A simple example of Dirichlet process mixture inconsistency for the number of components

Jeffrey W. Miller and Matthew T. Harrison Brown University, Division of Applied Math

Introduction

Summary

Dirichlet process mixtures (DPMs) are not consistent for the number of components in a finite mixture. However, there is a natural alternative that is consistent and has many of the attractive properties of DPMs.

Overview

For data assumed to come from a finite mixture with an unknown number of components, it has become common to use Dirichlet process mixtures (DPMs) not only for density estimation, but also for inferences about the number of components. The typical approach is to use the posterior distribution on the number of "occupied tables" — that is, the posterior on the number of components represented in the observed data. However, it turns out that this posterior is not consistent — it does not concentrate at the true number of components.

It is known that in many cases the DPM posterior is consistent for the density (Ghosal, 2010) as well as for the mixing distribution (Nguyen, 2013), however, the question of consistency for the number of components has been unanswered until now.

Our general theorem (arXiv:1309.0024) proves this inconsistency for Pitman–Yor process mixtures over a large class of continuous exponential families and essentially all discrete families. This result explains the tiny extra clusters that are often observed in posterior samples (e.g. West, Müller, and Escobar, 1994).

Solution?

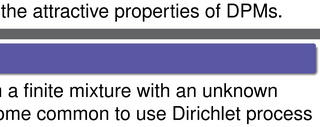
A natural alternative is to simply put a prior on the number of components in a finite mixture model, and this is known to be consistent (Nobile, 1994). It has been believed that inference in such a model is difficult and requires techniques such as reversible jump MCMC (Richardson & Green, 1997). To the contrary, we have found that this approach gives rise to a combinatorial stochastic process that closely parallels that of the DPM, and consequently, efficient approximate inference can be done in much the same way as for the DPM.

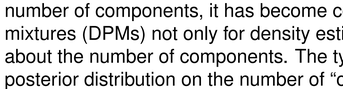
Caution! This remains highly sensitive to misspecification of the component distributions. In general, we urge researchers interested in the number of components to be wary of robustness issues.

Setup

Finite mixture model

$$(\pi_1,\dots,\pi_k) \sim \text{Dirichlet}(\gamma,\dots,\gamma)$$
$$\theta_1,\dots,\theta_k \overset{\text{iid}}{\sim} H$$
$$X_1,\dots,X_n \overset{\text{iid}}{\sim} f(x) = \sum_{i=1}^{k} \pi_i p_{\theta_i}(x)$$

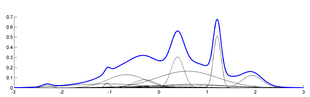




5 tables (i.e. components)
3 occupied tables

Dirichlet process mixture model

$$(\pi_1,\pi_2,\dots) \sim \text{Stick}(1,\alpha)$$
$$\theta_1,\theta_2,\dots \overset{\text{iid}}{\sim} H$$
$$X_1,\dots,X_n \overset{\text{iid}}{\sim} f(x) = \sum_{i=1}^{\infty} \pi_i p_{\theta_i}(x)$$



∞ tables (i.e. components)
4 occupied tables

Theoretical results

Background

It is known that the posterior concentrates at the true density f_0, i.e.

$$P(\|f - f_0\|_{L_1} < \varepsilon \mid X_{1:n}) \xrightarrow[n\to\infty]{} 1 \quad \forall \varepsilon > 0,$$

in many cases, for *any* sufficiently regular f_0 (usually at the minimax optimal rate, up to a logarithmic factor).
(Contributions by: Ghosal, van der Vaart, Scricciolo, Lijoi, Prünster, Walker, James, Tokdar, Dunson, Bhattacharya, Wu, Ghosh, Ramamoorthi, Ishwaran, and others.)

In fact, the posterior on the mixing distribution concentrates (in Wasserstein distance) at the true mixing distribution (Nguyen, 2013).

