominator satisfies

$$\liminf_{N\to\infty} e^{\beta N}\int\prod_{n=1}^N (\frac{p}{p_0})^{\alpha_N}(X_n)d\Pi(p) = \infty, \quad f_0\text{-a.s.}$$

Suppose V is a Hellinger neighborhood of f_0. Then by Lemma A.3 and since $V^c \supseteq U^c$, for some $\beta_0 > 0$, with f_0-probability 1,

$$e^{\beta_0 N}\int_{V^c}\prod_{n=1}^N (\frac{p}{p_0})^{\alpha_N}(X_n)d\Pi(p) \ge e^{\beta_0 N}\int_{U^c}\prod_{n=1}^N (\frac{p}{p_0})^{\alpha_N}(X_n)d\Pi(p) \overset{N\to\infty}{\longrightarrow} 0.$$

Finally, choose $\beta = \beta_0$, and so with f_0-probability 1, Equation (5) goes to 0 as $N \to \infty$.

We now prove the lemmas used for the proof above.

Lemma A.2. *Suppose that f_0 is in the KL support of the prior Π and and $\alpha_N \to \alpha \in (0,1)$ as $N \to \infty$, then for any $\beta > 0$,*

$$\liminf_{N\to\infty} e^{\beta N}\int\prod_{n=1}^N (\frac{p}{p_0})^{\alpha_N}(X_n)d\Pi(p) = \infty, \qquad f_0\text{-a.s.}$$

Proof. Let $K_\epsilon := \{f : \mathrm{KL}(f_0, f) < \epsilon\}$. First note that by monotonicity, the strong law of large numbers, and the KL condition, we can bound the integral as follows

$$\int\prod_{n=1}^N (\frac{p}{p_0})^{\alpha_N}(X_n)d\Pi(p) \ge \int_{K_\epsilon}\exp\left(\alpha_N\sum_{n=1}^N \log\left[\frac{p}{p_0}(X_n)\right]\right)d\Pi(p) \ge \int_{K_\epsilon}\exp\Big(-\alpha\epsilon N\Big)d\Pi(p).$$

Thus, Ghosh and Ramamoorthi (2003, Lemma 4.4.1) implies the the result. □

Lemma A.3. *Let $\alpha_N \to \alpha \in (0,1)$ as $N \to \infty$. Then for any Hellinger neighborhood V of f_0 and for some $\beta_0 > 0$,*

$$e^{\beta_0 N}\int_{V^c}\prod_{n=1}^N (\frac{p}{p_0})^{\alpha_N}(X_n)d\Pi(p) \overset{N\to\infty}{\longrightarrow} 0, \qquad f_0\text{-a.s.} \tag{6}$$

Proof. Denote the Hellinger transform as $\rho_\alpha(p, p_0) := \int p^\alpha p_0^{1-\alpha} d\mu$. By Fubini's theorem and because $f_0(p/p_0)^\alpha = \rho_\alpha(p, p_0)$, we have that

$$f_0\left(\int_{V^c} \prod_{n=1}^{N} (p/p_0)^{\alpha_N}(X_n) d\Pi(p)\right) = \int_{V^c} \rho_{\alpha_N}(p, p_0)^N d\Pi(p) \leq C \exp(-N\epsilon^2 \min(\alpha_N, 1-\alpha_N)).$$

The inequality above follows because for any $\alpha \in (0,1)$, $\min(\alpha, 1-\alpha)\, d_H^2(p, p_0) \leq -\log p_\alpha(p, p_0)$, where d_H denotes the Hellinger distance (Ghosal and van der Vaart, 2017, Lemma B.5) and because for any $p \in V^c$, $d_H^2(p, p_0) \geq \epsilon^2$. Applying Markov's inequality and Borel-Cantelli implies that for some $\beta_0 > 0$,

$$f_0\left(\int_{V^c} \prod_{n=1}^{N} (p/p_0)^{\alpha_N}(X_n) d\Pi(p) > e^{-\beta_0 N} \text{ i.o.}\right) = 0,$$

