Supplementary Material: Finite mixture models do not reliably learn the number of components

Diana Cai [*1] **Trevor Campbell** [*2] **Tamara Broderick** [3]

A. Background on posterior contraction

Weak topology and weak convergence. In this section, we review some definitions and results used in our work. Our treatment follows Ghosal and van der Vaart (2017, Appendix A), and we refer to this chapter for additional details on the topology of weak convergence.

Let $\mathbb{X}$ be a Polish space metrized by ρ. Below we define the Lévy-Prokhorov metric, which induces the weak topology on $\mathscr{P}(\mathbb{X})$.

Definition A.1. *Let $f, g \in \mathscr{P}(\mathbb{X})$. The Lévy-Prokhorov metric is defined as*

$$d(f,g) = \inf\{\epsilon > 0 : f(A) < g(A^\epsilon) + \epsilon, g(A) < f(A^\epsilon) + \epsilon\},$$

where $A^\epsilon := \{y : \rho(x,y) < \epsilon \text{ for some } x \in A\}$.

The Portmanteau theorem characterizes equivalent notions of weak convergence, and below we include the relevant portions of the Portmanteau theorem used in our proof. For a full statement of the theorem, see Ghosal and van der Vaart (2017, Theorem A.2).

Theorem A.2 (Portmanteau (partial statement))**.** *The following statements are equivalent for any $f_i, f \in \mathscr{P}(\mathbb{X})$:*

1. *$f_i \Rightarrow f$;*
2. *for all bounded, uniformly continuous $h : \mathbb{X} \to \mathbb{R}$,*
$$\int h df_i \longrightarrow \int h df;$$
3. *for every closed subset C, $\limsup_i f_i(C) \leq f(C)$.*

Prokhorov's theorem (Ghosal and van der Vaart, 2017, Theorem A.4) characterizes (weakly) compact subsets of $\mathscr{P}(\mathbb{X})$ in terms of a *tight* subset of measures. A subset $\Gamma \subseteq \mathscr{P}(\mathbb{X})$ is *tight* if for any $\epsilon > 0$, there exists a compact subset $K_\epsilon \subseteq \mathbb{X}$ such that for every $\psi \in \Gamma$, $\psi(K_\epsilon) \geq 1 - \epsilon$.

Theorem A.3 (Prokhorov)**.** *If $\mathbb{X}$ is a Polish space, then $\Gamma \subseteq \mathscr{P}(\mathbb{X})$ is relatively compact if and only if Γ is tight.*

Schwartz's theorem for weak consistency. Below, we state a result for posterior consistency with respect to the weak topology due to Schwartz (1965) (see also Ghosh and Ramamoorthi (2003, Theorem 4.4.2)). The result is a posterior consistency theorem for the density, and thus relies on the assumption that the space of models $\mathbb{F}$ is dominated by a σ-finite measure μ.

[*]Equal contribution [1]Department of Computer Science, Princeton University [2]Department of Statistics, University of British Columbia [3]CSAIL, Massachusetts Institute of Technology. Correspondence to: Diana Cai <dcai@cs.princeton.edu>, Trevor Campbell <trevor@stat.ubc.ca>, Tamara Broderick <tbroderick@csail.mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Theorem A.4 (Schwartz)**.** *Let* Π *be a prior on* $\mathbb{F}$ *and suppose* f_0 *is in the KL support of the prior* Π*. Then the posterior is* weakly consistent *at* f_0*: i.e., for any weak neighborhood* U *of* f_0 *the sequence of posterior distributions satisfies*

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

The above result assumes that the prior Π is fixed. Note that weak consistency also holds (with f_0-probability) for priors that vary with N, provided that the sequence Π_N satisfies the KL support condition stated in Assumption 5.1 (Ghosal and van der Vaart, 2017, Theorem 6.25).

B. Finite mixture models with an upper bound on the number of components

In this section we consider a modification of the setting from the main paper in which the prior Π has support on only those finite mixtures with at most $\tilde{k}$ components. We start by stating and proving our main result in this finite-support case. Then we discuss why our conditions have changed slightly from Theorem 2.1. Finally we demonstrate our finite-support theory in practice with an experiment.

B.1. Result and proof

Let $\mathbb{F}(k)$ be the set of finite mixtures with exactly k components for $k \leq \tilde{k}$. We can apply the same proof technique in Section 4 to the present case, provided that the mixture-density posterior concentrates on weak neighborhoods of some compact subset of $\tilde{k}$-mixtures.

