Dirichlet process mixtures are inconsistent for the number of components in a finite mixture

Jeffrey W. Miller
and
Matthew T. Harrison

Division of Applied Mathematics
182 George Street
Providence, RI 02912

ICERM, September 17, 2012

Outline of the talk

Outline of the talk

Notational preliminaries

- 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 $\theta_1, \theta_2, \ldots \in \Theta$ and δ_θ is the unit point mass at $\theta \in \Theta$.
- Let f_q denote the density of the resulting mixture, that is,

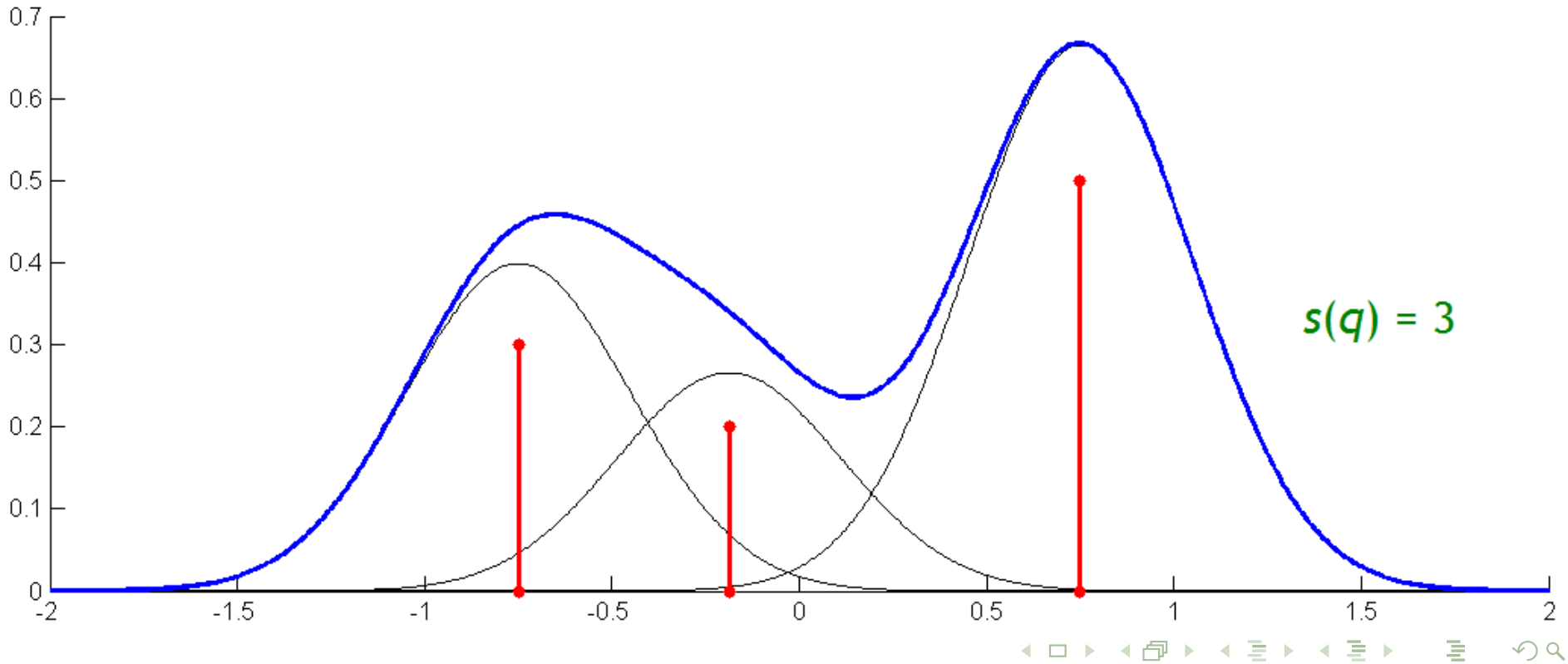
$$f_q(x) = \int_\Theta p_\theta(x)\, dq(\theta) = \sum_{i=1}^{\infty} \pi_i p_{\theta_i}(x).$$

- Let $s(q) = |\operatorname{support}(q)| \in \{1, 2, \ldots\} \cup \{\infty\}$.
- Assume identifiability in the sense that $f_q = f_{q'} \Rightarrow q = q'$ for any q, q' with finite support.

Notational preliminaries

$q = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$ (mixing distribution)
$f_q(x) = \sum \pi_i p_{\theta_i}(x)$ (density)
$s(q) = |\text{support}(q)|$ (number of components)

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



Two distributions

Notation: $q = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$, $f_q(x) = \sum \pi_i p_{\theta_i}(x)$, $s(q) = |\,\text{support}(q)|$.

Data distribution (the "true" distribution)

$X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_{q_0}$ for some q_0 with $s(q_0) < \infty$.

Model distribution

$Q \sim$ some prior on discrete measures q,
$X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_Q$ (given Q).

Two distributions

Notation: $q = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$, $f_q(x) = \sum \pi_i p_{\theta_i}(x)$, $s(q) = |\,\text{support}(q)|$.

Data distribution (the "true" distribution)

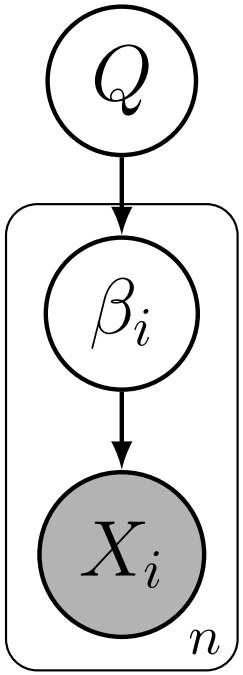
$X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_{q_0}$ for some q_0 with $s(q_0) < \infty$.

Model distribution

$Q \sim$ some prior on discrete measures q,
$X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_Q$ (given Q).

Model distribution (equivalent formulation)

$Q \sim$ some prior on discrete measures q,
$\beta_1, \beta_2, \ldots \overset{\text{iid}}{\sim} Q$ (given Q),
$X_i \sim p_{\beta_i}$ (given $Q, \beta_1, \beta_2, \ldots$) indep. for $i = 1, 2, \ldots$.



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

Many possible questions

Data: $X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_{q_0}$. Write $X_{1:n} = (X_1, \ldots, X_n)$.

Model: $Q \sim$ prior, $\beta_i \overset{\text{iid}}{\sim} Q$, $X_i \sim p_{\beta_i}$, and $T_n = \#\{\beta_1, \ldots, \beta_n\}$.

Is the posterior consistent (and at what rate of convergence) . . .

