Research Article

Jeffrey W. Miller*

Consistency of mixture models with a prior on the number of components

https://doi.org/10.1515/demo-2022-0150
received May 6, 2022; accepted October 5, 2022

Abstract: This article establishes general conditions for posterior consistency of Bayesian finite mixture models with a prior on the number of components. That is, we provide sufficient conditions under which the posterior concentrates on neighborhoods of the true parameter values when the data are generated from a finite mixture over the assumed family of component distributions. Specifically, we establish almost sure consistency for the number of components, the mixture weights, and the component parameters, up to a permutation of the component labels. The approach taken here is based on Doob's theorem, which has the advantage of holding under extraordinarily general conditions, and the disadvantage of only guaranteeing consistency at a set of parameter values that has probability one under the prior. However, we show that in fact, for commonly used choices of prior, this yields consistency at Lebesgue-almost all parameter values, which is satisfactory for most practical purposes. We aim to formulate the results in a way that maximizes clarity, generality, and ease of use.

Keywords: asymptotics, Bayesian statistics, clustering, nonparametric inference

MSC 2020: Primary: 62G20, Secondary: 62F15

1 Introduction

Many theoretical advances have been made in establishing posterior consistency and contraction rates for density estimation when using nonparametric mixture models (see [8] and many references therein) or finite mixture models with a prior on the number of components [12,24]. Elegant results have also been provided showing posterior consistency and contraction rates for estimation of the discrete mixing distribution [19] when using either class of models, as well as consistency for the number of components [9].

Meanwhile, it has long been known that Doob's theorem [4] can be used to prove almost sure consistency for the number of components as well as the mixture weights and the component parameters, up to a permutation [20]. Interestingly, in contrast to the modern theory mentioned previously, a Doob-type result can be extraordinarily general, holding under very minimal conditions. Doob's theorem has been criticized for only guaranteeing consistency on a set of probability one under the prior, and thus, a poorly chosen prior can lead to a useless result [22]; however, for many models, this is a straw man argument since a well-chosen prior can lead to a consistency guarantee at Lebesgue-almost all parameter values.

While the result of Nobile [20] was prescient and general, it has some disadvantages. First, Nobile [20] assumes some conditions that are not needed, specifically, (i) that there is a sigma-finite measure μ such that for all v, the component distribution F_v has a density f_v with respect to μ, (ii) that $v \mapsto f_v(x)$ is continuous for all x, and (iii) employing a somewhat complicated algorithm for mapping parameters

* **Corresponding author: Jeffrey W. Miller,** Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA, United States, e-mail: jwmiller@hsph.harvard.edu

into an identifiable space. Furthermore, it is difficult to use Nobile [20] as a reference since the exposition is quite technical and requires significant effort to unpack.

In this article, we present a Doob-type consistency result for mixtures, with the goal of maximizing clarity, generality, and ease of use. Our result generalizes upon the work of Nobile [20] in that we do not require conditions (i)–(iii). We formulate the result directly in terms of the original parameter space (rather than a transformed space as done by Nobile [20]), reflecting the way these models are used in practice. Furthermore, we provide conditions under which consistency holds almost everywhere with respect to Lebesgue measure, rather than just almost everywhere with respect to the prior as done by Nobile [20].

Compared to the modern theory, the limitation of a Doob-type result is that, for any given true parameter value, the theorem cannot tell us whether it is in the measure zero set where consistency may fail. Another important caveat is that the data are required to be generated from the assumed class of finite mixture models. Most consistency results are based on an assumption of model correctness, and the result we present is no different in that respect. However, unfortunately, the posterior on the number of components in a mixture model is especially sensitive to model misspecification [2,14], so any inferences about the number of components should be viewed with extreme skepticism. On the other hand, Miller and Harrison [15,16] show that popular nonparametric mixture models (such as Dirichlet process mixtures) are not even consistent for the number of components when the component family is correctly specified – and this lack of consistency is an even more fundamental concern than sensitivity to misspecification. Thus, although finite mixture models are rarely – if ever – exactly correct, having a consistency guarantee at least provides an assurance that the methodology is coherent.

In practice, mixture models with a prior on the number of components often provide useful insights into heterogeneous data, and, as the saying goes, "all models are wrong but some are useful" [1]. Mixtures are extensively used in a wide range of applications, and modern algorithms facilitate posterior inference when placing a prior on the number of components; see Miller and Harrison [17] and references therein. Thus, it is important to characterize the theoretical properties of these models as generally as possible.

The article is organized as follows. In Section 2, we describe the class of models under consideration and introduce the conditions to be assumed. In particular, we state conditions on the component distributions (Condition 2.1) and prior (Condition 2.2) enabling Lebesgue-almost everywhere consistency, and we provide common examples satisfying these conditions. In Section 3, we state our main results, and Section 4 contains the proofs.

