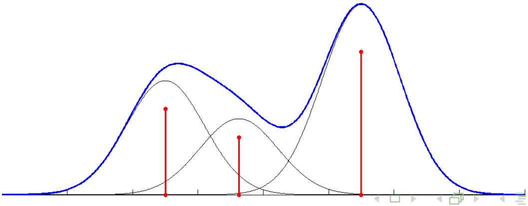
Posterior consistency for the number of components in a finite mixture

Jeffrey W. Miller
and
Matthew T. Harrison

Brown University
Division of Applied Mathematics
Providence, RI

NIPS, December 7, 2012



Summary

1. Dirichlet process mixtures (DPMs) are not consistent for the number of components in a finite mixture.

2. However, there is a natural alternative that is consistent and exhibits many of the attractive properties of DPMs.

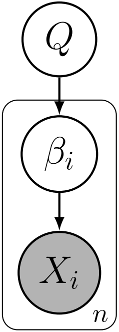
Dirichlet process mixture (DPM) model

Generic DPM model

$Q \sim \mathrm{DP}(\alpha, H)$

$\beta_1, \beta_2, \ldots \overset{\text{iid}}{\sim} Q$ (given Q)

$X_i \sim p_{\beta_i}$ independent for $i = 1, 2, \ldots$ (given $Q, \beta_1, \beta_2, \ldots$)
for some parametric family $\{p_\theta : \theta \in \Theta\}$.



Let $T_n = \#\{\beta_1, \ldots, \beta_n\}$.
That is, T_n is the number of distinct components so far (i.e. the number of clusters).

Mixture of finite mixtures (MFM)

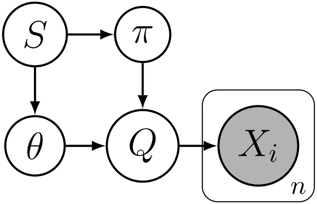
Many authors have considered the following natural alternative to DPMs.

e.g. Nobile (1994, 2000, 2004, 2005, 2007), Richardson & Green (1997, 2001), Stephens (2000), Zhang et al. (2004), Kruijer (2008), Rousseau (2010), Kruijer, Rousseau, & van der Vaart (2010).

Instead of $Q \sim \mathrm{DP}(\alpha, H)$, choose Q as follows:

A mixture over finite mixtures

$S \sim p(s)$, a p.m.f. on $\{1, 2, \dots\}$
$\pi \sim \mathrm{Dirichlet}(\gamma_{s1}, \dots, \gamma_{ss})$ (given $S = s$)
$\theta_1, \dots, \theta_s \overset{\text{iid}}{\sim} H$ (given $S = s$)
$Q = \sum_{i=1}^{S} \pi_i \delta_{\theta_i}$



For mathematical convenience, we suggest:

- H as a conjugate prior for $\{p_\theta\}$
- $p(s) = \mathrm{Poisson}(s - 1 \mid \lambda)$
- $\gamma_{ij} = \gamma > 0$ for all i, j

Questions of convergence

For data from a finite mixture, is the posterior consistent ...

(and at what rate of convergence) ...

	DPMs	MFMs
... for the density?	Yes (optimal rate)	Yes (optimal rate)

DPMs: Ghosal & van der Vaart (2001, 2007), and others.
MFMs: Doob's theorem gives a.e. consistency. Kruijer et al. (2008, 2010) prove rates.

	DPMs	MFMs
... for the mixing distribution?	Yes (optimal rate)	Yes

DPMs: Nguyen (2012)
MFMs: Doob's theorem gives a.e. consistency. Optimal rate?

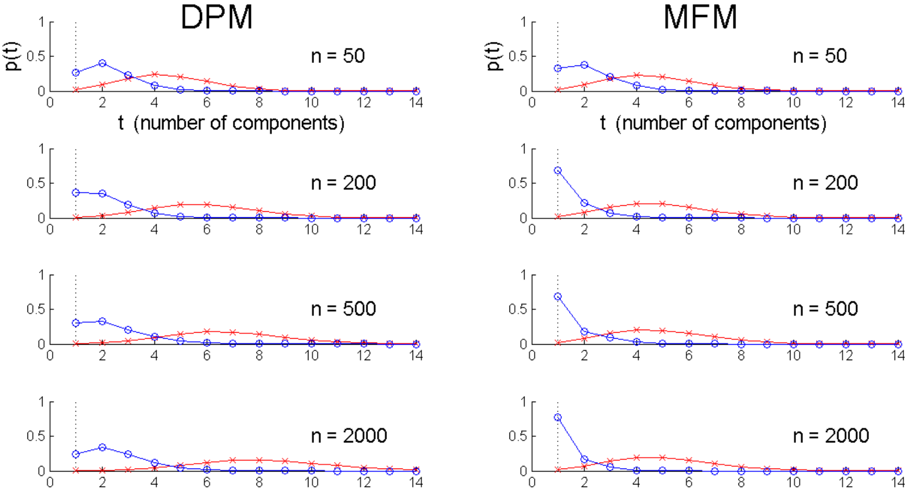
	DPMs	MFMs
... for the number of components?	**Not consistent**	**Yes**

DPMs: This is our contribution.
MFMs: Doob's theorem gives a.e. consistency (see e.g. Nobile (1994)).

(Note: Ignoring tiny clusters might fix this issue.)