1. . . . for the density?

   i.e. $P_{\text{model}}(\text{dist}(f_Q, f_{q_0}) < \varepsilon \,|\, X_{1:n}) \xrightarrow[n\to\infty]{P_{\text{data}}} 1 \;\; \forall \varepsilon > 0$?

   (Also, does this hold at any sufficiently smooth density, even when it is not a mixture from $\{p_\theta : \theta \in \Theta\}$?)

2. . . . for the mixing distribution?

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

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: We use T_n instead of $s(Q)$ since $s(Q) \overset{\text{a.s.}}{=} \infty$ in a DPM.)

Answers for Dirichlet process mixtures (DPMs)

In a DPM, $Q \sim \mathrm{DP}(\alpha H)$.
Is the posterior consistent (and at what rate of convergence)...

	DPMs
...for the density?	Yes (optimal rate) (Ghosal & van der Vaart 2001, 2007)
...for the mixing distribution?	Yes (optimal rate) (Nguyen 2012)
...for the number of components?	Not consistent

This holds for any sufficiently smooth density (in a certain sense).
Contributions also by: Lijoi, Prünster, Walker, James, Tokdar, Dunson, Bhattacharya, Ghosh, Ramamoorthi, Wu, Khazaei, Rousseau, Balabdaoui, Tang

(Note: Ignoring tiny components when computing T_n might fix this issue.)

Outline of the talk

Mixture of finite mixtures (MFM)

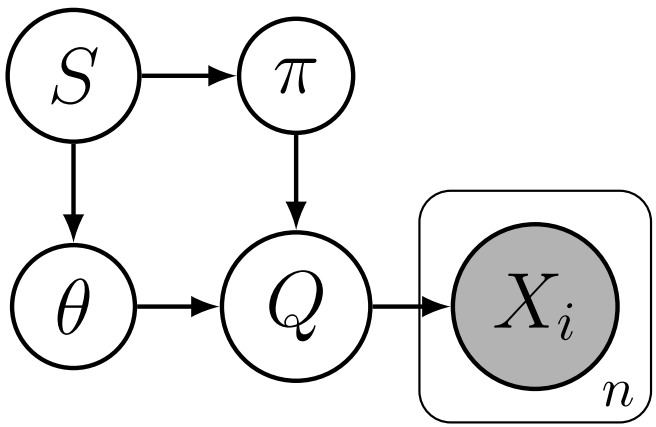
Many authors have considered the following natural alternative to DPMs.

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 $\{1, 2, \dots\}$
$\pi \sim \mathrm{Dirichlet}(\alpha_{s1}, \dots, \alpha_{ss})$ (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 $\{p_\theta\}$
- $p(s) = \mathrm{Poisson}(s - 1 \,|\, \lambda)$
- $\alpha_{ij} = \alpha > 0$ for all i, j

Answers for MFM models

Is the posterior consistent (and at what rate of convergence)...

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

Doob's theorem gives consistency at Lebesgue almost-all mixing distributions q_0.
For any sufficiently smooth density, convergence at the optimal rate was proven by Kruijer (2008) and Kruijer, Rousseau, & van der Vaart (2010) (in the same sense as for DPMs).

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

Doob's theorem guarantees consistency, as before. Optimal rate?

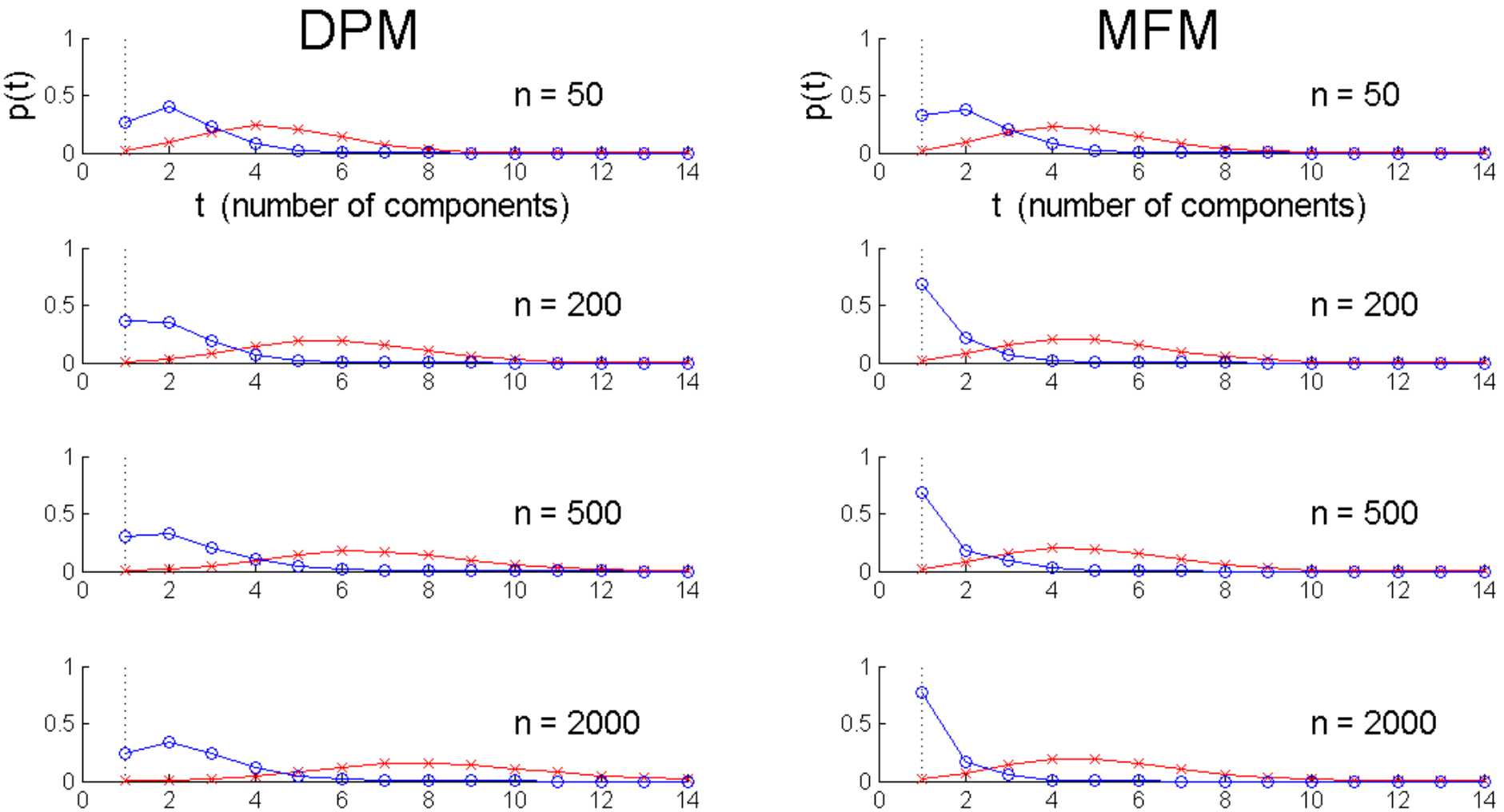
...for the number of components?	Not consistent	**Yes**

By Doob's theorem, again.