2 Model

Let $(F_v : v \in \mathcal{V})$ be a family of probability measures on $\mathcal{X}$, where $\mathcal{V} \subseteq \mathbb{R}^D$ is measurable and $\mathcal{X}$ is a Borel measurable subset of a complete separable metric space, equipped with the Borel sigma-algebra. For all d, we give $\mathbb{R}^d$ the Euclidean topology and the resulting Borel sigma-algebra. For $k \in \{1, 2, \ldots\}$, define $\Delta_k \coloneqq \{w \in (0,1)^k : \sum_{i=1}^k w_i = 1\} \subseteq \mathbb{R}^k$. For $w \in \Delta_k$ and $v \in \mathcal{V}^k$, define a probability measure

$$P_{w,v} = \sum_{i=1}^{k} w_i F_{v_i} \tag{1}$$

on $\mathcal{X}$. Thus, $P_{w,v}$ is the mixture with weights w_i and component parameters v_i.

Let π, D_k, and G_k be probability measures on $\{1, 2, \ldots\}$, Δ_k, and $\mathcal{V}^k$, respectively. Consider the following model:

$$\begin{aligned} \text{(number of components)} \quad & K \sim \pi \\ \text{(mixture weights)} \quad & W|K = k \sim D_k \text{ where } W = (W_1, \ldots, W_k) \\ \text{(component parameters)} \quad & V|K = k \sim G_k \text{ where } V = (V_1, \ldots, V_k) \\ \text{(observed data)} \quad & X_1, \ldots, X_n|W, V \sim P_{W,V} \text{ i.i.d.} \end{aligned} \tag{2}$$

We use uppercase letters to denote random variables, such as K, and lowercase to denote particular values, such as k.

2.1 Conditions

Condition 2.1. (Family of component distributions).
(1) For all measurable $A \subseteq \mathcal{X}$, the function $v \mapsto F_v(A)$ is measurable on $\mathcal{V}$.
(2) (Finite mixture identifiability) For all $k, k' \in \{1, 2, \ldots\}$, $w \in \Delta_k$, $w' \in \Delta_{k'}$, $v \in \mathcal{V}^k$, and $v' \in \mathcal{V}^{k'}$, if $P_{w,v} = P_{w',v'}$, then $\sum_{i=1}^{k} w_i \delta_{v_i} = \sum_{i=1}^{k'} w_i' \delta_{v_i'}$.

Here, δ_x denotes the unit point mass at x. Roughly, Condition 2.1(1) is that $(F_v : v \in \mathcal{V})$ is a measurable family and Condition 2.1(2) is that the discrete mixing distribution $\sum_{i=1}^{k} w_i \delta_{v_i}$ is uniquely determined by $P_{w,v}$. Condition 2.1(2) is a standard definition of finite mixture identifiability [25]. Let S_k denote the set of permutations of $\{1, \ldots, k\}$.

Condition 2.2. (Prior). Under the model in equation (2), for all $k \in \{1, 2, \ldots\}$,
(1) $\mathbb{P}(K = k) > 0$,
(2) for all $A \subseteq \Delta_k$ measurable, if $\mathbb{P}(W \in A | K = k) = 0$, then $\{w_{1:k-1} : w \in A\}$ has Lebesgue measure zero,
(3) for all $A \subseteq \mathcal{V}^k$ measurable, if $\sum_{\sigma \in S_k} \mathbb{P}(V_\sigma \in A | K = k) = 0$, then A has Lebesgue measure zero,
(4) $\mathbb{P}(V_i = V_j | K = k) = 0$ for all $1 \leq i < j \leq k$.

Here, $w_{1:k-1} = (w_1, \ldots, w_{k-1})$ and $V_\sigma = (V_{\sigma_1}, \ldots, V_{\sigma_k})$. Roughly, Conditions 2.2(1–3) state that the prior gives positive mass to all k and all sets with nonzero Lebesgue measure, for some permutation of the component labels. Condition 2.2(4) is that the component parameters are distinct with prior probability 1. Note that we do not assume that $W|k$ and $V|k$ have densities with respect to Lebesgue measure.

2.2 Examples

The conditions in Section 2.1 hold for many commonly used mixture models.

2.2.1 Family of component distributions

For the component distributions F_v, there are many commonly used choices that satisfy Condition 2.1(2), including the multivariate normal [26] and, more generally, many elliptical families, such as multivariate t distributions [10]. Several discrete families, such as the Poisson, geometric, negative binomial, and many other power-series distributions, also satisfy Condition 2.1(2) [23]. In each of these cases, Condition 2.1(1) can be easily verified using Folland [7, Theorem 2.37].

