Posterior consistency for the number of components in a finite mixture

Jeffrey W. Miller
Division of Applied Mathematics
Brown University
Providence, RI 02912
Jeffrey_Miller@brown.edu

Matthew T. Harrison
Division of Applied Mathematics
Brown University
Providence, RI 02912
Matthew_Harrison@brown.edu

Abstract

We will present our recent results proving that Dirichlet process mixtures (DPMs) are not consistent for the number of components in a finite mixture. Further, we will describe a natural alternative to DPMs that is consistent and exhibits many of the attractive properties of DPMs.

1 Introduction

Dirichlet process mixtures (DPMs) [1] are often applied when the data is assumed to come from a mixture with finitely many components, but one does not know the number of components s. In many such cases, one desires to make inferences about s, and it is common practice to use the posterior distribution on the number of components occurring so far. It turns out that this posterior is not consistent for s. That is, we have proven that given unlimited i.i.d. data from a finite mixture with s_0 components, the posterior probability of s_0 does not converge to 1.

Motivated by this finding, we examine an alternative approach to Bayesian nonparametric mixtures, which we refer to as a mixture of finite mixtures (MFM). In addition to being consistent for the number of components, MFMs are very natural and possess many of the attractive features of DPMs, including: efficient approximate inference (with MCMC), consistency for the density (at the optimal rate, under certain conditions), and appealing equivalent formulations (exchangeable distribution on partitions, "restaurant process", stick-breaking, and random discrete measures). Our findings suggest that MFMs may be preferable to DPMs when the data comes from a finite mixture.

2 Posterior consistency

Suppose $\{p_\theta : \theta \in \Theta\}$ is a parametric family, with $\Theta \subset \mathbb{R}^k$. We will be interested in discrete probability measures of the form

$$q = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$$

where $\theta_1, \theta_2, \ldots \in \Theta$, $\pi_1, \pi_2, \ldots \geq 0$ such that $\sum_i \pi_i = 1$, and δ_θ is the unit point mass at $\theta \in \Theta$. Let f_q denote the density of the resulting mixture, that is,

$$f_q(x) = \int_\Theta p_\theta(x)\, dq(\theta) = \sum_{i=1}^{\infty} \pi_i p_{\theta_i}(x).$$

Let $s(q) = |\operatorname{support}(q)| \in \{1, 2, \ldots\} \cup \{\infty\}$, that is, $s(q)$ is the number of components in the mixture f_q. For finite mixtures, assume identifiability in the sense that for any q, q' such that $s(q), s(q') < \infty$, if $f_q = f_{q'}$ then $q = q'$. (Note that we are not constraining the π's to be positive, nor are we constraining the θ's to be distinct and ordered, so the π's and θ's will be non-identifiable. However, we are only assuming identifiability of the measure q, and this is satisfied for

many commonly-used families $\{p_\theta\}$, such as multivariate Gaussian, Gamma, Poisson, Exponential, and Cauchy [9].)

When analyzing the posterior consistency of a Bayesian model, there are two probability distributions under consideration: the data distribution (the "true" distribution) and the model distribution. We will assume:

- Data: $X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_{q_0}$ for some q_0 with $s(q_0) < \infty$.
- Model: $Q \sim$ some prior on discrete measures q, and $X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_Q$ (given Q).

(In a DPM, one chooses $Q \sim \mathrm{DP}(\alpha H)$ in the model.) We will be interested the number of components sampled so far (the number of "occupied tables" in the Chinese restaurant process terminology), and in order to introduce this hidden variable, we can equivalently define the model distribution as follows: $Q \sim$ prior, $\beta_1, \beta_2, \ldots \overset{\text{iid}}{\sim} Q$ (given Q), and $X_i \sim p_{\beta_i}$ (given $Q, \beta_1, \beta_2, \ldots$) independently for $i = 1, 2, \ldots$. This allows us to define

$$T_n = \#\{\beta_1, \ldots, \beta_n\},$$

the number of distinct components used in drawing $X_1, \ldots, X_n$.