Outline of the talk

Toy example #1: One normal component

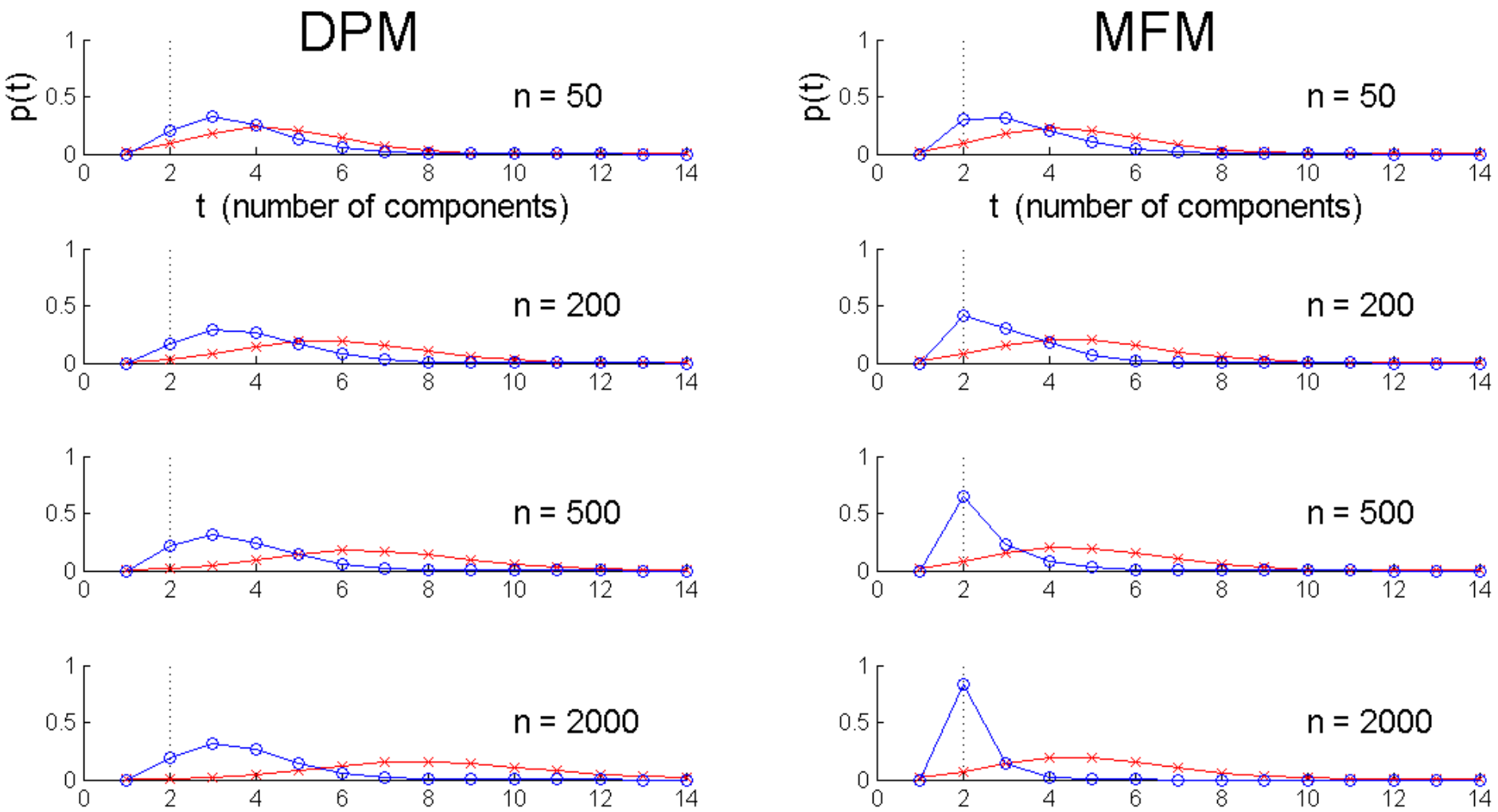
Prior (x) and estimated posterior (o) of T_n



