

Introduction

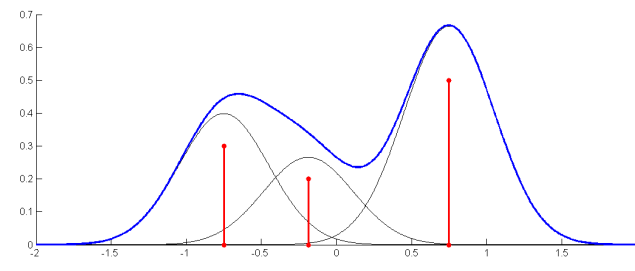
Summary

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)

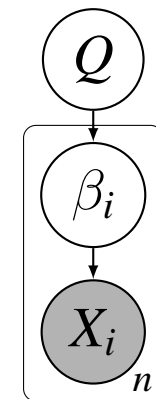
$X_1, X_2, \dots \stackrel{\text{iid}}{\sim} f_{q_0}$ for some q_0 with finite support.

Model distribution

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

Note that $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} f_Q$ (given Q).

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



Questions of convergence

In a DPM, $Q \sim \text{DP}(\alpha, H)$. 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)
... for the mixing measure?	Yes (optimal rate)	Yes
... for the number of components?	Not consistent	Yes

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

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

A Consistent Alternative

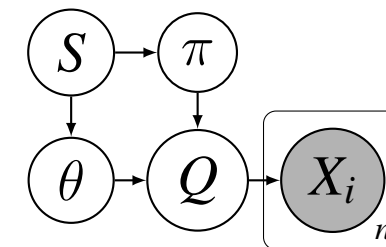
A mixture of finite mixtures (MFM)

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

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

MFM model

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



For convenience, we suggest $p(s) = \text{Poisson}(s - 1 | \lambda)$ 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 $\{1, \dots, n\}$ into t parts, then

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

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

- Here, $s_{(t)} = s(s-1) \dots (s-t+1)$ and $s^{(n)} = s(s+1) \dots (s+n-1)$.
- 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	$ c $	$(c + 1)\kappa(n, t)$
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

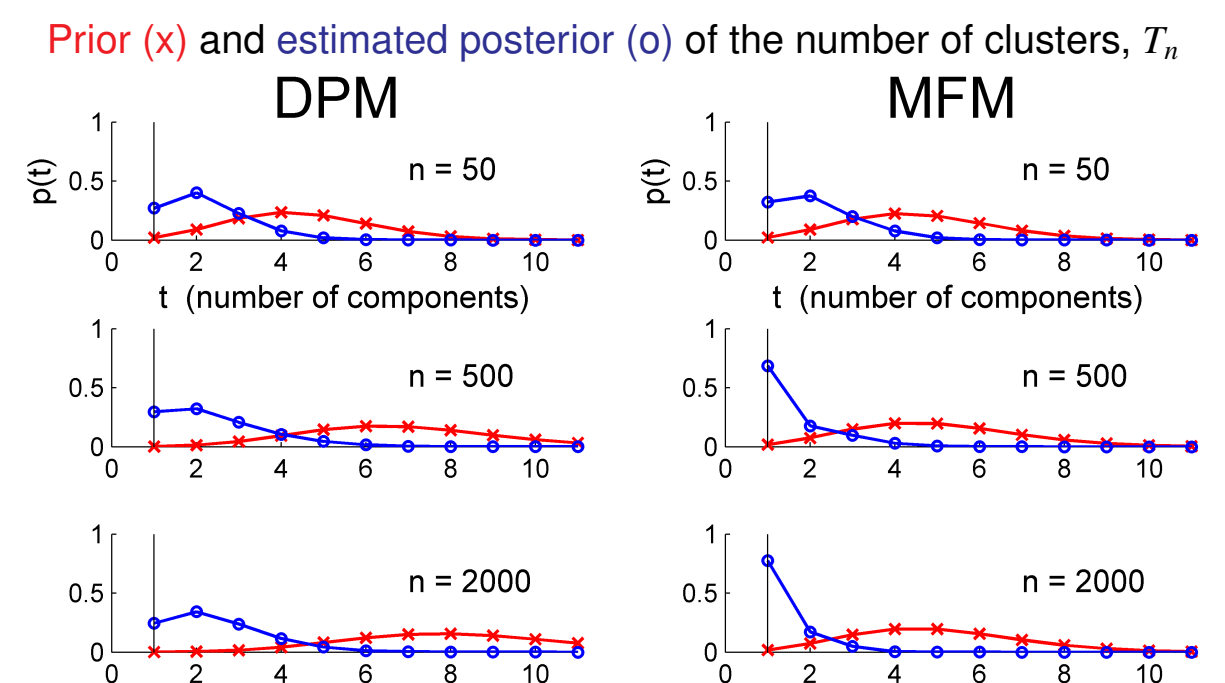
When $\gamma = 1$, the marginal distribution of π is beautifully simple:

