Combinatorial stochastic processes for variable-dimension models

Jeffrey W. Miller

Joint work with Matt Harrison

(Research supported by DARPA and the NSF)

Duke University
Department of Statistical Science

Texas A&M Statistics Colloquium
Oct 31, 2014

Nonparametric Bayesian models have found many applications . . .

- astronomy
- epidemiology
- gene expression profiling
- haplotype inference
- medical image analysis
- survival analysis
- extreme value analysis
- meteorology

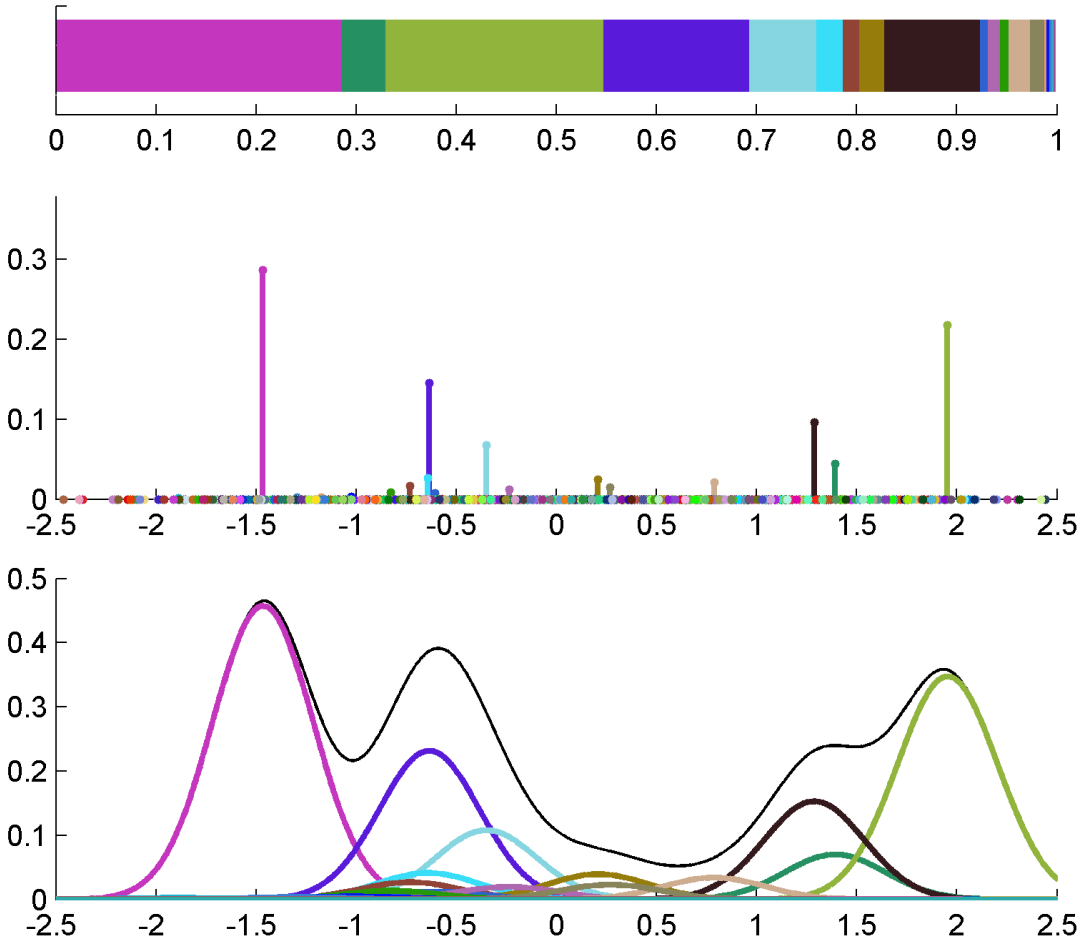
. . . . . .

- econometrics
- phylogenetics
- species delimitation
- computer vision
- classification
- document modeling
- cognitive science
- natural language processing

. . . . . .

- Many nonparametric models are an infinite-dimensional limit of a family of finite-dimensional models.
- Another way to construct a flexible Bayesian model is to put a prior on the dimension — i.e., to use a variable-dimension model.
- For example, putting a prior on the number of components in a finite mixture.