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: Two normal components

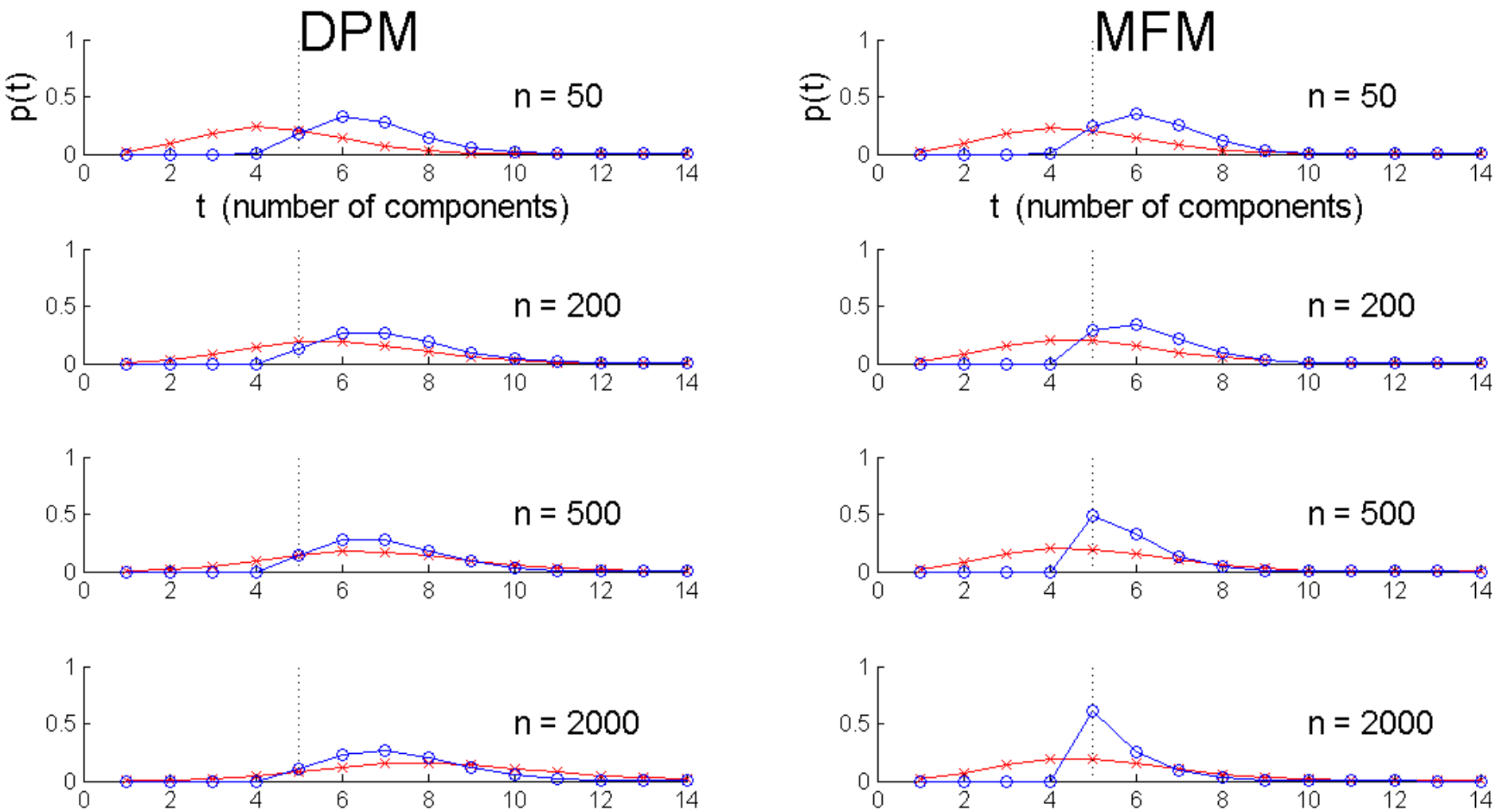
Prior (x) and estimated posterior (o) of T_n



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 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.

Outline of the talk

Inconsistency results

Theorem (Exponential families)

If:

- $\{p_\theta : \theta \in \Theta\}$ *is an 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 distribution q_0 *with* $s(q_0) < \infty$*, the DPM posterior on* T_n *is not consistent, that is,* $P_{DPM}(T_n = s(q_0) \,|\, X_{1:n})$ *does not converge to* 1*.*

Remarks:

- To be precise, the theorem applies to any regular full-rank exponential family in natural form, where Θ is the natural parameter space.
- For instance, this covers: multivariate Gaussian, Gamma, Poisson, Exponential, Geometric, Laplace, and others.

Inconsistency results

"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.

Theorem (The posterior can be "badly" inconsistent)

If $X_1, X_2, \ldots \overset{iid}{\sim} \mathcal{N}(0, 1)$ (i.e. there is one standard normal component), then

$$P_{DPM}(T_n = 1 \,|\, X_{1:n}) \xrightarrow[n\to\infty]{Pr} 0$$

under a standard normal DPM with any fixed value of $\alpha > 0$.

We conjecture that more generally: for data from any sufficiently regular density, $P_{\text{DPM}}(T_n = t \,|\, X_{1:n}) \to 0$ for all t.

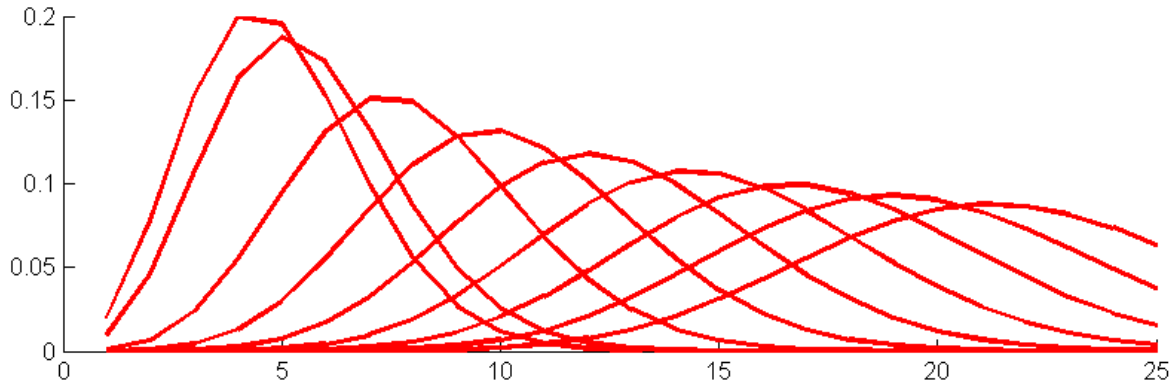
The wrong intuition

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

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

where $\left[\begin{matrix} n \\ t \end{matrix} \right]$ 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!

Replacing the prior on T_n doesn't fix the problem

- For each $n = 1, 2, \ldots$ let $p_n(t)$ be a p.m.f. on $\{1, \ldots, n\}$.
- Define the "tilted" model:

$$P_{\text{TILT}}(X_{1:n}, T_n = t) = P_{\text{DPM}}(X_{1:n} \,|\, T_n = t)\, p_n(t).$$

- Call the sequence 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.

(Recall "Standard normal DPM": $p_\theta(x) = \mathcal{N}(x \,|\, \theta, 1)$ and H is $\mathcal{N}(0, 1)$.)

The right intuition

- Let $A = (A_1, \ldots, A_t)$ be an ordered partition of $\{1, \ldots, n\}$.
  Let $K = (K_1, \ldots, K_t)$ where $K_i = |A_i|$ and assume $K_1, \ldots, K_t > 0$
  (e.g. $A = (\{3, 5\}, \{1\}, \{2, 4, 6\})$, $K = (2, 1, 3)$).
- The distributions over A and $K|T_n = t$ in a DPM are

$$P_{\text{DPM}}(A) = \frac{1}{n!\, t!} \prod_{i=1}^{t} (K_i - 1)! \quad \text{and} \quad P_{\text{DPM}}(K = k | T_n = t) \propto \frac{1}{k_1 \cdots k_t}.$$

- This distribution heavily favors partitions with many small k's.
- It turns out that the likelihood is not strong enough to overcome this effect — the likelihood "does not mind" adding tiny superfluous parts.

The right intuition

- If the likelihood "does not mind" adding tiny superfluous parts, then how is it possible for MFM models to be consistent?
- The answer is that MFM models put negligible prior mass on such partitions.

