Posterior Consistency for the Number of Components in a Finite Mixture

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

Introduction

Summary

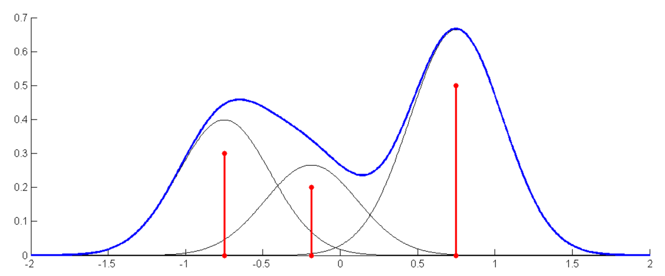
Dirichlet process mixtures (DPMs) are not consistent for the number of components in a finite mixture. However, there is a natural alternative that is consistent and exhibits many of the attractive properties of DPMs.

Setup

- Parametric family: $\{p_\theta : \theta \in \Theta\}$, with $\Theta \subset \mathbb{R}^k$.
- Mixing measure: $q = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$, where $\theta_i \in \Theta$ and $\sum \pi_i = 1$
- Mixture density: $f_q(x) = \sum_{i=1}^{\infty} \pi_i p_{\theta_i}(x)$

  (Note: f_q is a finite mixture when q has finite support.)
- Assume identifiability: $f_q = f_{q'} \Rightarrow q = q'$ for any q, q' with finite support.

For example, $\{p_\theta : \theta \in \Theta\}$ might be univariate normals with $\theta = (\mu, \sigma^2)$:



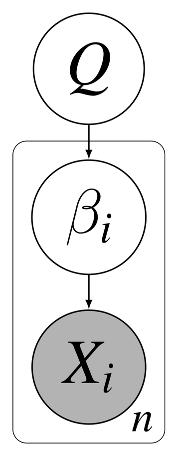
Two distributions

Data distribution (the "true" distribution)

$X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_{q_0}$ for some q_0 with finite support.

Model distribution

$Q \sim$ some prior on mixing measures q, (e.g. $Q \sim \mathrm{DP}(\alpha, H)$)
$\beta_1, \beta_2, \ldots \overset{\text{iid}}{\sim} Q$ (given Q),
$X_i \sim p_{\beta_i}$ (given $Q, \beta_1, \beta_2, \ldots$) independent for $i = 1, 2, \ldots$.



Note that $X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_Q$ (given Q).
Let $T_n = \#\{\beta_1, \ldots, \beta_n\}$ (i.e. the number of distinct components so far).

Questions of convergence

In a DPM, $Q \sim \mathrm{DP}(\alpha, H)$. Alternatively, we could use a MFM (see next panel).
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.		
…for the mixing measure?	Yes (optimal rate)	Yes
DPMs: Nguyen (2012) MFMs: Doob's theorem gives a.e. consistency. Optimal rate?		
…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)).		

A Consistent Alternative

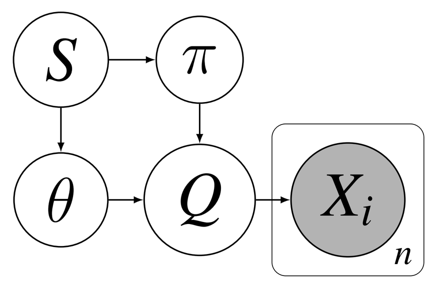
A mixture of finite mixtures (MFM)

Many authors have considered the following natural alternative to DPMs.
(e.g. Nobile (1994, 2007), Richardson & Green (1997, 2001), Stephens (2000), etc.)

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

MFM model

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



For convenience, we suggest $p(s) = \mathrm{Poisson}(s-1 \mid \lambda)$ and $\gamma_{ij} = \gamma > 0\ \forall i, j$.

Exchangeable partition probability function (EPPF)

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

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

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

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

where $\kappa(n, t) = \mathbb{E}(S_{(t)}/S^{(n)})$.

- Here, $s_{(t)} = s(s-1)\cdots(s-t+1)$ and $s^{(n)} = s(s+1)\cdots(s+n-1)$.
- The numbers $\kappa(n, t)$ can be efficiently precomputed.