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

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

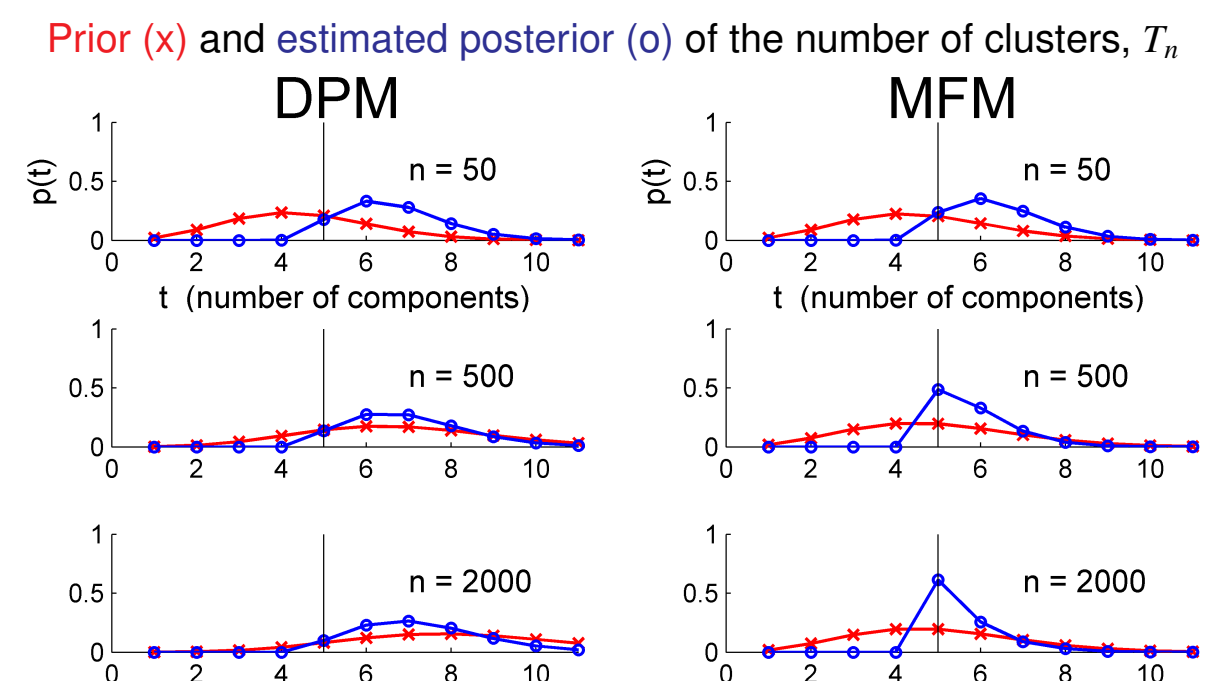
Empirical Results

Toy example #1: One normal component



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

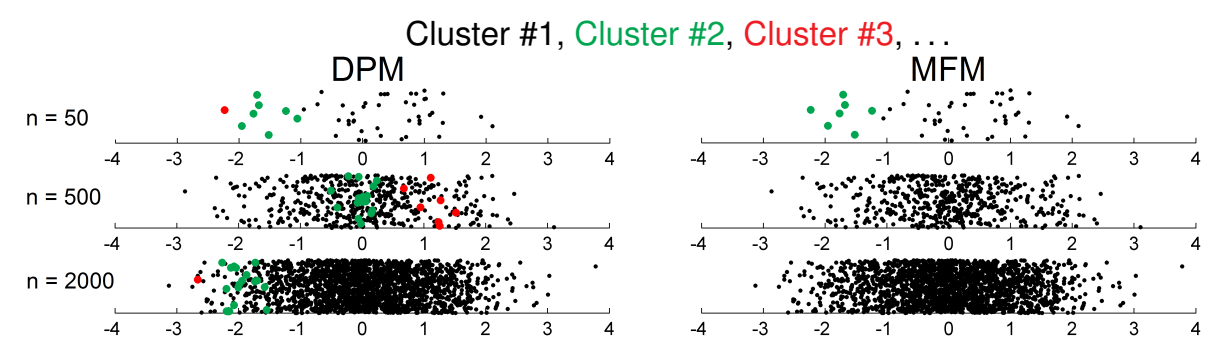
Toy example #2: Five normal components



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.



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

Theoretical Results

Inconsistency of DPMs

Theorem (Exponential families)

If

- $\{p_\theta : \theta \in \Theta\}$ is a (sufficiently regular) exponential family,
- the base measure H is a conjugate prior, and
- 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).

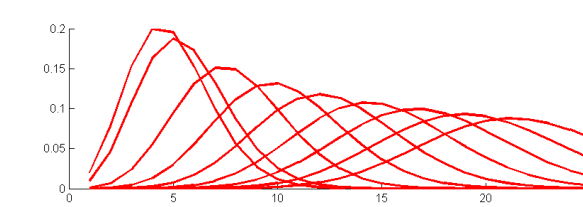