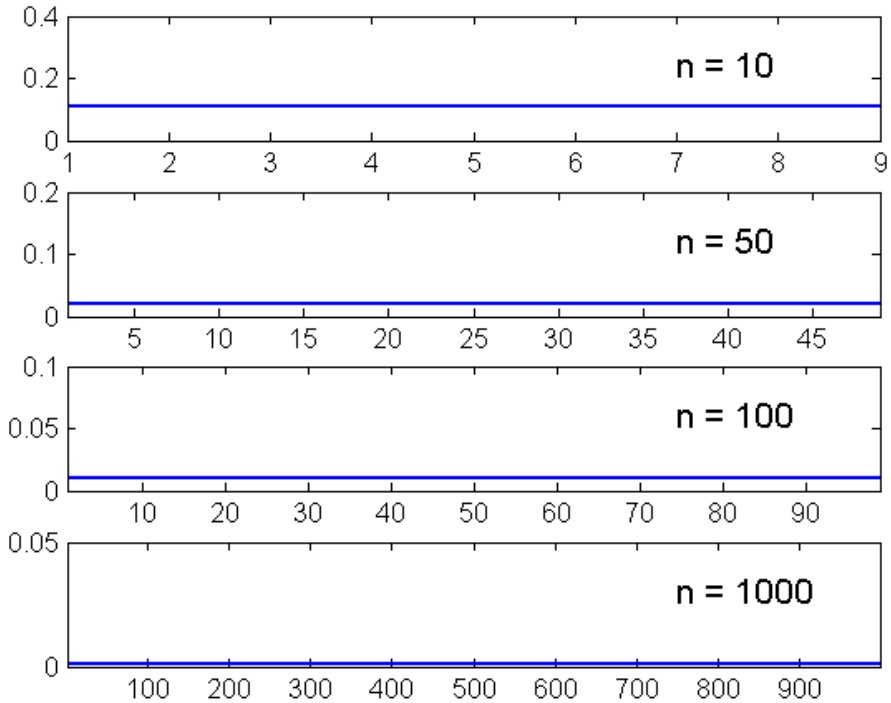
$$P_{\text{MFM}}(k|T_n = t) \propto k_1^{\alpha-1} \cdots k_t^{\alpha-1}$$

$$P_{\text{MFM}}(K_1 \leq n^{\varepsilon} \,|\, T_n = 2) \xrightarrow[n\to\infty]{} 0$$

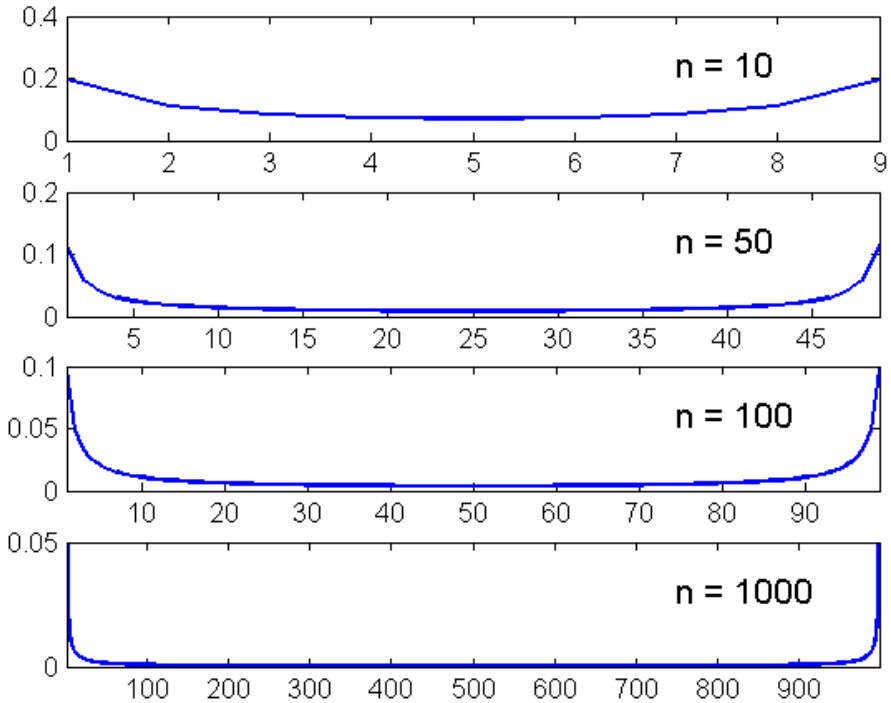
$$P_{\text{DPM}}(k|T_n = t) \propto k_1^{-1} \cdots k_t^{-1}$$

$$P_{\text{DPM}}(K_1 \leq n^{\varepsilon} \,|\, T_n = 2) \xrightarrow[n\to\infty]{} \varepsilon/2$$

$P_{\text{MFM}}(k_1 \,|\, T_n = 2)$



$P_{\text{DPM}}(k_1 \,|\, T_n = 2)$



Outline of the talk

Appropriate and inappropriate usage of DPMs

Appropriate usage:

- for density estimation
  (. . . and not for inferences about the number of components)
  or
- for data assumed to come from a DPM
  (. . . and in particular, there are infinitely many components)
  (A possible example here is topic models.)

Inappropriate usage:

- for inferences about the number of components in a finite mixture

  (Many publications use DPMs in this manner.)

Applications that may be problematic, in retrospect

Population structure / species delimitation

- In population genetics, an important problem is identification of subpopulations of organisms.
- For example, geographic barriers divide populations and genetic drift occurs.
- DPMs are being used to infer the number of groups:
  - Proposals to use DPMs
    - Huelsenbeck & Andolfatto (2007) — 134 citations (as of 9/7/2012)
    - Pella & Masuda (2006) — 54 citations (as of 9/7/2012)
  - Popular software package
    - "Structurama" — Huelsenbeck, Andolfatto, & Huelsenbeck (2011)
  - Methods using DPMs
    - Onogi, Nurimoto, & Morita (2011)
    - Fogelqvist, Niittyvuopio, Agren, Savolainen, & Ascoux (2010)
    - Hausdorf & Hennig (2010)
  - Applications to real-world scientific problems
    - West African forest geckos — Leaché & Fujita (2010)
    - Sardines — Gonzales & Zardoya (2007)
    - Avocados — Chen, Morrell, Ashworth, de la Cruz, & Clegg (2009)
    - Apples — Richards, Volk, Reilley, Henk, Lockwood, Reeves, & Forsline (2009)

Applications that may be problematic, in retrospect

Haplotype inference and founder estimation

- Xing, Sohn, Jordan, & Teh (2006)