and so the conclusion holds. □

A.3 Extension to priors that vary with N

Our main result (i.e., Theorem 2.1) applies to the setting of a fixed prior Π. However, it is often natural to specify a prior distribution that changes with N (e.g., Roeder and Wasserman (1997), Richardson and Green (1997), and Miller and Harrison (2018, Section 7.2.1)). Corollary A.5 below demonstrates that a result nearly identical to Theorem 2.1 holds for priors that are allowed to vary with N, provided that f_0 is in the KL-support of the *sequence* of priors Π_N. The only difference is that our result in this case is slightly weaker: we show that the posterior number of components diverges in probability rather than almost surely.

Assumption A.4. *For all $\epsilon > 0$, the sequence of prior distributions Π_N satisfies*

$$\liminf_{N\to\infty} \Pi_N(f : \mathrm{KL}(f_0, f) < \epsilon) > 0.$$

Corollary A.5. *Suppose in the setting of Theorem 2.1 we replace Assumption 3.1 with Assumption A.4. Then for any $\alpha \in (0, 1]$, the α-posterior on the number of components diverges in f_0-probability: i.e., for all $k \in \mathbb{N}$,*

$$\Pi^{(\alpha)}(k \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 0 \quad \textit{in } f_0\textit{-probability}.$$

Proof. Since for any $\epsilon > 0$, $\liminf_{N\to\infty} \Pi_N(f : \mathrm{KL}(f_0, f) < \epsilon) > 0$, Ghosal and van der Vaart (2017, Theorem 6.54, Lemma 6.26) imply that the α-posterior is weakly consistent at f_0 in probability: i.e., for any weak neighborhood U of f_0,

$$\Pi^{(\alpha)}(U \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1 \quad \text{in } f_0\text{-probability}.$$

Assumption A.4 also implies that for sufficiently large N, f_0 is a weak limit of finite mixtures in $\mathbb{F}$. The remainder of the proof is identical to that of Theorem 2.1. □

B Example component families

Consider the multivariate Gaussian family $\Psi = \{\mathcal{N}(\nu, \Sigma) : \nu \in \mathbb{R}^d, \Sigma \in \mathbb{S}_{++}^d\}$ with parameter space $\Theta = \mathbb{R}^d \times \mathbb{S}_{++}^d$, equipped with the topology induced by the Euclidean metric. Let $(\lambda_j(\Sigma))_{j=1}^d$ denote the eigenvalues of the covariance matrix $\Sigma \in \mathbb{S}_{++}^d$ that satisfy $\infty > \lambda_1(\Sigma) \geq \ldots \geq \lambda_d(\Sigma) > 0$. Since the family of Gaussians is continuous and mixture-identifiable (Yakowitz and Spragins, 1968, Proposition 2), the main condition we need to verify is that the family has degenerate limits (Definition 3.5). A useful fact is that if a sequence of Gaussian distributions is tight, then the sequence of means and the eigenvalues of the covariance matrix is bounded.

Lemma B.1. *Let $(\psi_i)_{i\in\mathbb{N}}$ be a sequence of Gaussian distributions with mean $\nu_i \in \mathbb{R}^d$ and covariance $\Sigma_i \in \mathbb{S}_{++}^d$. If $(\psi_i)_{i\in\mathbb{N}}$ is a tight sequence of measures, then the sequences $(\nu_i)_{i\in\mathbb{N}}$ and $(\lambda_1(\Sigma_i))_{i\in\mathbb{N}}$ are bounded.*

Proof. Let Y_i denote a random variable with distribution ψ_i. For each covariance matrix Σ_i, consider its eigenvalue decomposition $\Sigma_i = U_i \Lambda_i U_i^\top$, where $U_i \in \mathbb{R}^{d\times d}$ is an orthonormal matrix and $\Lambda_i \in \mathbb{R}^{d\times d}$ is a diagonal matrix. Then the random variable $Z_i = U_i^\top Y_i$ has distribution $\mathcal{N}(U_i^\top \nu_i, \Lambda_i)$. If either $\|\nu_i\|_2 = \|U_i^\top \nu_i\|_2$ is unbounded or $\|\Lambda_i\|_F$ is unbounded, then Z_i is not tight (Billingsley (1986, Example 25.10)). Since Z_i and Y_i lie in any ball centered at the origin with the same probability, Y_i is not tight. □