Consider a “standard normal DPM”: $p_\theta(x) = \mathcal{N}(x | \theta, 1)$ and H is $\mathcal{N}(0, 1)$.

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 it is diverging as $n \rightarrow \infty$.



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

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

Theorem (Tilted models)

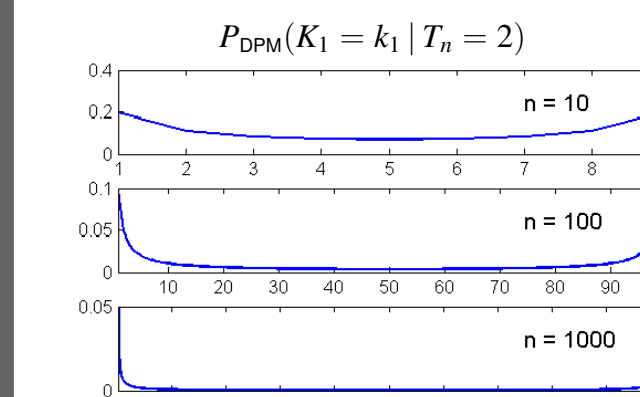
For any non-degenerate sequence p_n , under the tilted model P_{TILT} based on the standard normal DPM, the posterior of T_n is not consistent.

The right intuition

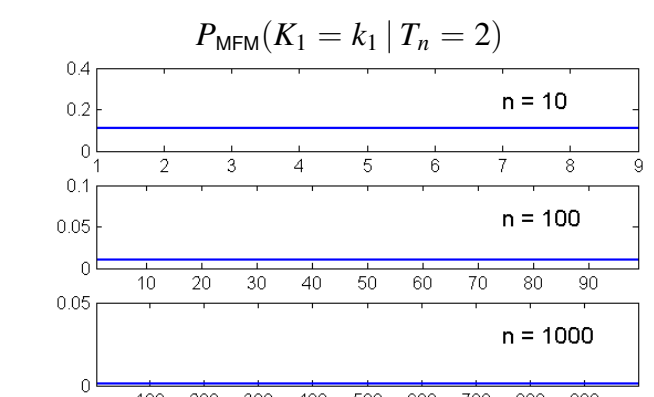
Let K_i be the size of cluster i .

$$P_{\text{DPM}}(K = k | T_n = t) \propto k_1^{-1} \dots k_t^{-1}$$

$$P_{\text{MFM}}(K = k | T_n = t) \propto k_1^{\gamma-1} \dots k_t^{\gamma-1}$$



DPMs heavily favor having many small clusters.



MFMs put negligible mass on such partitions.