Network communities

- Baskerville, Dobson, Bedford, Allesina, Anderson, & Pascual (2011)

Epidemiology

- Choi, Lawson, Cai & Hossain (2011)

Heterotachy (i.e. mutation rates in phylogenetic trees)

- Lartillot & Philippe (2004)
- Rodrigue, Philippe, & Lartillot (2008)
- Zhou, Brinkmann, Rodrigue, Lartillot, & Philippe (2010)
- Huelsenbeck, Jain, Frost, & Pond (2006)

Gene expression profiling

- Medvedovic & Sivaganesan (2002)
- Qin (2006)
- Rasmussen, de la Cruz, Ghahramani, & Wild (2009)

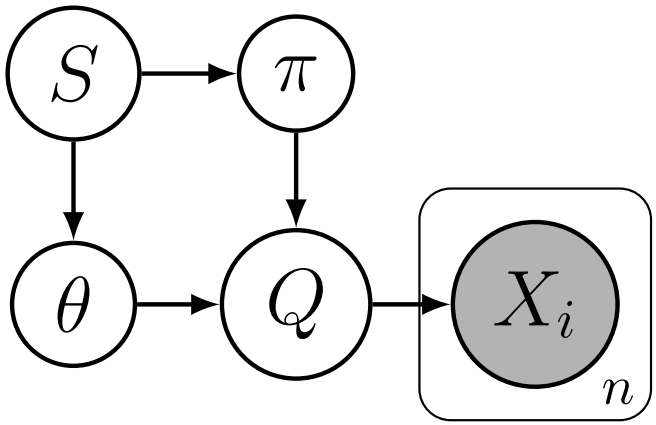
Outline of the talk

Mixture of finite mixtures (MFM)

Recall:

MFM model (Poisson case)

$$S \sim \text{Poisson}(\lambda) + 1$$
$$\pi \sim \text{Dirichlet}_s(\alpha, \ldots, \alpha) \text{ (given } S = s)$$
$$\theta_1, \ldots, \theta_s \overset{\text{iid}}{\sim} H \text{ (given } S = s)$$
$$Q = \sum_{i=1}^{S} \pi_i \delta_{\theta_i}$$
$$X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_Q \text{ (given } Q).$$



MFMs vs DPMs

Similarities between MFMs and DPMs:

- Efficient approximate inference (via Gibbs sampling)
- Appealing equivalent formulations:
  - exchangeable distribution on partitions
  - restaurant process
  - stick-breaking
  - random discrete measures
- Consistent for any sufficiently smooth density (at the optimal rate, in a certain sense)

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

- MFMs are a natural Bayesian extension of finite mixtures.
- Consistency (a.e.) for S, π, θ, and f_Q is automatically guaranteed under very general conditions (by Doob's theorem).

Disadvantages of MFMs (vs DPMs):

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

Properties of MFMs

For clarity, set $\alpha = 1$ in both MFM and DPM.

Exchangeable distribution on partitions (MFM vs DPM)

Let $\mathcal{C}$ be an (unordered) partition of $\{1, \ldots, n\}$ into t parts (e.g. $\mathcal{C} = \{\{3,5\},\{1\},\{2,4,6\}\}$). Then

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

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

- Here, $s_{(t)} = s(s-1)\cdots(s-t+1)$ and $s^{(n)} = s(s+1)\cdots(s+n-1)$.
- The numbers $\kappa(n,t)$ can be efficiently precomputed using $\kappa(n,t) = \kappa(n-1,t-1) - (n+t-2)\,\kappa(n,t-1)$, and $\kappa(n,0) = \mathbb{E}(1/S^{(n)}) = P(S > n)/\lambda^n$

  (the last equality holding only in the Poisson case).

Properties of MFMs

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

Restaurant process (MFM vs DPM)

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

	MFM	DPM
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.

- This is easily verified using the recursion for $\kappa(n, t)$.
- This yields a simple Gibbs sampling scheme ...

Approximate inference with MCMC

- Gibbs sampling for MFMs is nearly identical to Gibbs sampling for DPMs.
- Sampling from $P(\mathcal{C}|x_{1:n}) \propto P(x_{1:n}|\mathcal{C})P(\mathcal{C})$ proceeds as follows.
- Let $\mu(\mathcal{C}) = P(x_{1:n}|\mathcal{C})$. (This is the same for both models.)

Gibbs sampling (MFM vs DPM)

Suppose $\mathcal{C}$ is the current partition, not including customer k.
Reseat customer k...

	MFM	DPM
at table $c \in \mathcal{C}$ with probability $\propto$	$(\lvert c\rvert + 1)\kappa(n, t)\, \mu(\mathcal{C}_c)$	$\lvert c\rvert\, \mu(\mathcal{C}_c)$
or at a new table with probability $\propto$	$\kappa(n, t+1)\, \mu(\mathcal{C}_*)$	$\mu(\mathcal{C}_*)$

where

- $t = |\mathcal{C}|$ is the number of occupied tables (excluding customer k),
- $\mathcal{C}_c$ is the partition formed by assigning k to table c, and
- $\mathcal{C}_*$ is the partition formed by assigning k to a new table.

Approximate inference with MCMC

For both models,

$$\mu(\mathcal{C}) = P(x_{1:n}|\mathcal{C}) = \prod_{c\in\mathcal{C}} m(x_c) \quad \text{where} \quad m(x_c) = \int \prod_{i\in c} p_\theta(x_i)\, dH(\theta).$$

As usual, $\mu(\mathcal{C})$ can be computed analytically when H is a conjugate prior.

Gibbs sampling (MFM vs DPM)

Suppose $\mathcal{C}$ is the current partition, not including customer k.
Reseat customer k...

	MFM	DPM
at table $c \in \mathcal{C}$ with probability $\propto$	$(\lvert c\rvert + 1)\kappa(n, t)\, \mu(\mathcal{C}_c)$	$\lvert c\rvert\, \mu(\mathcal{C}_c)$
or at a new table with probability $\propto$	$\kappa(n, t+1)\, \mu(\mathcal{C}_*)$	$\mu(\mathcal{C}_*)$

where

- $t = |\mathcal{C}|$ is the number of occupied tables (excluding customer k),
- $\mathcal{C}_c$ is the partition formed by assigning k to table c, and
- $\mathcal{C}_*$ is the partition formed by assigning k to a new table.

Stick-breaking construction

Recall that $S \sim \text{Poisson}(\lambda) + 1$ and $\pi | S = s \sim \text{Dirichlet}_s(\alpha, \ldots, \alpha)$.
When $\alpha = 1$, the marginal distribution of π is beautifully simple:

Stick-breaking for MFM (Poisson-Uniform case)

Let $Y_1, Y_2, \ldots \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$.
Let $\pi_k = \min\{Y_k,\ 1 - \sum_{i=1}^{k-1} \pi_i\}$ for $k = 1, 2, \ldots$.