Proposition B.2. *Let $\Psi = \{\mathcal{N}(\nu, \Sigma) : \nu \in \mathbb{R}^d, \Sigma \in \mathbb{S}_{++}^d\}$ be the multivariate Gaussian family, where $\mathbb{S}_{++}^d := \{\Sigma \in \mathbb{R}^{d\times d} : \Sigma = \Sigma^\top, \Sigma \succ 0\}$ is the set of $d \times d$ symmetric, positive definite matrices. Then Ψ satisfies Assumption 3.6.*

We now show that the multivariate Gaussian family has degenerate limits.

Proof of Proposition B.2. If the parameters $(\theta_i)_{i\in\mathbb{N}}$ are not a relatively compact subset of Θ, then either some coordinate of the sequence of means ν_i diverges, $\lambda_1(\Sigma_i) \to \infty$, or $\lambda_d(\Sigma_i) \to 0$. If some coordinate of the mean ν_i diverges or the maximum eigenvalue diverges, i.e., $\lambda_1(\Sigma_i) \to \infty$, then the sequence (ψ_{θ_i}) is not tight by Lemma B.1. On the other hand, if $\lambda_d(\Sigma_i) \to 0$ as $i \to \infty$, then ψ_{θ_i} converges weakly to a sequence of degenerate Gaussian measures that concentrate on $C_i = \{x \in \mathbb{R}^d : (x - \nu_i)^\top u_{d,i} = 0\}$, where $u_{d,i}$ is the d^{th} eigenvector of Σ_i. Note that $\mu(C_i) = 0$ for Lebesgue measure μ; so if we define $C = \cup_i C_i$ in the setting of Definition 3.4, the sequence is not μ-wide.

□

We can generalize Proposition B.2 beyond multivariate Gaussians to mixture-identifiable location-scale families, as shown in Proposition B.3. Examples of such families include the multivariate Gaussian family, the Cauchy family, the logistic family, the von Mises family, and generalized extreme value families. The proof is similar to that of Proposition B.2.

Proposition B.3. *Suppose Ψ is a location-scale family that is mixture-identifiable and absolutely continuous with respect to Lebesgue measure μ, i.e.,*

$$\frac{d\Psi}{d\mu} = \left\{ |\Sigma|^{-1/2} \varphi\left(\Sigma^{-1/2}(x - \nu)\right) : \nu \in \mathbb{R}^d, \Sigma \in \mathbb{S}_{++}^d \right\},$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a probability density function. Then Ψ satisfies Assumption 3.6.

C Additional experiments

C.1 Gaussian mixture data

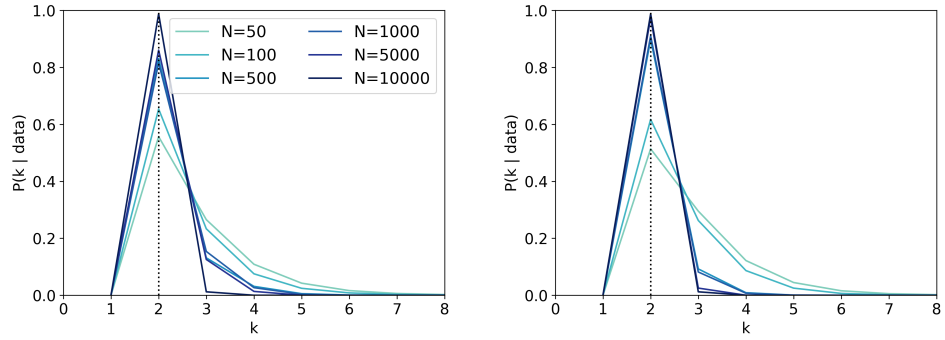
We generated data sets of increasing size $N \in \{50, 100, 500, 1000, 5000, 10000\}$ from 2-component univariate Gaussian mixture models with means $(-5, 5)$, scales $(1.5, 1)$, and mixing weights $(0.4, 0.6)$. In Figure 3, we see that the posterior number of components concentrates on the generating number of components, and for $\alpha = 0.2$, the posterior is less concentrated when there is less data.