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 $\propto$	$\lvert c \rvert$	$(\lvert c \rvert + 1)\kappa(n, t)$
or at a new table with probability $\propto$	1	$\kappa(n, t+1)$

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

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

Stick-breaking construction for MFMs

When $\gamma = 1$, the marginal distribution of π is beautifully simple:

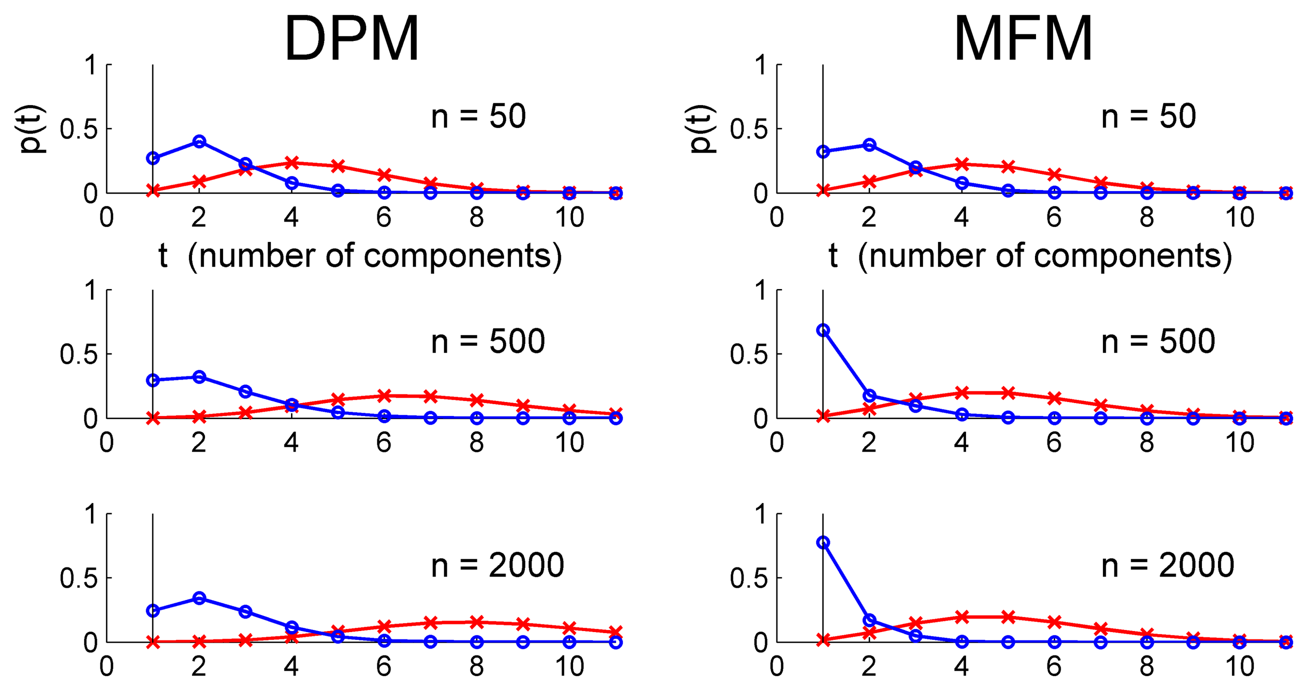
Start with a stick of unit length.
Break off i.i.d. $\mathrm{Exponential}(\lambda)$ pieces until you run out of stick.

Note that this corresponds to a Poisson process on the unit interval.