Theorem B.1. *Suppose that the prior* Π *has support on only those mixtures with at most* $\tilde{k}$ *components. Assume that:*

1. *The posterior concentrates on weak neighborhoods of a weak-compact subset of* $\mathbb{F}(\tilde{k})$*, and*
2. Ψ *is continuous, is mixture-identifiable, and has degenerate limits.*

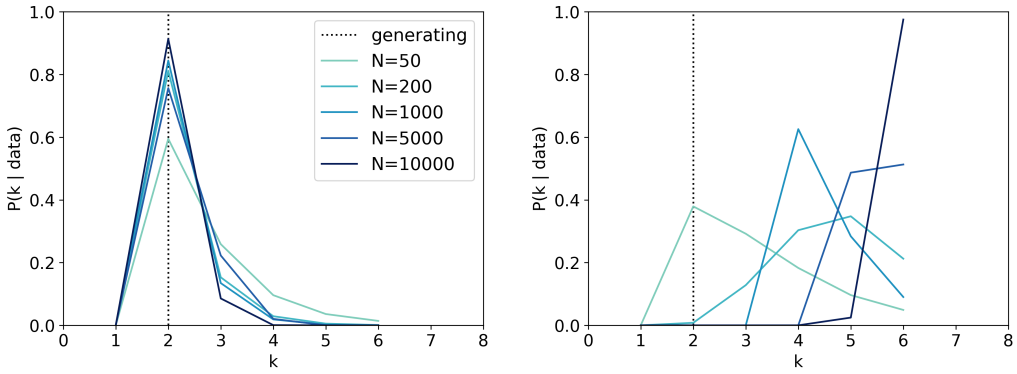
Then the posterior on the number of components concentrates on $\tilde{k}$*:*

$$\Pi(\tilde{k} \mid X_{1:N}) \overset{N\to\infty}{\longrightarrow} 1 \quad f_0\text{-a.s.} \tag{2}$$

Proof Sketch. By assumption, the posterior concentrates on weak neighborhoods of some weak-compact subset $\mathcal{A} \subseteq \mathbb{F}(\tilde{k})$. It remains to show that there exists a weak neighborhood U of $\mathcal{A}$ that, for all $k < \tilde{k}$, contains no k-mixtures of the family of Ψ. Suppose the contrary, i.e., that every such neighborhood contains a mixture of strictly less than $\tilde{k}$ components; then we can construct a sequence $(f_i)_{i=1}^{\infty}$ of mixtures of strictly less than $\tilde{k}$ components such that $f_i \Rightarrow \mathcal{A}$ (in the sense that the infimum of the weak metric between f_i and elements of $\mathcal{A}$ converges to 0). Let g_i be the corresponding sequence of mixing measures such that $f_i = F(g_i)$. Now we follow step 2 of the proof of the main theorem, with some slight modifications to account for the fact that f_i converges weakly to a set rather than a single density. Suppose that $g_i(\Theta \setminus K) \to 0$ for some compact subset $K \subseteq \Theta$. Then following the proof of the main theorem, we have that $\mathbb{F}_K$ and $\mathbb{G}_K$ are compact, and so there is a weak-convergent subsequence of $F(\hat{g}_{i,K})$ that converges to some f_0; since $\mathcal{A}$ is weak-closed, $f_0 \in \mathcal{A}$. The remainder of this branch of the proof then follows the proof main theorem directly. Now for the other branch, suppose $g_i(\Theta \setminus K) \not\to 0$ for any compact $K \subseteq \Theta$. Then as in the main proof there is a sequence of parameters that is not relatively compact; so the corresponding sequence of components ψ_i is either not tight or not μ-wide. Since $\mathcal{A}$ is weak-compact, by Prokhorov's theorem $\mathcal{A}$ is tight, so f_i must be tight, so ψ_i must be tight. On the other hand, ψ_i also must be μ-wide, since otherwise replacing it with the singular sequence ϕ_i shows that f_i would not converge weakly to $\mathcal{A}$. This concludes the second branch of the proof, and the result follows. □

B.2. Discussion of the weak concentration condition

Our main result in Theorem 2.1 uses a KL support condition to guarantee weak concentration of the posterior. In contrast, in Theorem B.1, we do not impose any KL support condition and instead just directly assume weak posterior concentration for the mixture density. First we discuss why this assumption remains reasonable and then discuss why we chose to change the condition.