C.2 Results for $\alpha_N \to \alpha$

Using the same data of Section 5, we examined the behavior of the α_N-posterior number of components. Below, we consider sequences $\alpha_N = (1 - 1/N)\alpha$, where $\alpha \in \{1, 0.8, 0.5, 0.2\}$. We plot the α_N-posterior number of components for the Laplace mixture data in Figure 4 and the galaxy data in Figure 5. As in the fixed α case, we observe that the α_N posterior number of components diverges.

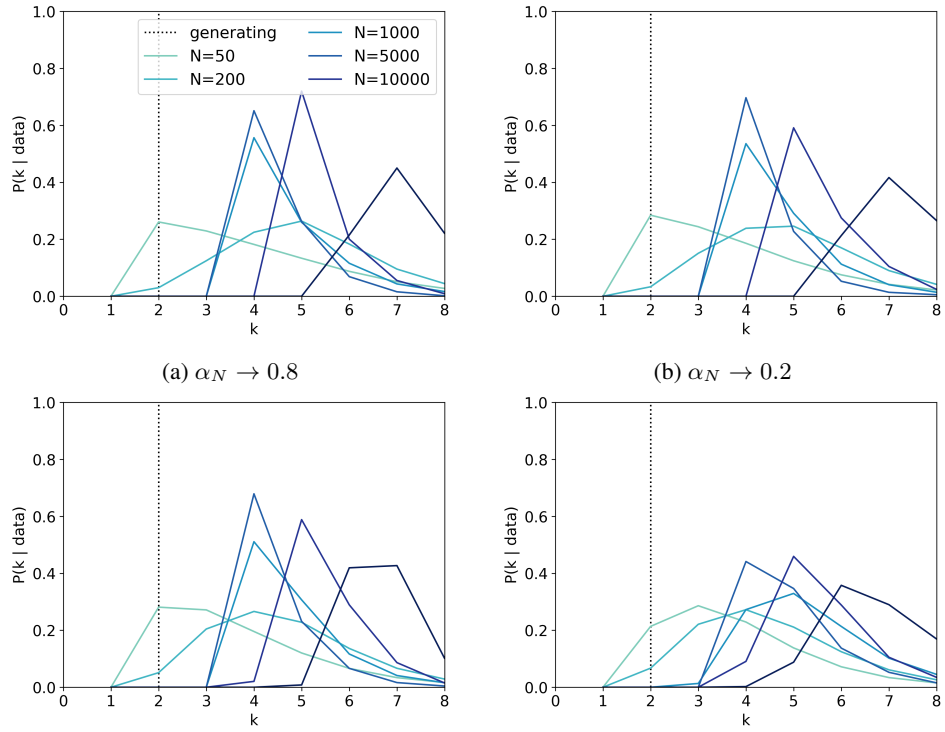
C.3 Gene expression data

Computational biologists are interested in classifying cell types by applying clustering techniques to gene expression data (Yeung et al., 2001; Medvedovic and Sivaganesan, 2002; McLachlan et al., 2002; Medvedovic et al., 2004; Rasmussen et al., 2008; de Souto et al., 2008; McNicholas and Murphy, 2010).



(a) $\alpha = 1$ (b) $\alpha = 0.2$

Figure 3: Gaussian mixture data. Results for fixed α-posterior number of components. The vertical black dotted line denotes the generating number of components.