2.2.2 Prior

For the prior on the mixture weights $W|k$, Condition 2.2(2) is satisfied by choosing $W|k \sim \text{Dirichlet}(\alpha_{k1}, \ldots, \alpha_{kk})$ for any $\alpha_{k1}, \ldots, \alpha_{kk} > 0$, since this has a density with respect to $(k-1)$-dimensional Lebesgue measure $dw_1 \cdots dw_{k-1}$ and this density is strictly positive on Δ_k. More generally, for the same reason, Condition 2.2(2) is satisfied if $W|k$ is defined as follows: let $Z_i \sim \text{Beta}(a_{ki}, b_{ki})$ independently for $i \in \{1, \ldots, k-1\}$, where $a_{ki}, b_{ki} > 0$, then set $W_i = Z_i \prod_{j=1}^{i-1}(1 - Z_j)$ for $i \in \{1, \ldots, k-1\}$ and $W_k = 1 - \sum_{i=1}^{k-1} W_i$; this is called the generalized Dirichlet distribution [3,11].

For the prior on the component parameters $V|k$, perhaps the most common situation is that $V_1, \dots, V_k$ are i.i.d. from some distribution G_0; in this case, Conditions 2.2(3) and 2.2(4) are satisfied if G_0 has a density with respect to Lebesgue measure and this density is strictly positive on $\mathcal{V}$ except for a set of Lebesgue measure zero. A more interesting example is the case of repulsive mixtures, which use a non-independent prior on component parameters to favor well-separated mixture components. For instance, Petralia et al. [21] propose defining $V|k$ to have a density (with respect to Lebesgue measure) proportional to $h(v)\prod_{i=1}^{k} g_0(v_i)$, where g_0 is a probability density on $\mathcal{V}$ and $h : \mathcal{V}^k \to \mathbb{R}$ is either $h(v) = \prod_{1\le i<j\le k}\rho(\|v_i - v_j\|)$ or $h(v) = \min_{1\le i<j\le k}\rho(\|v_i - v_j\|)$, where $\rho : [0, \infty) \to \mathbb{R}$ is a strictly increasing, bounded function with $\rho(0) = 0$. Then Conditions 2.2(3) and 2.2(4) are satisfied as long as g_0 is strictly positive on $\mathcal{V}$ except for a set of Lebesgue measure zero. This holds not only when v_i consists of location parameters but also, in general, for any form of component parameter, such as both location and scale parameters.

3 Main results

We show that for any model as in equation (2) satisfying Conditions 2.1 and 2.2, the posterior is consistent for k, w, and v up to a permutation of the component labels, except on a set of Lebesgue measure zero. More generally, if only Conditions 2.1 and 2.2(4) are satisfied, then the result holds except on a set of prior measure zero.

Define $\Theta_k \coloneqq \Delta_k \times \mathcal{V}^k$ and $\Theta \coloneqq \bigcup_{k=1}^{\infty}\Theta_k$, noting that $\Theta_1, \Theta_2, \dots$ are disjoint sets. Thus, for any $\theta \in \Theta$, we have $\theta = (w, v)$ for some unique $w \in \Delta_k$, $v \in \mathcal{V}^k$, and $k \in \{1, 2, \dots\}$; let $k(\theta)$ denote this value of k. In terms of θ, the data distribution is $P_\theta = P_{w,v}$, where $P_{w,v}$ is defined in equation (1).

We define a metric on Θ as follows: for $\theta, \theta' \in \Theta$, let

$$d_\Theta(\theta, \theta') = \begin{cases} \min\{\|\theta - \theta'\|, 1\} & \text{if } k(\theta) = k(\theta'), \\ 1 & \text{otherwise,} \end{cases} \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm on $\Delta_k \times \mathcal{V}^k \subseteq \mathbb{R}^{k+kD}$. Propositions A.1 and A.2 show that d_Θ is indeed a metric and Θ is a Borel measurable subset of a complete separable metric space; we give Θ the resulting Borel sigma-algebra. Recall that S_k denotes the set of permutations of $\{1, \dots, k\}$. For $\sigma \in S_k$ and $\theta \in \Theta_k$, let $\theta[\sigma]$ denote the transformation of θ obtained by permuting the component labels, that is, if $\theta = (w, v)$, then $\theta[\sigma] \coloneqq (w_\sigma, v_\sigma)$, where $w_\sigma = (w_{\sigma_1}, \dots, w_{\sigma_k})$ and $v_\sigma = (v_{\sigma_1}, \dots, v_{\sigma_k})$. For $\theta_0 \in \Theta_k$ and $\varepsilon > 0$, define

$$\tilde{B}(\theta_0, \varepsilon) = \bigcup_{\sigma \in S_k} \{\theta \in \Theta : d_\Theta(\theta, \theta_0[\sigma]) < \varepsilon\}. \tag{4}$$

Consider the model in equation (2) and define the random variable $\theta \coloneqq (W, V)$.