(a) Gaussian mixture data

(b) Laplace mixture data

Figure B1. Well-specified and misspecified component families that use a prior with an upper bound on the number of components given by $k \sim \text{Unif}\{1, \ldots, 6\}$. Posterior values for component counts k with $k > 6$ are all zero, so we do not plot them.

Reasonableness of the condition. Note that the new weak-concentration assumption is actually weaker than the KL condition in the main paper—albeit potentially substantially more difficult to verify. As a simple example of why this assumption is reasonable, suppose we obtain data generated from a Laplace distribution, and we use a mixture model with Gaussian components and a prior that asserts that the mixture has at most 10 components. Then we expect the posterior to concentrate on mixture densities that have exactly 10 components, and in particular, the set of KL-closest mixture densities to the Laplace. Although many examples will have a single closest such density, we state Theorem B.1 in such a way that it allows for the case where the posterior concentrates on a *compact set* of densities (usually due to symmetry in the model).

Why change the condition. In the main text, we assume—via the KL support condition, Assumption 3.1—that the infimum of the KL divergence from the data generating distribution f_0 to mixture distributions from the model is 0. In other words, we must be able to approximate f_0 arbitrarily well using mixture distributions from the model. However, in the setting with a bounded number of components, this assumption typically does not hold. In particular, the infimum KL from the data-generating distribution f_0 to mixture distributions in the model is nonzero. For example, in the previous Laplace versus Gaussian mixture example, we require an unbounded number of components to achieve a vanishing KL divergence. If we are limited to 10 components, the infimum KL is nonzero.

Demonstrating weak consistency with a reasonable amount of generality when the KL support condition does not hold is challenging; see for instance, Kleijn (2003, Lemma 2.8) and Ramamoorthi et al. (2015, Remark 4). Thus, we opt to require that weak concentration be verified directly for each particular applied setting of interest, rather than attempting to develop a general set of sufficient conditions. The fact that we directly require weak convergence also means that we do not need to make any stipulations about how data are generated. Therefore, in contrast to the main theorem, we do not impose any such assumptions.

B.3. Experiments

Now we demonstrate that the asymptotic behavior described by our theory occurs in practice. In order to the study both the well-specified and misspecified cases, we consider the same 2-component Gaussian and Laplace data described in Section 7.1. Here, we set the prior on the number of components to be a uniform distribution on $\{1, \ldots, 6\}$. The resulting posterior number of components appears in Figure B1. Here the well-specified model (Gaussian data) is consistent and concentrates on the true generating number of components as N grows (Rousseau and Mengersen, 2011). On the other hand, in the misspecified model (Laplace data), the posterior concentrates on the largest possible number of components under the prior, in this case given by $\tilde{k} = 6$.

C. Proof of Proposition 2.2

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 C.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. □

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

Proof of Proposition 2.2. 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 C.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 $C_i = \left\{x \in \mathbb{R}^d : (x - \nu_i)^\top u_{d,i} = 0\right\}$, 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 2.2 beyond multivariate Gaussians to mixture-identifiable location-scale families, as shown in Proposition C.2. 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 2.2.

Proposition C.2. *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.

D. Additional related work

Priors for microclustering behavior have been a recent focus in the Bayesian nonparametrics literature (Zanella et al., 2016; Klami and Jitta, 2016). Since having a fixed number of components across dataset sizes N would be incompatible with sublinear growth (in N) of cluster size across all clusters, we expect divergence issues similar in flavor to those in Miller and Harrison (2013; 2014).

References

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

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

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

A. Klami and A. Jitta. Probabilistic size-constrained microclustering. In *Proceedings of the Conference of Uncertainty in Artificial Intelligence*, 2016.

B. Kleijn. *Bayesian asymptotics under misspecification*. PhD thesis, Vrije Universiteit Amsterdam, 2003.

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.

R. Ramamoorthi, K. Sriram, and R. Martin. On posterior concentration in misspecified models. *Bayesian Analysis*, 10(4): 759–789, 2015.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

L. Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie*, 4:10–26, 1965.

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

G. Zanella, B. Betancourt, H. Wallach, J. Miller, A. Zaidi, and R. C. Steorts. Flexible models for microclustering with application to entity resolution. In *Advances in Neural Information Processing Systems*, 2016.