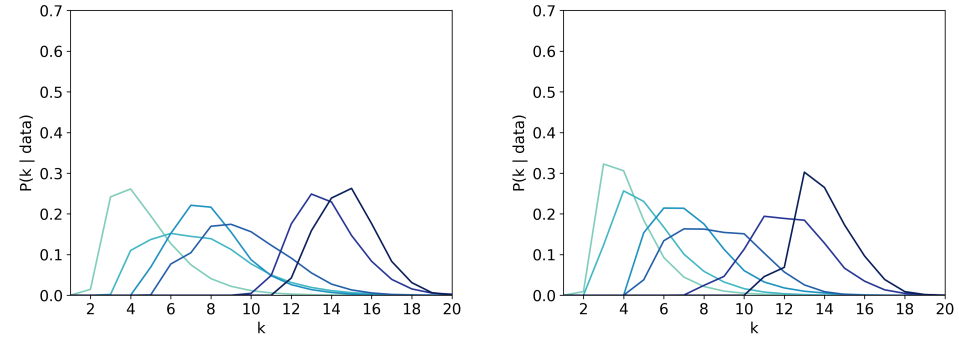
(a) $\alpha_N \to 0.8$ (b) $\alpha_N \to 0.2$

(c) $\alpha_N \to 0.8$ (d) $\alpha_N \to 0.2$

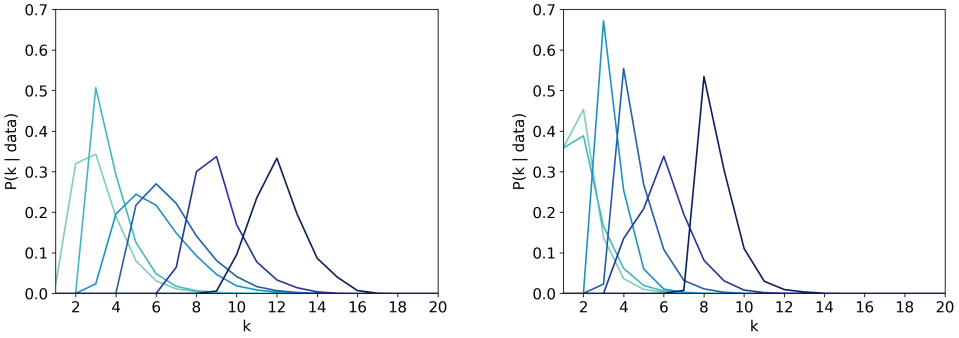
Figure 4: Laplace mixture data. Results for α_N-posterior number of components, where $\alpha_N \to \alpha \in (0, 1)$.

In this section, we examine the behavior of the usual posterior ($\alpha = 1$) on gene expression data using a mixture of multivariate Gaussians with axis-aligned covariances. In particular, consider number of components k, mixture weights $p \in \mathbb{R}^k$, Gaussian component precisions $\tau \in \mathbb{R}_+^{k \times D}$ and means $\theta \in \mathbb{R}^{k \times D}$, labels $Z \in \{1, \ldots, k\}^N$, and data $X \in \mathbb{R}^{N \times D}$. Then the probabilistic generative model is

$$
\begin{aligned}
k &\sim \text{Geom}(r) & w &\sim \text{Dirichlet}_k(\gamma, \ldots, \gamma) \\
\tau_{jd} &\overset{\text{i.i.d.}}{\sim} \text{Gam}(\alpha, \beta) & \theta_{jd} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(m, \kappa_{jd}^{-1}) \\
Z_n &\overset{\text{i.i.d.}}{\sim} \text{Categorical}(w) & X_{nd} &\overset{\text{ind}}{\sim} \mathcal{N}(\theta_{z_n d}, \tau_{z_n d}^{-1}),
\end{aligned}
$$