Theorem 3.1. *Assume Conditions* 2.1 *and* 2.2(4) *hold. There exists* $\Theta_* \subseteq \Theta$ *such that* $\mathbb{P}(\theta \in \Theta_*) = 1$ *and for all* $\theta_0 \in \Theta_*$, *if* $X_1, X_2, \dots \sim P_{\theta_0}$ *i.i.d., then for all* $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\theta \in \tilde{B}(\theta_0, \varepsilon) | X_1, \dots, X_n) = 1 \quad a.s.\left[P_{\theta_0}\right] \tag{5}$$

and

$$\lim_{n\to\infty} \mathbb{P}(K = k(\theta_0) | X_1, \dots, X_n) = 1 \quad a.s.\left[P_{\theta_0}\right]. \tag{6}$$

Here, the conditional probabilities are under the assumed model in equation (2); note that $\theta|X_1, \dots, X_n$ has a regular conditional distribution by Durrett [6, Theorems 1.4.12 and 4.1.6]. Now, define a measure λ on Θ as follows. Let $\lambda_{\mathcal{V}^k}$ denote Lebesgue measure on $\mathcal{V}^k$, and let λ_{Δ_k} denote the measure on Δ_k such that, for all $A \subseteq \Delta_k$ measurable, $\lambda_{\Delta_k}(A)$ equals the Lebesgue measure of $\{w_{1:k-1} : w \in A\} \subseteq \mathbb{R}^{k-1}$. Define $\lambda(A) \coloneqq \sum_{k=1}^{\infty}(\lambda_{\Delta_k} \times \lambda_{\mathcal{V}^k})(A \cap \Theta_k)$ for all measurable $A \subseteq \Theta$. In essence, λ can be thought of as Lebesgue measure on Θ.

Theorem 3.2. *If Conditions* 2.1 *and* 2.2 *hold, then the set* Θ_* *in* Theorem 3.1 *can be chosen such that* $\lambda(\Theta \setminus \Theta_*) = 0$.

In other words, for λ-almost all values of θ_0 in Θ, if $X_1, X_2, \ldots \sim P_{\theta_0}$ i.i.d., then for all $\varepsilon > 0$, equations (5) and (6) hold P_{θ_0}-almost surely.

4 Proofs

Proof of Theorem 3.1. The basic idea of the proof is to use Doob's theorem on posterior consistency [4,13]. However, Doob's theorem cannot be directly applied since it requires identifiability, and while we assume identifiability of $\sum_{i=1}^{k} w_i \delta_{v_i}$ in Condition 2.1(2), this does not imply identifiability of (w, v) due to (a) invariance of $P_{w,v}$ with respect to permutation of the component labels and (b) the existence of points in Θ where $v_i = v_j$. To handle this, we consider a certain restricted parameter space on which identifiability holds for (w, v), apply Doob's theorem to a collapsed model on this restricted space, and then show that this implies the claimed result on all of Θ.

Identifiability constraints. We constrain the component parameters as follows to obtain identifiability of (w, v). Putting the dictionary order (also known as lexicographic order) on elements of $\mathcal{V} \subseteq \mathbb{R}^D$, define

$$\mathcal{V}_k \coloneqq \{(v_1, \ldots, v_k) \in \mathcal{V}^k : v_1 \prec \cdots \prec v_k\} \subseteq \mathbb{R}^{kD}.$$

Here, $v_i \prec v_j$ denotes that v_i precedes v_j and $v_i \neq v_j$. There is nothing particularly special about using the dictionary order here, aside from being a well-known total order on multivariate spaces and producing order-constrained sets $\mathcal{V}_k$ that are Borel measurable. Define $\tilde{\Theta}_k \coloneqq \Delta_k \times \mathcal{V}_k$ and $\tilde{\Theta} \coloneqq \bigcup_{k=1}^{\infty} \tilde{\Theta}_k$. Then, $\tilde{\Theta}$ is a Borel measurable subset of a complete separable metric space under the metric d_Θ as defined in equation (3); this follows from Propositions A.1 and A.2 by taking $\mathcal{X}_k = \mathbb{R}^{k+kD}$, $d_k(x, y) = \|x - y\|$ for $x, y \in \mathcal{X}_k$, and $A_k = \tilde{\Theta}_k$ for $k \in \{1, 2, \ldots\}$.