For any particular choice of prior on Q, there are several important questions regarding consistency: Is the posterior consistent (and at what rate of convergence)...

(1) ...for the density? i.e. $P_{\text{model}}(\mathrm{dist}(f_Q, f_{q_0}) < \varepsilon \,|\, X_{1:n}) \xrightarrow[n\to\infty]{P_{\text{data}}} 1 \;\; \forall \varepsilon > 0$?
(Also, does this hold at any smooth density, even when it is not a mixture from $\{p_\theta\}$?)

(2) ...for the mixing distribution? i.e. $P_{\text{model}}(\mathrm{dist}(Q, q_0) < \varepsilon \,|\, X_{1:n}) \xrightarrow[n\to\infty]{P_{\text{data}}} 1 \;\; \forall \varepsilon > 0$?

(3) ...for the number of components? i.e. $P_{\text{model}}(T_n = s(q_0) \,|\, X_{1:n}) \xrightarrow[n\to\infty]{P_{\text{data}}} 1$?
(Note: It is customary to use T_n instead of $s(Q)$ since $s(Q) \overset{\text{a.s.}}{=} \infty$ in a DPM.)

(We denote $X_{1:n} = (X_1, \ldots, X_n)$.) It has been shown that DPMs exhibit (1) consistency (in the L^1 metric) for any sufficiently smooth density (at the optimal rate, in a certain sense) [2], and (2) consistency (in the Wasserstein metric) for the mixing distribution (also at the optimal rate) [4]. However, our results answer question (3) in the negative: DPMs are not consistent for the number of components.

3 Inconsistency results

We will present the following results and the intuition behind them. (Note: In (B)-(D) below, by a "standard normal DPM" we mean a DPM using univariate normal components of unit variance, with a standard normal prior on the component means.)

(A) For any DPM using an exponential family for the component distributions with a conjugate prior on component parameters and any fixed value of the concentration parameter, the posterior of T_n is inconsistent for the number of components.

(B) For a standard normal DPM, inconsistency remains when a Gamma prior is put on the concentration parameter.

(C) For a standard normal DPM, inconsistency remains when the prior on T_n is modified to prevent it from diverging (e.g. by choosing $\alpha = c/\log n$ as the concentration parameter).

(D) The posterior can be "badly inconsistent": For a standard normal DPM with any fixed value of the concentration parameter, if the true number of components is 1 then the posterior probability of $T_n = 1$ goes to 0.

To be precise, (A) applies to any regular full-rank exponential family $\{p_\theta : \theta \in \Theta\}$ in natural form, where Θ is the natural parameter space. For instance, this covers: multivariate Gaussian, Gamma, Poisson, Exponential, Geometric, Laplace, and others.