On data from a finite mixture, does the posterior on the number of occupied tables concentrate at the true number of components?

General inconsistency

Theorem

Under mild regularity conditions, if $X_1, X_2, \dots$ are i.i.d. from a finite mixture with k_0 components, then the DPM posterior on the number of occupied tables T_n satisfies

$$\limsup_{n\to\infty} \; P(T_n = k_0 \mid X_{1:n}) < 1$$

with probability 1.

- The model is assumed to use the same family of component distributions as in the data distribution.
- This implies inconsistency of Dirichlet process mixtures over essentially all discrete families and a large class of continuous exponential families (including multivariate Gaussian).
- We assume the concentration parameter α is fixed.
- This generalizes to Pitman–Yor process mixtures.
- See Miller & Harrison (2013) arXiv:1309.0024 for details.

Severe inconsistency

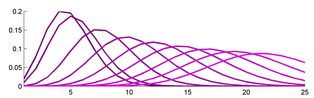
Consider a "standard normal DPM": $p_\theta(x) = \mathcal{N}(x \mid \theta, 1)$ and H is $\mathcal{N}(0,1)$.

Theorem

If $X_1, X_2, \dots \sim \mathcal{N}(0,1)$ i.i.d. then $P(T_n = 1 \mid X_{1:n}) \xrightarrow{\text{Pr}} 0$ as $n \to \infty$, under the standard normal DPM with concentration parameter $\alpha = 1$.

The wrong intuition

It is tempting to think that the prior on T_n is the culprit, since it is diverging as $n \to \infty$.

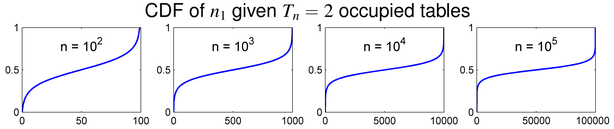


However, this is not the main reason why inconsistency occurs.

The right intuition

Given t occupied tables, the conditional distribution of their sizes $n_1,\dots,n_t$ is $P(n_1,\dots,n_t \mid T_n = t) \;\propto\; n_1^{-1}\cdots n_t^{-1}\, I(\sum n_i = n)$.

CDF of n_1 given $T_n = 2$ occupied tables



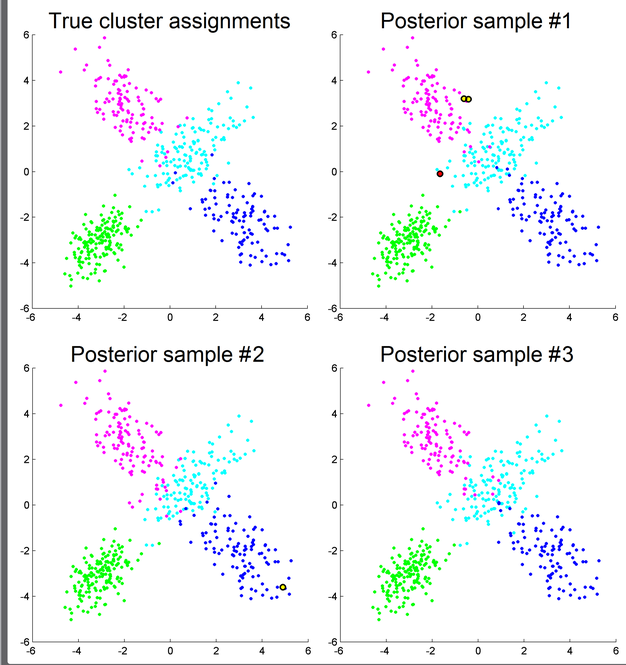
Key observation

As n grows, this becomes concentrated in the "corners". In other words, the DPM really likes to have one or more tables with very few customers.