Collapsed model. For $\theta \in \Theta_k$, define $T(\theta) = \theta[\sigma]$, where $\sigma \in S_k$ is chosen such that $\theta[\sigma] \in \tilde{\Theta}_k$ if possible, and otherwise $\theta[\sigma] = \theta$. Then, $\mathbb{P}(T(\theta) \in \tilde{\Theta}) = 1$ since the subset of $\mathcal{V}^k$ where two or more v_i's coincide has prior probability zero, by Condition 2.2(4). Denoting $B[\sigma] = \{\theta[\sigma] : \theta \in B\}$, note that by the definition of T, for all $B \subseteq \tilde{\Theta}_k$,

$$T^{-1}(B) = \{\theta \in \Theta : T(\theta) \in B\} = \bigcup_{\sigma \in S_k} B[\sigma]. \tag{7}$$

Letting $\tilde{Q}$ denote the distribution of $T(\theta)$, restricted to $\tilde{\Theta}$, we have

$$\begin{aligned} &T(\theta) \sim \tilde{Q} \\ &X_1, \ldots, X_n | T(\theta) \sim P_{T(\theta)} \quad \text{i.i.d.} \end{aligned} \tag{8}$$

by Dudley [5, Theorem 10.2.1] since $P_\theta = P_{T(\theta)}$ and for all $A \subseteq \mathcal{X}^n$ and $B \subseteq \tilde{\Theta}$ measurable, $\mathbb{P}(X_{1:n} \in A, T(\theta) \in B) = \mathbb{P}(X_{1:n} \in A, \theta \in T^{-1}(B)) = \int_B P_\theta^{(n)}(A) \mathrm{d}\tilde{Q}(\theta)$, where $X_{1:n} = (X_1, \ldots, X_n)$; measurability of $\theta \mapsto P_\theta^{(n)}(A)$ for $A \subseteq \mathcal{X}^n$ follows from measurability of $\theta \mapsto P_\theta(A)$ for $A \subseteq \mathcal{X}$ (shown below at equation (9)) along with Miller [13, Lemma 5.2]. We refer to equation (8) as the collapsed model.

Applying Doob's theorem. We show that the collapsed model in equation (8) satisfies the conditions of Doob's theorem [13]. First, we check identifiability. Let $\theta, \theta' \in \tilde{\Theta}$ such that $P_\theta = P_{\theta'}$. By Condition 2.1(2), $\sum_{i=1}^{k} w_i \delta_{v_i} = \sum_{i=1}^{k'} w'_i \delta_{v'_i}$, where $\theta = (w, v)$, $\theta' = (w', v')$, $k = k(\theta)$, and $k' = k(\theta')$. By the definition of $\tilde{\Theta}$, $v_1, \ldots, v_k$ are all distinct, $v'_1, \ldots, v'_{k'}$ are all distinct, $w_1, \ldots, w_k > 0$, and $w'_1, \ldots, w'_{k'} > 0$. This implies that $k = k'$, $w = w'_\sigma$, and $v = v'_\sigma$ for some $\sigma \in S_k$. Furthermore, because $v_1 \prec \cdots \prec v_k$ and $v'_1 \prec \cdots \prec v'_k$ by the definition of $\tilde{\Theta}$, it must be the case that σ is the identity permutation, so $w = w'$ and $v = v'$, that is, $\theta = \theta'$. Therefore, $\theta = (w, v)$ is identifiable on the restricted space $\tilde{\Theta}$.

Next, we check measurability. Let $A \subseteq \mathcal{X}$ be measurable. Then, for any $k \in \{1, 2, \ldots\}$,

$$\theta \mapsto P_\theta(A) = \sum_{i=1}^{k} w_i F_{v_i}(A) \tag{9}$$

is measurable as a function on $\Theta_k = \Delta_k \times \mathcal{V}^k$, since the projections $(w, v) \mapsto w_i$ and $(w, v) \mapsto v_i$ are measurable, and $v_i \mapsto F_{v_i}(A)$ is measurable on $\mathcal{V}$ by Condition 2.1(1). Therefore, $\theta \mapsto P_\theta(A)$ is measurable as a function on $\tilde{\Theta}_k = \Delta_k \times \mathcal{V}_k \subseteq \Delta_k \times \mathcal{V}^k$. It follows that it is measurable as a function on $\tilde{\Theta}$ (since the preimage of a measurable subset of $\mathbb{R}$ is a union of measurable subsets of $\tilde{\Theta}_1, \tilde{\Theta}_2, \dots$, respectively, and is thus measurable by Proposition A.2).