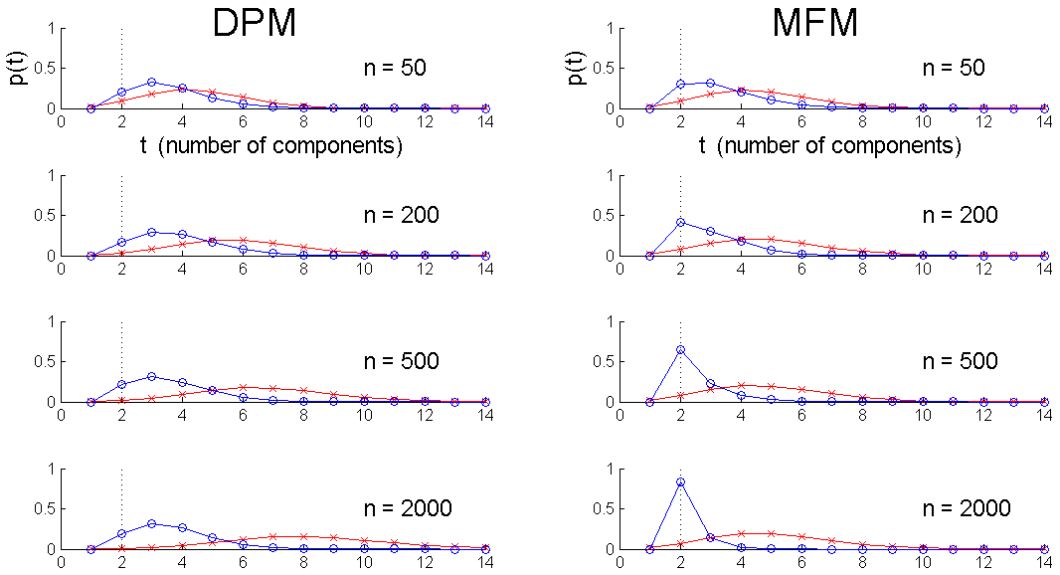

Figure 1: Prior (red x) and estimated posterior (blue o) of T_n, for data from the two-component univariate normal mixture $\frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(6,1)$. Each plot is the average over 5 datasets.

4 A consistent alternative: Mixture of finite mixtures

A natural alternative to a DPM, which we refer to as a "mixture of finite mixtures" (MFM), has been considered by many authors (e.g. [5, 6, 7, 8, 3]). Instead of $Q \sim \mathrm{DP}(\alpha H)$, in a MFM we choose Q as follows:

$$S \sim p(s), \text{ a p.m.f. on } \{1, 2, \dots\}$$

$$\pi \sim \mathrm{Dirichlet}(\alpha_{s1}, \dots, \alpha_{ss}) \text{ (given } S = s)$$

$$\theta_1, \dots, \theta_s \overset{\text{iid}}{\sim} H \text{ (given } S = s)$$

$$Q = \textstyle\sum_{i=1}^{S} \pi_i \delta_{\theta_i}.$$

Then, as in the generic model above, $X_1, X_2, \dots \overset{\text{iid}}{\sim} f_Q$ (given Q). For mathematical convenience, we suggest choosing H to be a conjugate prior for $\{p_\theta\}$, choosing $p(s) = \mathrm{Poisson}(s-1 \mid \lambda)$, and setting $\alpha_{ij} = \alpha > 0$ for all i, j.

MFMs are consistent for the number of components, and furthermore, they share many of the nice properties of DPMs: consistency for the density and the mixing distribution, efficient approximate inference, and interesting equivalent formulations.

Consistency

Under very general conditions, MFMs are consistent for the density, the mixing distribution, and the number of components: using Doob's theorem, consistency can be shown to hold for all true mixing distributions q_0 with finite support, except possibly in a set of Lebesgue measure zero (when Lebesgue measure is extended in a natural way to this space). When using a certain location-scale family with a shared scale parameter, DPMs have been shown [2] to exhibit consistency at any sufficiently smooth density (at the optimal rate) even when it is not a mixture from $\{p_\theta\}$, and a similar result has been proven for MFMs [3].

Exchangeable distribution on partitions

In what follows, to keep things simple we use $\alpha = 1$ in both the MFM and DPM. If $\mathcal{C}$ is a partition of $\{1, \dots, n\}$ into t parts (e.g. $\mathcal{C} = \{\{3,5\}, \{4\}, \{1,2,6\}\}$, $n = 6$, $t = 3$), then

$$P_{\text{MFM}}(\mathcal{C}) = \kappa(n,t) \prod_{c \in \mathcal{C}} |c|! \quad \text{and} \quad P_{\text{DPM}}(\mathcal{C}) = \frac{1}{n!} \prod_{c \in \mathcal{C}} (|c|-1)!$$

where $\kappa(n,t) = \mathbb{E}(S_{(t)}/S^{(n)})$. Here, $s_{(t)} = s(s-1)\cdots(s-t+1)$ is the falling factorial, and $s^{(n)} = s(s+1)\cdots(s+n-1)$ is the rising factorial. The numbers $\kappa(n,t)$ can be efficiently precomputed using

$$\kappa(n,t) = \kappa(n-1,t-1) - (n+t-2)\,\kappa(n,t-1) \quad \text{and} \quad \kappa(n,0) = \mathbb{E}(1/S^{(n)}).$$

In general, $\kappa(n,0)$ can be easily approximated to arbitrary precision. If $p(s) = \text{Poisson}(s-1 \mid \lambda)$, then $\kappa(n,0)$ can be computed analytically using $\kappa(n,0) = P(S > n)/\lambda^n$.