Then $S := \#\{k : \pi_k > 0\} \sim \text{Poisson}(\lambda) + 1$
and $(\pi_1, \ldots, \pi_s) | S = s \sim \text{Dirichlet}_s(1, \ldots, 1)$.

In other words, we have the following stick-breaking construction:

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.

Outline of the talk

Open questions

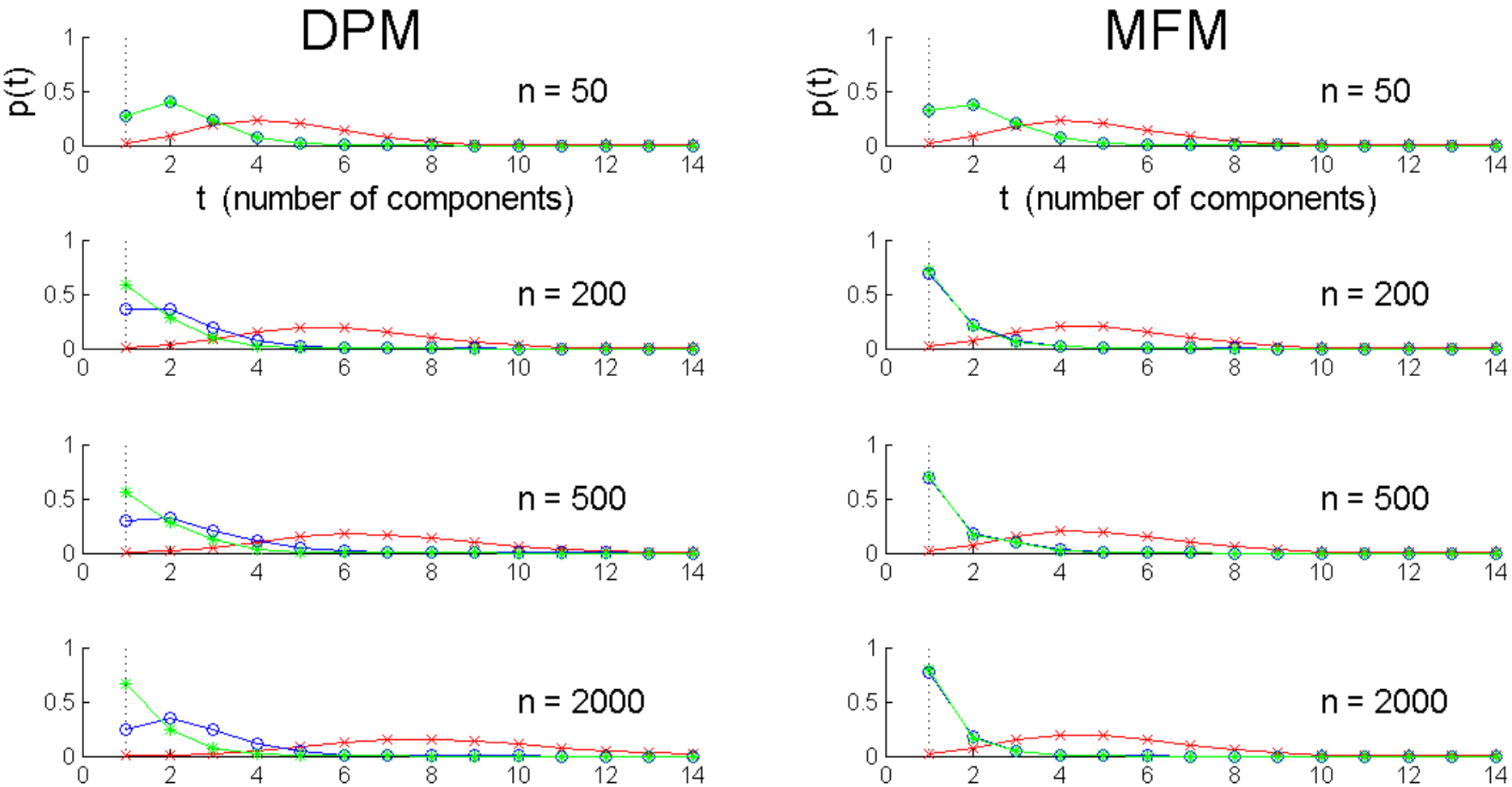
1. Does "pruning" tiny DPM components result in consistency?
2. Does the DPM posterior of T_n diverge?
   i.e. does $P_{\text{DPM}}(T_n = t \,|\, X_{1:n})$ always go to 0 for all t?
3. What rate of convergence do MFMs have for the mixing distribution?
   . . . for the number of components?
4. How well do MFMs perform in practice, compared to DPMs?