Thus, by Doob's theorem [13], there exists $\tilde{\Theta}_* \subseteq \tilde{\Theta}$ such that $\mathbb{P}(T(\theta) \in \tilde{\Theta}_*) = 1$ and the collapsed model is consistent at all $T(\theta_0) \in \tilde{\Theta}_*$; that is, for any neighborhood $B \subseteq \tilde{\Theta}$ of $T(\theta_0)$, we have $\mathbb{P}(T(\theta) \in B | X_{1:n}) \to 1$ a.s.$[P_{T(\theta_0)}]$, where $X_{1:n} = (X_1, \dots, X_n)$. Define Θ_* to be the set of all points in Θ that can be obtained by permuting the mixture components of a point in $\tilde{\Theta}_*$, that is, $\Theta_* := \bigcup_{k=1}^{\infty} \bigcup_{\sigma \in S_k} (\tilde{\Theta}_* \cap \tilde{\Theta}_k)[\sigma]$. Then, by equation (7),

$$\mathbb{P}(\theta \in \Theta_*) = \mathbb{P}(T(\theta) \in \tilde{\Theta}_*) = 1.$$

Putting the pieces together. Let $\theta_0 \in \Theta_*$ and define $k_0 = k(\theta_0)$. Let $X_1, X_2, \dots \sim P_{\theta_0}$ i.i.d., let $\varepsilon \in (0, 1)$, and define $B := \{\theta \in \tilde{\Theta} : d_\Theta(\theta, T(\theta_0)) < \varepsilon\} \subseteq \tilde{\Theta}_{k_0}$. Referring to equation (4), observe that $\cup_{\sigma \in S_{k_0}} B[\sigma] \subseteq \tilde{B}(\theta_0, \varepsilon)$. Hence, by equation (7),

$$\mathbb{P}(\theta \in \tilde{B}(\theta_0, \varepsilon) | X_{1:n}) \geq \mathbb{P}\left(\theta \in \cup_{\sigma \in S_{k_0}} B[\sigma] | X_{1:n}\right) = \mathbb{P}(T(\theta) \in B | X_{1:n}) \xrightarrow[n \to \infty]{\text{a.s.}} 1, \tag{10}$$

where $X_{1:n} = (X_1, \dots, X_n)$, since $P_{\theta_0} = P_{T(\theta_0)}$ and the collapsed model is consistent at all $T(\theta_0) \in \tilde{\Theta}_*$. This proves equation (5). Equation (6) follows directly from equation (10), since $\varepsilon < 1$ implies $\tilde{B}(\theta_0, \varepsilon) \subseteq \Theta_{k_0}$, and therefore,

$$\mathbb{P}(K = k_0 | X_{1:n}) = \mathbb{P}\left(\theta \in \Theta_{k_0} | X_{1:n}\right) \geq \mathbb{P}(\theta \in \tilde{B}(\theta_0, \varepsilon) | X_{1:n}) \xrightarrow[n \to \infty]{\text{a.s.}} 1. \qquad \square$$

Proof of Theorem 3.2. Define Θ_* as in the proof of Theorem 3.1. Since $\mathbb{P}(\theta \in \Theta_*) = 1$,

$$0 = \mathbb{P}(\theta \in \Theta \setminus \Theta_*) = \sum_{k=1}^{\infty} \mathbb{P}(\theta \in \Theta_k \setminus \Theta_* | K = k)\, \mathbb{P}(K = k).$$

Since $\mathbb{P}(K = k) > 0$ for all k by Condition 2.2(1), $\mathbb{P}(\theta \in \Theta_k \setminus \Theta_* | K = k) = 0$ for all k.

For $\sigma \in S_k$, let D_k^σ and G_k^σ denote the distributions of $W_\sigma | k$ and $V_\sigma | k$, respectively, under the model. Note that for all $\sigma \in S_k$, $(\Theta_k \setminus \Theta_*)[\sigma] = \Theta_k \setminus \Theta_*$. Thus,

$$(D_k^\sigma \times G_k^\sigma)(\Theta_k \setminus \Theta_*) = (D_k \times G_k)(\Theta_k \setminus \Theta_*) = \mathbb{P}(\theta \in \Theta_k \setminus \Theta_* | K = k) = 0. \tag{11}$$

Note that λ_{Δ_k} is invariant under permutations $\sigma \in S_k$, since by Folland [7, Theorem 2.47], Lebesgue measure $dw_1 \cdots dw_{k-1}$ on $\{w_{1:k-1} \in (0, 1)^{k-1} : \sum_{i=1}^{k-1} w_i < 1\}$ is invariant under transformations of the form $g(w_{1:k-1}) = (w_{\sigma_1}, \dots, w_{\sigma_{k-1}})$, where $w_k = 1 - \sum_{i=1}^{k-1} w_i$, because the Jacobian determinant is ± 1. Conditions 2.2(2) and 2.2(3) are that $\lambda_{\Delta_k} \ll D_k$ and $\lambda_{\mathcal{V}^k} \ll \sum_{\sigma \in S_k} G_k^\sigma$, respectively, where $\ll$ denotes absolute continuity. Thus, by Folland [7, Exercise 3.2.12],

$$\lambda_{\Delta_k} \times \lambda_{\mathcal{V}^k} \ll \lambda_{\Delta_k} \times \sum_{\sigma \in S_k} G_k^\sigma = \sum_{\sigma \in S_k} \lambda_{\Delta_k}^\sigma \times G_k^\sigma \ll \sum_{\sigma \in S_k} D_k^\sigma \times G_k^\sigma. \tag{12}$$