Restaurant process and Gibbs sampling

This leads to a simple "restaurant process" closely resembling the Chinese restaurant process:

> The first customer sits at a table. (At this point, $\mathcal{C} = \{\{1\}\}$.)
>
> The n^{th} customer sits ...
>
> | | MFM | CRP |
> |---|---|---|
> | at table $c \in \mathcal{C}$ with probability $\propto$ | $(\lvert c\rvert + 1)\kappa(n,t)$ | $\lvert c\rvert$ |
> | or at a new table with probability $\propto$ | $\kappa(n,t+1)$ | 1 |
>
> where $t = |\mathcal{C}|$ is the number of occupied tables so far.

Using the recursion for $\kappa(n,t)$, it is easy to show that this process yields the exchangeable distribution on partitions described above. From this restaurant process, one immediately obtains a collapsed Gibbs sampler that is nearly identical to such a sampler for a DPM.

Stick-breaking construction for MFM

The marginal distribution of π is beautifully simple when $p(s) = \text{Poisson}(s-1 \mid \lambda)$ and $\alpha = 1$ (i.e. $S \sim \text{Poisson}(\lambda) + 1$ and $\pi \sim \text{Dirichlet}_s(1, \dots, 1)$ given $S = s$), due to the following fact:

> If $Y_1, Y_2, \dots \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$ and $\pi_k = \min\{Y_k,\ 1 - \sum_{i=1}^{k-1} \pi_i\}$ for $k = 1, 2, \dots$
> then $S := \#\{k : \pi_k > 0\} \sim \text{Poisson}(\lambda) + 1$ and $(\pi_1, \dots, \pi_s) \sim \text{Dirichlet}_s(1, \dots, 1)$ given $S = s$.

In other words, we have the following stick-breaking construction for π: start with a unit-length stick, and break off i.i.d. $\text{Exponential}(\lambda)$ pieces until you run out of stick. (It is interesting to note that this corresponds to a Poisson process on the unit interval.)

Empirical demonstrations

We will present empirical demonstrations, such as in Figure 1, corroborating our theoretical results on consistency of MFMs and inconsistency of DPMs with respect to the number of components.

Open questions

Finally, we will discuss some interesting open questions relating to this work.

References

[1] M. D. Escobar and M. West, *Bayesian density estimation and inference using mixtures*, Journal of the American Statistical Association **90** (1995), no. 430, 577–588.

[2] S. Ghosal and A. W. Van Der Vaart, *Posterior convergence rates of Dirichlet mixtures at smooth densities*, The Annals of Statistics **35** (2007), no. 2, 697–723.

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

[4] X. Nguyen, *Convergence of latent mixing measures in nonparametric and mixture models*, (2012), arXiv:1109.3250.

[5] A. Nobile, *Bayesian analysis of finite mixture distributions*, Ph.D. thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 1994.

[6] A. Nobile and A. T. Fearnside, *Bayesian finite mixtures with an unknown number of components: The allocation sampler*, Statistics and Computing **17** (2007), no. 2, 147–162.

[7] S. Richardson and P. J. Green, *On Bayesian analysis of mixtures with an unknown number of components*, Journal of the Royal Statistical Society. Series B **59** (1997), no. 4, 731–792.

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

[9] S. J. Yakowitz and J. D. Spragins, *On the identifiability of finite mixtures*, The Annals of Mathematical Statistics **39** (1968), no. 1, 209–214.