Empirical Results

Toy example #1: One normal component

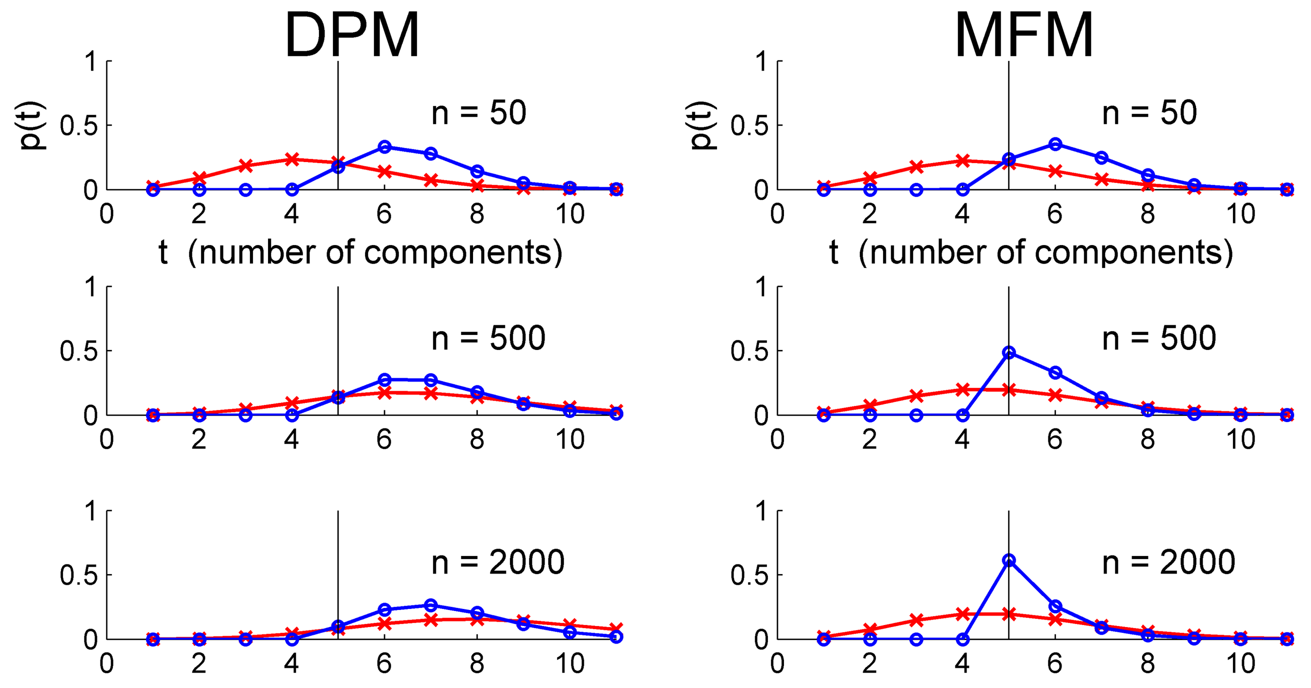
Prior (x) and estimated posterior (o) of the number of clusters, T_n



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: Five normal components

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

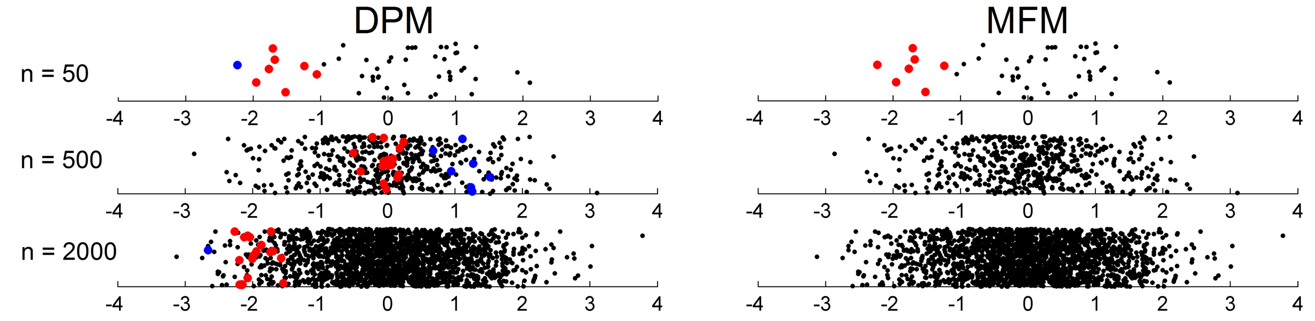


Data: $\sum_{k=-2}^{2} \frac{1}{5}\mathcal{N}(4k, \frac{1}{2})$. Each plot is the average over 5 datasets. Burn-in: 10,000 sweeps, Sample: 100,000 sweeps.

Typical cluster assignments for Example #1

Empirically, DPMs like to have several tiny clusters in addition to the "dominant" ones, while MFMs prefer only dominant clusters.

Cluster #1, Cluster #2, Cluster #3, …



Data: $\mathcal{N}(0, 1)$. (For visualization purposes, the datapoints have been vertically jittered.)