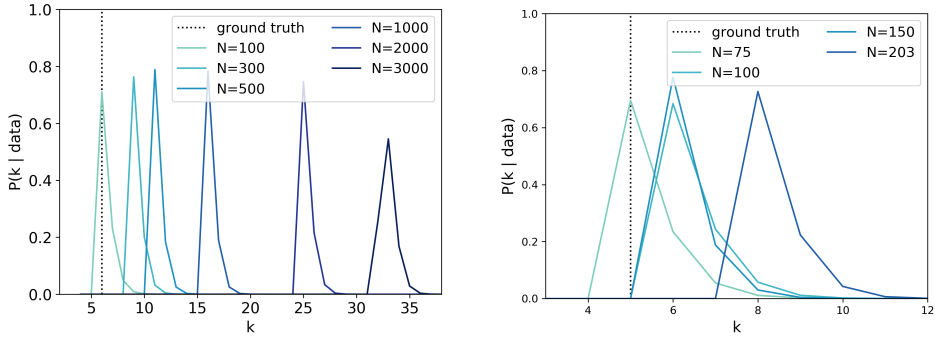


(a) $\alpha \to 1$ (b) $\alpha \to 0.8$



(c) $\alpha \to 0.5$ (d) $\alpha \to 0.2$

Figure 5: Shapley galaxy data. Results for α_N-posterior number of components, where $\alpha_N \to \alpha \in (0, 1)$.



(a) Mouse single-cell data (b) Lung gene expression data

Figure 6: Posterior probability of the number of components k for Gaussian mixture models, fit to (a) mouse cortex single-cell RNA sequencing data and (b) lung tissue gene expression data.

where j ranges from $1, \ldots, k$, d ranges from $1, \ldots, D$, and n ranges from $1, \ldots, N$.

In our next set of experiments, we apply the Gaussian finite mixture model to two gene expression data sets: (1) single-cell RNA sequencing data from mouse cortex and hippocampus cells (Zeisel et al., 2015) with the same feature selection as Prabhakaran et al. (2016) ($N = 3008$, $D = 558$, 11,000 Gibbs sampling steps with 1,000 of those as burn-in) and (2) mRNA expression data from human lung tissue (Bhattacharjee et al., 2001) ($N = 203$, $D = 1543$, and 10,000 Gibbs sampling steps with 1,000 of those burn-in). Our experiments here represent a simplified version of previous mixture model analyses for these and other related data sets (de Souto et al., 2008; Prabhakaran et al., 2016; Armstrong et al., 2001; Miller and Harrison, 2018). For both data sets, we used hyperparameters $\alpha = 1$, $\beta = 1$, $m = 0$, $\kappa_{jd} = \tau_{jd}$, $r = 0.1$, and $\gamma = 1$.

As these gene expression data sets contain counts, we first transformed the data to real numerical values. In particular, we used a base-2 log transform followed by standardization—such that each dimension of the data had zero mean and unit variance—per standard practices (e.g., Miller and Harrison (2018)). Then to examine the effect of increasing data set size on inferential results, we randomly sampled subsets of increasing size without replacement; each smaller subset was contained in the next larger data set.

For the single-cell RNAseq data set, the posterior on the number of components is shown in Figure 6a. Here the ground truth number of clusters is captured when the data set size is $N = 100$. But as predicted by our theory, as we increase the number of data points, the posterior number of components diverges.

The α-posterior on the number of components for the lung gene expression data is shown in Figure 6b. Again we find that on the smallest data subsets, the posterior appears to capture the ground truth number of clusters, but that as we examine more and more data, the posterior diverges. While diagonal covariance Gaussian components are likely not rich enough to model the cluster shapes, our purpose here is to capture the effect of model misspecification on the posterior on the number of components. Thus, these examples suggest the need for more robust analyses.

References

S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. d. Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41, 2001.

J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49 (3):803–821, 1993.

A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. Mark, E. Lander, W. Wong, B. Johnson, T. Golub, D. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24):13790–13795, 2001.

P. Billingsley. *Probability and Measure*. John Wiley and Sons, third edition, 1986.

C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, and T. B. Kepler. Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A*, 73(8):693–701, 2008.

J. Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1): 221–233, 1995.

M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, 9(1):497, 2008.