Empirical evidence

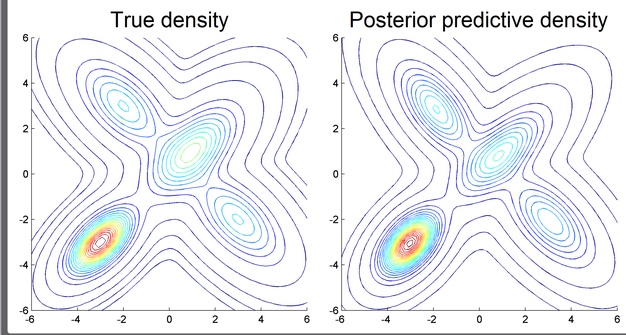
Posterior clustering samples

Tiny extra clusters often appear in posterior samples.



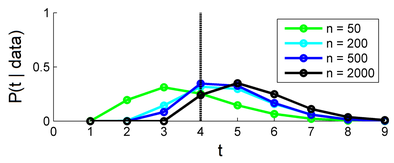
Posterior predictive density

The posterior predictive is accurate, unharmed by such tiny clusters.



Posterior on the number of occupied tables

But the posterior on the number of occupied tables puts significant mass above the true number of components (four, in this case).



Similar behavior occurs for a wide variety of component families.

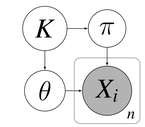
A consistent alternative

A mixture of finite mixtures (MFM)

There is a natural alternative to DPMs that is consistent.
(Nobile (1994, 2007), Richardson & Green (1997, 2001), Stephens (2000), etc.)

MFM model

$$K \sim q(k), \text{ a p.m.f. on } \{1,2,\dots\}$$
$$\pi \sim \text{Dirichlet}(\gamma_{k1},\dots,\gamma_{kk}) \text{ (given } K = k)$$
$$\theta_1,\dots,\theta_k \overset{\text{iid}}{\sim} H \text{ (given } K = k)$$
$$X_1,\dots,X_n \overset{\text{iid}}{\sim} f(x) = \textstyle\sum_{i=1}^{K} \pi_i p_{\theta_i}(x) \text{ (given } K, \pi, \theta).$$



For convenience, we suggest $q(k) = \text{Poisson}(k-1 \mid \lambda)$ and $\gamma_{ij} = \gamma > 0$.

Exchangeable partition probability function (EPPF)

This yields an EPPF of Gibbs form (Gnedin and Pitman, 2005).

EPPF (DPM vs MFM) (. . . with $\alpha = 1$ and $\gamma = 1$ for simplicity)

If $\mathcal{C}$ is a partition of $\{1,\dots,n\}$ into t parts, then

$$P_{\text{DPM}}(\mathcal{C}) = \frac{1}{n!}\prod_{c\in\mathcal{C}}(|c|-1)! \qquad P_{\text{MFM}}(\mathcal{C}) = v_n(t)\prod_{c\in\mathcal{C}}|c|!$$

where $v_n(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{k^{(n)}}\, q(k)$.

- Here, $k_{(t)} = k(k-1)\cdots(k-t+1)$ and $k^{(n)} = k(k+1)\cdots(k+n-1)$.
- The numbers $v_n(t)$ can be efficiently precomputed, numerically.

Restaurant process and Gibbs sampling

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

Restaurant process (DPM vs MFM)

The first customer sits at a table. (At this point, $\mathcal{C} = \{\{1\}\}$.)
The n^{th} customer sits. . .

	DPM	MFM
at table $c \in \mathcal{C}$ with probability $\propto$	$\lvert c\rvert$	$(\lvert c\rvert + 1)\, v_n(t)$
or at a new table with probability $\propto$	1	$v_n(t+1)$

where $t = |\mathcal{C}|$ is the number of occupied tables so far.

Thus, Gibbs sampling for MFMs and DPMs is **nearly identical**.

Empirical results

As n grows, we observe the posterior concentrating at the true number of components, the posterior predictive density converging to the true density, and posterior clustering samples rarely having extra clusters.