By equation (11), $(D_k^\sigma \times G_k^\sigma)(\Theta_k \setminus \Theta_*) = 0$ for all $\sigma \in S_k$, and thus, $(\lambda_{\Delta_k} \times \lambda_{\mathcal{V}^k})(\Theta_k \setminus \Theta_*) = 0$ by equation (12). Therefore, $\lambda(\Theta \setminus \Theta_*) = \sum_{k=1}^{\infty} (\lambda_{\Delta_k} \times \lambda_{\mathcal{V}^k})(\Theta_k \setminus \Theta_*) = 0$. $\square$

5 Conclusion

There are several directions that could be pursued in future work. First, it is straightforward to generalize to cases in which the true parameter is known to be in a subset of the space and the prior is restricted

accordingly – for instance, if it is known that the number of components is less than some maximal number. A related generalization would be to handle partially identified mixtures, that is, to show consistency in cases where the mixtures are only identifiable up to a certain maximum number of components, such as mixtures of binomials [25, Proposition 4].

Another interesting direction would be to handle overfitted mixtures, in which the true distribution is a finite mixture from the assumed family, but the model uses a fixed number of components that is greater than the true value. An extension to establish consistency for the emission distributions of a hidden Markov model would also be interesting, if possible.

Finally, perhaps the biggest weakness of the Doob-type result is the unknown measure zero set on which consistency may fail. By adding regularity conditions, it might be possible to eliminate this limitation (i.e., to fully characterize the measure zero set) while still retaining broad generality, using a proof technique that augments Doob's theorem.

Acknowledgments: Thanks to Matthew Harrison for helpful comments on an early version of this manuscript.

Funding information: The author states that there is no funding involved.

Author contributions: The author has accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The author states that there is no conflict of interest.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analyzed during this study.

Appendix

A Supporting results

Proposition A.1. *If $\mathcal{X}_1, \mathcal{X}_2, \ldots$ is a sequence of disjoint, complete separable metric spaces with metrics $d_1, d_2, \ldots$, respectively, then $\mathcal{X} = \bigcup_{i=1}^{\infty} \mathcal{X}_i$ is a complete separable metric space under the metric*

$$d(x, y) = \begin{cases} \min\{d_i(x, y), 1\} & \text{if } x, y \in \mathcal{X}_i \text{ for some } i, \\ 1 & \text{if } x \in \mathcal{X}_i, y \in \mathcal{X}_j, \quad \text{and} \quad i \neq j, \end{cases}$$

and the topology induced by this metric coincides with the disjoint union topology.

The disjoint union topology is the smallest topology that contains all the open sets of all the $\mathcal{X}_i$'s. Equivalently, it is the topology consisting of all unions of the form $\bigcup_{i=1}^{\infty} A_i$, where A_i is open in $\mathcal{X}_i$ for $i \in \{1, 2, \ldots\}$.

Proof. First, we show that d is a metric on $\mathcal{X}$. It is easy to see that $d(x, y) = d(y, x)$, $d(x, y) \geq 0$, and $d(x, y) = 0 \Leftrightarrow x = y$. To prove the triangle inequality, let $x, y, z \in \mathcal{X}$ and suppose $x \in \mathcal{X}_i$, $y \in \mathcal{X}_j$, and $z \in \mathcal{X}_k$. Using the fact that $\bar{d}(x, y) \coloneqq \min\{d(x, y), 1\}$ is a metric [18, Theorem 20.1], it is simple to check that $d(x, y) \leq d(x, z) + d(z, y)$ in each of the following cases: (1) $i = j = k$, (2) $i = j \neq k$, and (3) $i \neq j$.

Next, we show that $\mathcal{X}$ is complete under d. Let $x_1, x_2, \ldots \in \mathcal{X}$ be a Cauchy sequence. Choose N such that for all $n, m \geq N$, $d(x_n, x_m) \leq 1/2$. Suppose i is the index such that $x_N \in \mathcal{X}_i$. Then, $x_n \in \mathcal{X}_i$ for all $n \geq N$, and

$d(x_n, x_m) = d_i(x_n, x_m)$ for all $n, m \geq N$. Thus, $(x_N, x_{N+1}, \ldots)$ is a Cauchy sequence in $\mathcal{X}_i$ under d_i, so it converges (under d_i) to some $x \in \mathcal{X}_i$ since $\mathcal{X}_i$ is complete. Hence, it also converges to x under d. Therefore, $\mathcal{X}$ is complete.