Toy example #1: One normal component

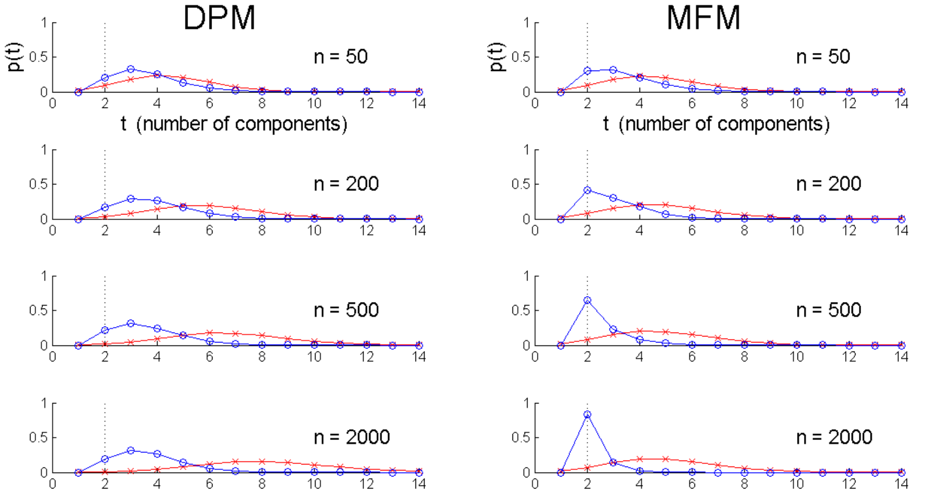
Prior (x) and estimated posterior (o) of the number of clusters



Data: $\mathcal{N}(0,1)$. Each plot is the average over 5 datasets. Burn-in: 10,000 sweeps, Sample: 100,000 sweeps.

Toy example #2: Two normal components

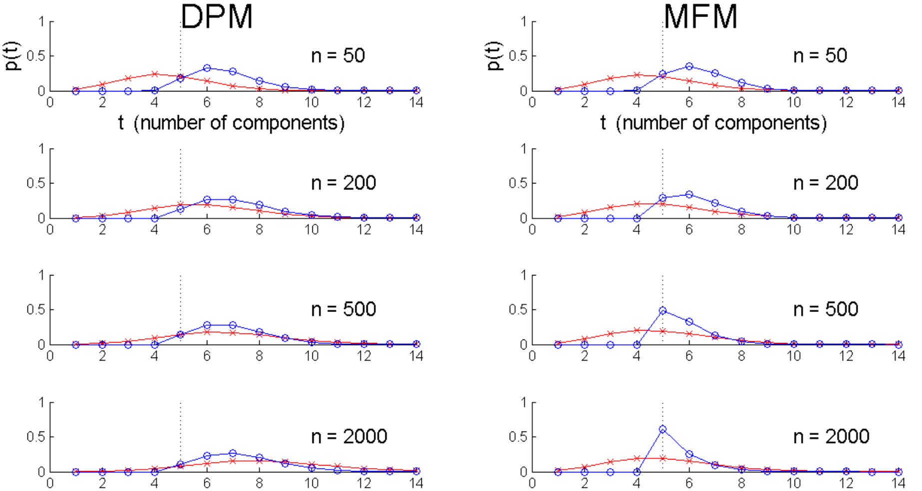
Prior (x) and estimated posterior (o) of the number of clusters



Data: $\frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(6,1)$. Each plot is the average over 5 datasets. Burn-in: 10,000 sweeps, Sample: 100,000 sweeps.

Toy example #3: Five normal components

Prior (x) and estimated posterior (o) of the number of clusters



Data: $\sum_{k=-2}^{2} \frac{1}{5}\mathcal{N}(4k, \frac{1}{2})$. Each plot is the average over 5 datasets. Burn-in: 10,000 sweeps, Sample: 100,000 sweeps.

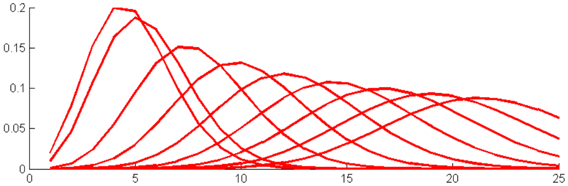
The wrong intuition

It is tempting to think that the prior on the number of clusters T_n is the culprit. After all, when e.g. $\alpha = 1$,

$$P_{\text{DPM}}(T_n = t) = \frac{1}{n!} \begin{bmatrix} n \\ t \end{bmatrix} \sim \frac{1}{n} \frac{(\log n)^{t-1}}{(t-1)!} = \text{Poisson}(t-1 | \log n)$$

where $\left[{n \atop t} \right]$ is an (unsigned) Stirling number of the first kind, and $a_n \sim b_n$ means that $a_n / b_n \to 1$ as $n \to \infty$. Hence, $P_{\text{DPM}}(T_n = t) \to 0$ for any t.

$P_{\text{DPM}}(T_n = t)$ for increasing n



However, this is **not** the fundamental reason why inconsistency occurs. Even if we replace the prior on T_n by something that is not diverging, inconsistency remains!

Comparing DPMs to MFMs

Similarities between DPMs to MFMs:

- Efficient approximate inference (via Gibbs sampling)
- Appealing equivalent formulations:
  - exchangeable distribution on partitions. . . e.g. when $\alpha = 1$ and $\gamma = 1$:

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

  - restaurant process
  - stick-breaking
  - random discrete measures
- Consistent at any sufficiently smooth density (at optimal rate, in a certain sense)

Advantages of MFMs (for data from a finite mixture):

- MFMs are a natural Bayesian extension of finite mixtures
- Consistency (a.e.) for S, π, θ, and the density is automatically guaranteed

Disadvantages of MFMs:

- More parameters (. . . you have to choose $p(s)$)
- (Slightly) more complicated sampling formulas

Thank you!

Jeff Miller
Brown University

jeffrey_miller@brown.edu
www.dam.brown.edu/people/jmiller/