Additional material

Additional material

Toy example #1: One normal component

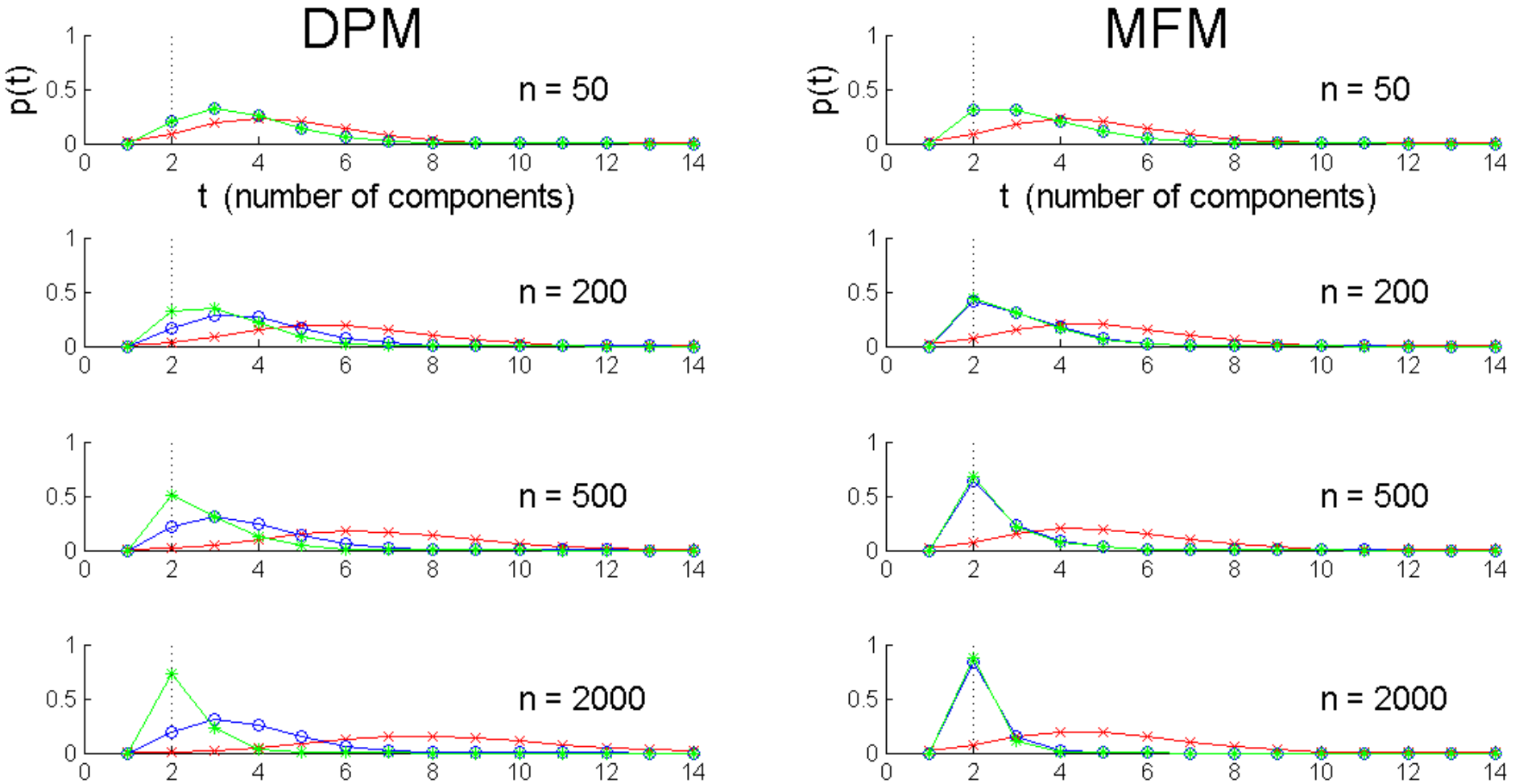
Prior (x) of T_n, estimated posterior (o) of T_n, and estimated posterior ($*$) of $T_{n,\delta}$ with $\delta = 0.01$



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

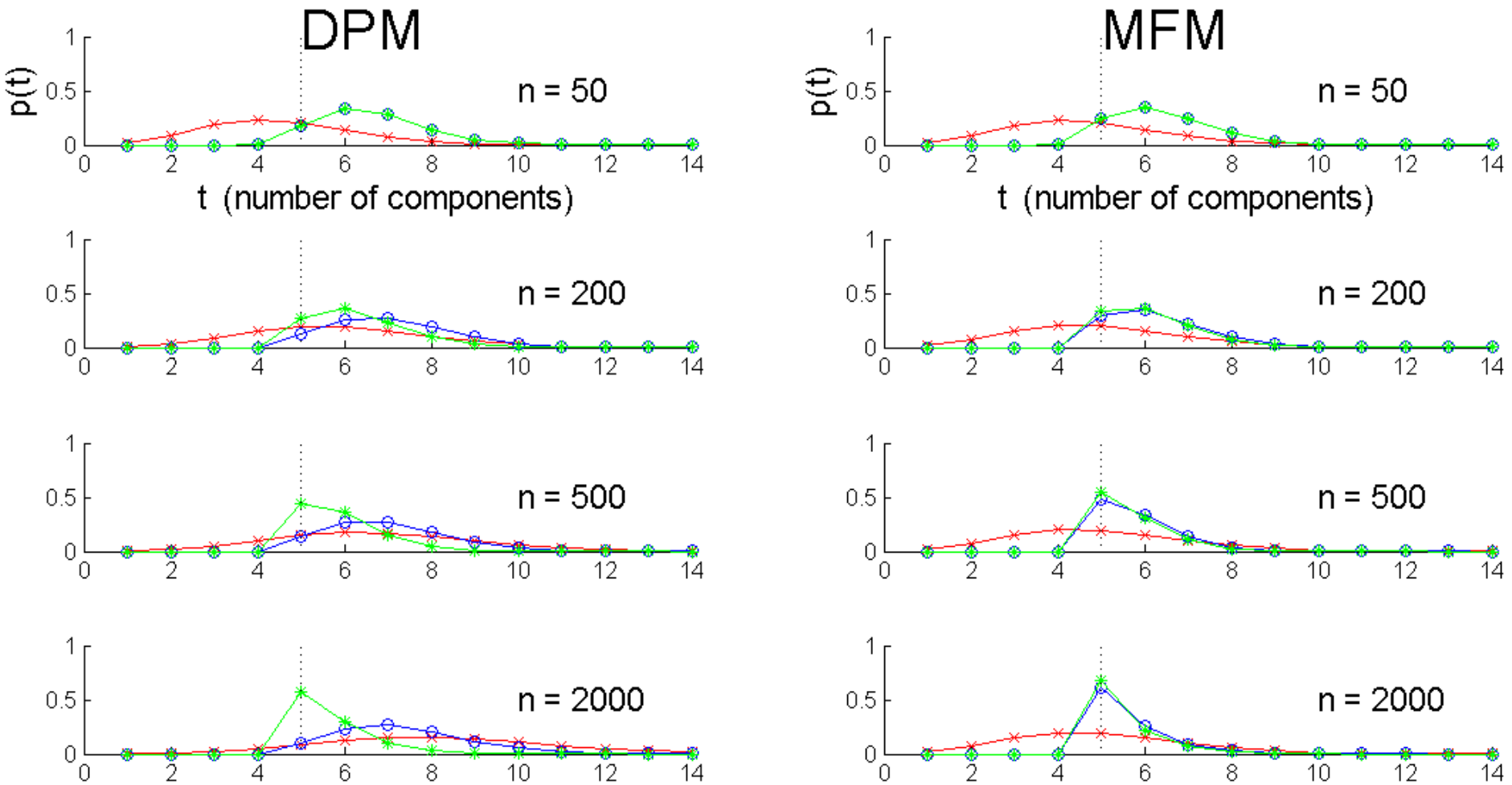
Prior (x) of T_n, estimated posterior (o) of T_n, and estimated posterior ($*$) of $T_{n,\delta}$ with $\delta = 0.01$



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) of T_n, estimated posterior (o) of T_n, and estimated posterior ($*$) of $T_{n,\delta}$ with $\delta = 0.01$



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.