Theoretical Results

Inconsistency results

Theorem (Exponential families)

If
1. $\{p_\theta : \theta \in \Theta\}$ is an exponential family,
2. the base measure H is a conjugate prior, and
3. the concentration parameter $\alpha > 0$ is any fixed value,

then for any "true" mixing measure q_0 with finite support, the DPM posterior on the number of clusters T_n is not consistent (that is, it does not converge to the true number of components).

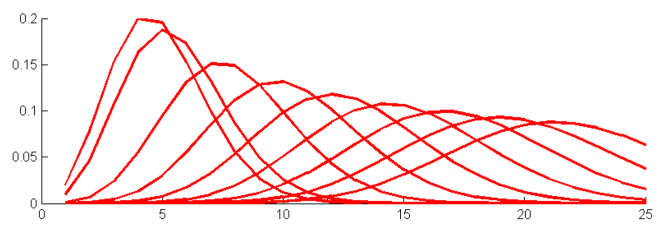
Consider a "standard normal DPM": $p_\theta(x) = \mathcal{N}(x \mid \theta, 1)$ and H is $\mathcal{N}(0, 1)$.

Theorem (Prior on the concentration parameter)

For a standard normal DPM, this inconsistency remains when the concentration parameter α is given a Gamma prior.

The wrong intuition

It is tempting to think that the prior on T_n is the culprit, since $P_{\text{DPM}}(T_n = t) \sim \mathrm{Poisson}(t-1 \mid \alpha \log n)$ as $n \to \infty$.



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!

- For each $n = 1, 2, \ldots$ let $p_n(t)$ be a p.m.f. on $\{1, \ldots, n\}$.
- Define the "tilted" model: $P_{\text{TILT}}(X_{1:n}, T_n = t) = P_{\text{DPM}}(X_{1:n} \mid T_n = t)\, p_n(t)$.
- Call $\{p_n\}$ "non-degenerate" if for all $t = 1, 2, \ldots$, $\liminf_{n \to \infty} p_n(t) > 0$.

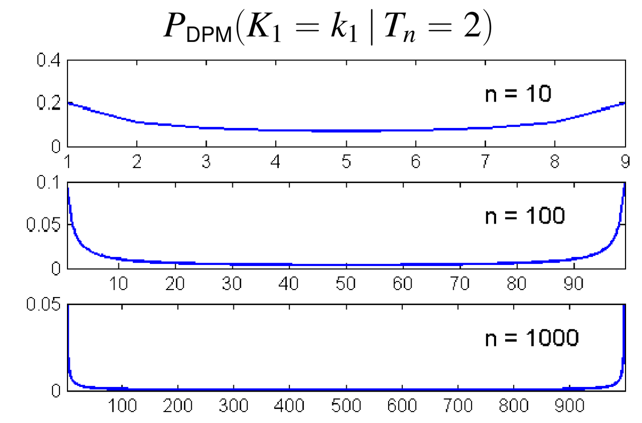
Theorem (Tilted models)

For any non-degenerate sequence p_n, under the tilted model P_{TILT} based on the standard normal DPM, the posterior of T_n is not consistent.

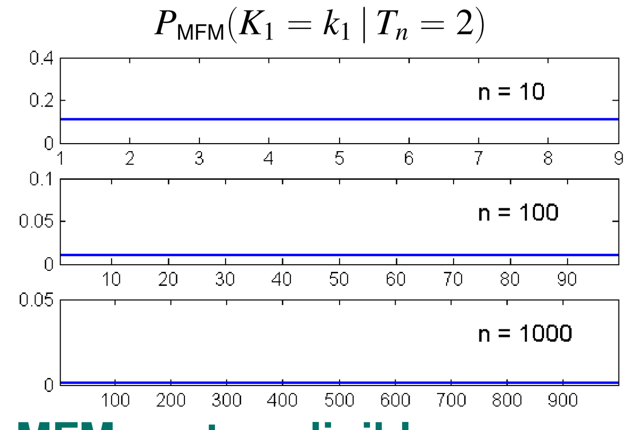
The right intuition

Let K_i be the size of cluster i.

$$P_{\text{DPM}}(K = k \mid T_n = t) \propto k_1^{-1} \cdots k_t^{-1} \qquad P_{\text{MFM}}(K = k \mid T_n = t) \underset{\sim}{\propto} k_1^{\gamma - 1} \cdots k_t^{\gamma - 1}$$



DPMs heavily favor having many small clusters.



MFMs put negligible mass on such partitions.