M. J. Drinkwater, Q. A. Parker, D. Proust, E. Slezak, and H. Quintana. The large scale distribution of galaxies in the Shapley supercluster. *Publications of the Astronomical Society of Australia*, 21(1): 89–96, 2004.

S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Series in Statistics, 2006.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017.

S. Ghosal, J. Ghosh, and R. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

J. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics, 2003.

S. Ghosh and E. B. Sudderth. Nonparametric learning for layered segmentation of natural images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2279. IEEE, 2012.

C. Grazian, C. Villa, and B. Liseo. On a loss-based prior for the number of components in mixture models. *Statistics & Probability Letters*, 158:108656, 2020.

P. J. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375, 2001.

P. Grünwald and T. v. Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

P. D. Grünwald. Bayesian inconsistency under misspecification. In *World Meeting of the International Society for Bayesian Analysis*, 2006.

A. Guha, N. Ho, and X. Nguyen. On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *arXiv preprint arXiv:1901.05078*, 2019.

P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018.

N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271–307, 2016.

C. Holmes and S. Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

J. P. Huelsenbeck and P. Andolfatto. Inference of population structure under a Dirichlet process model. *Genetics*, 175(4):1787–1802, 2007.

H. Ishwaran, L. F. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96(456):1316–1332, 2001.

E. D. Lorenzen, P. Arctander, and H. R. Siegismund. Regional genetic structuring and evolutionary history of the impala aepyceros melampus. *Journal of Heredity*, 97(2):119–132, 2006.

G. J. McLachlan, R. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3):413–422, 2002.

P. D. McNicholas and T. B. Murphy. Model-based clustering of microarray expression data via latent gaussian mixture models. *Bioinformatics*, 26(21):2705–2712, 2010.

M. Medvedovic and S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.

M. Medvedovic, K. Y. Yeung, and R. E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.

J. W. Miller and D. B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

J. W. Miller and M. T. Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pages 199–206, 2013.

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

J. W. Miller and M. T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

S. Mukherjee, E. D. Feigelson, G. J. Babu, F. Murtagh, C. Fraley, and A. Raftery. Three types of gamma-ray bursts. *The Astrophysical Journal*, 508(1):314, 1998.

J. Nemec and A. F. L. Nemec. Mixture models for studying stellar populations. I. Univariate mixture models, parameter estimation, and the number of discrete population components. *Publications of the Astronomical Society of the Pacific*, 103(659):95, 1991.

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

A. Nobile. *Bayesian analysis of finite mixture distributions*. PhD thesis, Carnegie Mellon University, 1994.

A. Nobile. On the posterior distribution of the number of components in a finite mixture. *Annals of Statistics*, 32(5):2044–2073, 2004.

A. Nobile. Bayesian finite mixtures: a note on prior specification and posterior computation. *arXiv preprint arXiv:0711.0458*, 2007.

A. Nobile and A. T. Fearnside. Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Statistics and Computing*, 17(2):147–162, 2007.

E. Otranto and G. M. Gallo. A nonparametric Bayesian approach to detect the number of regimes in Markov switching models. *Econometric Reviews*, 21(4):477–496, 2002.

F. Petralia, V. Rao, and D. B. Dunson. Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2012.

S. Prabhakaran, E. Azizi, A. Carr, and D. Pe'er. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, pages 1070–1079, 2016.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

C. Rasmussen, B. de la Cruz, Z. Ghahramani, and D. Wild. Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):615–628, 2008.

S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997.

K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902, 1997.

R. Royall and T.-S. Tsou. Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):391–404, 2003.

W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

M. Stephens. Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74, 2000.

H. Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248, 1961.

H. Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 1265–1269, 1963.

S. T. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, pages 90–110, 2006.

S. Walker and N. L. Hjort. On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821, 2001.

Y. Wu and S. Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331, 2008.

E. P. Xing, K.-A. Sohn, M. I. Jordan, and Y.-W. Teh. Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. In *International Conference on Machine Learning*, pages 1049–1056, 2006.

S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001.

A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015.