Furthermore, $\mathcal{X}$ is separable, since if $C_i \subseteq \mathcal{X}_i$ is a countable dense subset of $\mathcal{X}_i$ under d_i, then it is also dense in $\mathcal{X}_i$ under d, so $\bigcup_{i=1}^{\infty} C_i$ is a countable dense subset of $\mathcal{X}$ under d.

Finally, d induces the disjoint union topology on $\mathcal{X}$, since the collection of open balls

$$\{B_\varepsilon(x) : \varepsilon \in (0, 1), x \in \mathcal{X}_i, i = 1, 2, \ldots\},$$

where $B_\varepsilon(x) = \{y \in \mathcal{X} : d(x, y) < \varepsilon\}$, is a base for both the disjoint union topology and the d-metric topology. □

Proposition A.2. *Suppose $\mathcal{X}_1, \mathcal{X}_2, \ldots$, and $\mathcal{X}$ are defined as in* Proposition A.1. *If $A_1, A_2, \ldots$ are Borel measurable subsets of $\mathcal{X}_1, \mathcal{X}_2, \ldots$, respectively, then $\bigcup_{i=1}^{\infty} A_i$ is a Borel measurable subset of $\mathcal{X}$.*

Proof. For a topological space Y, let $\mathcal{T}_Y$ denote its topology and let $\mathcal{B}_Y = \sigma(\mathcal{T}_Y)$ denote its Borel sigma-algebra. Since $\mathcal{T}_{\mathcal{X}_i} \subseteq \mathcal{T}_{\mathcal{X}}$ (by the definition of the disjoint union topology), then $\mathcal{B}_{\mathcal{X}_i} \subseteq \mathcal{B}_{\mathcal{X}}$, and therefore, $A_i \in \mathcal{B}_{\mathcal{X}_i} \subseteq \mathcal{B}_{\mathcal{X}}$ for all $i = 1, 2, \ldots$. Hence, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}_{\mathcal{X}}$. □

References

[1] Box, G. E. (1979). Robustness in the strategy of scientific model building. In: *Robustness in statistics* (pp. 201–236). Cambridge, MA: Elsevier Inc.

[2] Cai, D., Campbell, T., & Broderick, T. (2021). Finite mixture models do not reliably learn the number of components. In: *International Conference on Machine Learning, PMLR*, (pp. 1158–1169).

[3] Connor, R. J., & Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, *64*(325), 194–206.

[4] Doob, J. L. (1949). Application of the theory of martingales. In: *Actes du Colloque International Le Calcul des Probabilités et ses applications* (*Lyon, 28 Juin – 3 Juillet, 1948*) (pp. 23–27). Paris: CNRS.

[5] Dudley, R. M. (2002). *Real analysis and probability*. Cambridge, UK: Cambridge University Press.

[6] Durrett, R. (1996). *Probability: Theory and examples* (Second edition). Belmont, CA: Wadsworth Publishing Company.

[7] Folland, G. B. (2013). *Real analysis: Modern techniques and their applications*. New York, NY: John Wiley & Sons.

[8] Ghosal, S., & Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge, UK: Cambridge University Press.

[9] Guha, A., Ho, N., & Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, *27*(4), 2159–2188.

[10] Holzmann, H., Munk, A., & Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics*, *33*(4), 753–763.

[11] Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

[12] Kruijer, W., Rousseau, J., & Van Der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, *4*, 1225–1257.

[13] Miller, J. W. (2018). *A detailed treatment of Doob's theorem*. arXiv: http://arXiv.org/abs/arXiv:1801.03122.

[14] Miller, J. W., & Dunson, D. B. (2018). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, *114*, 1113–1125.

[15] Miller, J. W., & Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. *Advances in Neural Information Processing Systems*, *26*.

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

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

[18] Munkres, J. R. (2000). *Topology* (Second edition). Upper Saddle River: Prentice Hall.

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

[20] Nobile, A. (1994). *Bayesian analysis of finite mixture distributions*. (PhD thesis), Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
[21] Petralia, F., Rao, V., & Dunson, D. (2012). Repulsive mixtures. *Advances in Neural Information Processing Systems*, *25*.
[22] Roeder, K., & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, *92*(439), 894–902.
[23] Sapatinas, T. (1995). Identifiability of mixtures of power-series distributions and related characterizations. *Annals of the Institute of Statistical Mathematics*, *47*(3), 447–459.
[24] Shen, W., Tokdar, S. T., & Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, *100*(3), 623–640.
[25] Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *34*, 1265–1269.
[26] Yakowitz, S. J., & Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *39*(1), 209–214.