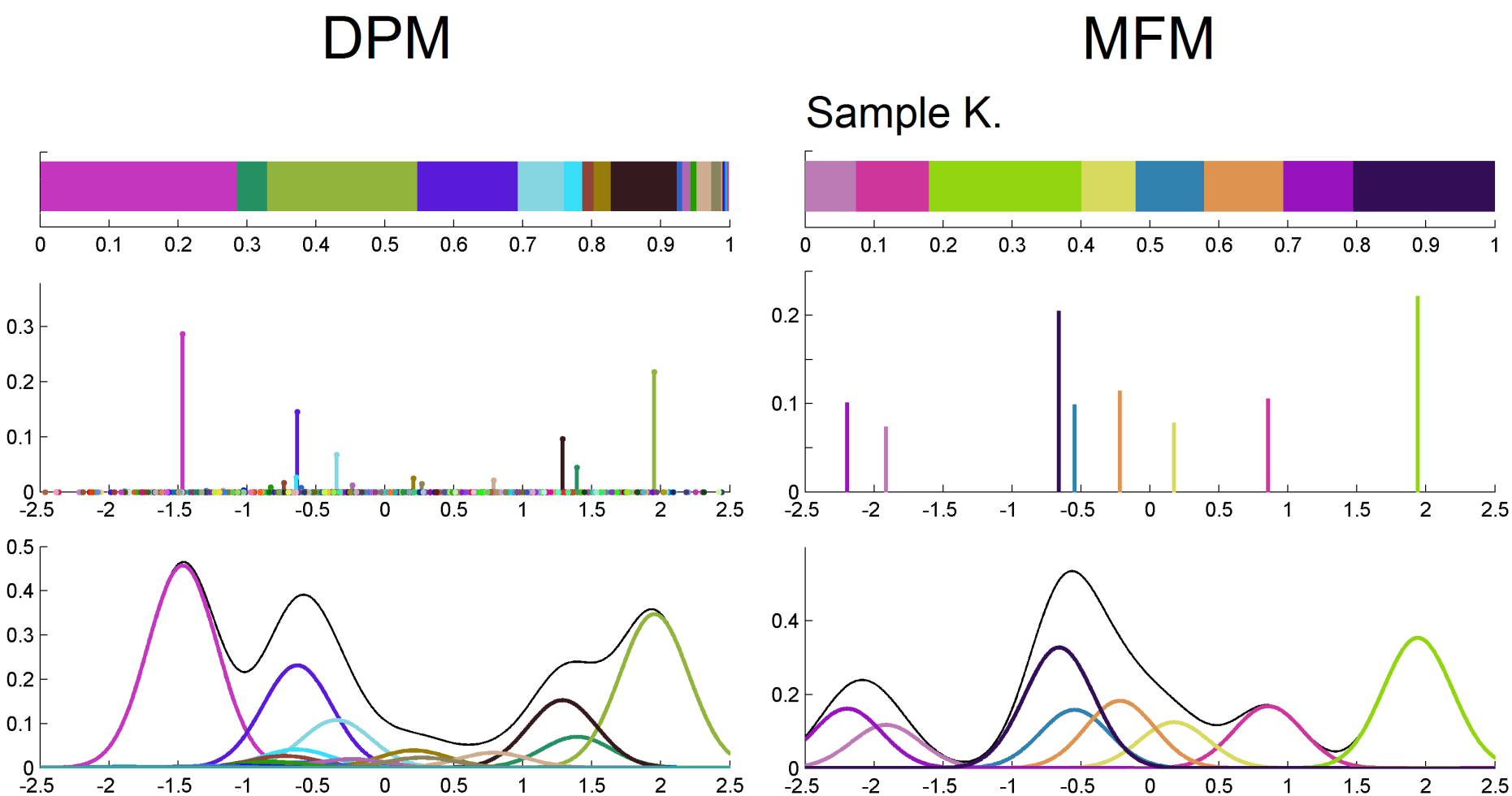
Dirichlet process mixture (DPM)



Ferguson (1983), Lo (1984), Sethuraman (1994),

West, Müller, and Escobar (1994), MacEachern (1994), ...

Mixture of finite mixtures (MFM)



Nobile (1994, 2007), Richardson & Green (1997, 2001), Stephens (2000), . . .

Why use a variable-dimension model?

1. Control
   - . . . over the distribution of the number of clusters/topics/features
   - . . . over the distribution of the relative sizes of clusters/topics/features
2. Interpretability
   - Cleaner clusters/topics/features (no tendency to make tiny superfluous groups)
   - Natural Bayesian approach for a data distribution of unknown complexity (if something is unknown, put a prior on it)
3. Theory can be simpler
   - The parameter space is a countable union of finite-dimensional spaces, rather than an infinite-dimensional space.

The main disadvantage, it seems, is the difficulty of doing inference.

How to do inference in a variable-dimension model?

- Reversible jump MCMC (Green, 1995) is the standard approach.
- Reversible jump is very general and has been used in many applications, but it is not a "black box".
- In contrast, a nice aspect of many of the nonparametric samplers is that they are fairly generic.
  - Green & Richardson (2001): "*In view of the intimate correspondence between DP and [MFM] models discussed above, it is interesting to examine the possibilities of using either class of MCMC methods for the other model class. We have been unsuccessful in our search for incremental Gibbs samplers for the [MFM] models . . .* "
- The key to such samplers is that the model can be characterized by a nice distribution on combinatorial structures (e.g., the Chinese restaurant process, in the case of the DPM).

This talk

The main point of this talk is that similar distributions on combinatorial structures exist for certain variable-dimension models:

- mixture of finite mixtures (similar to the DPM),
- hierarchical mixture of finite mixtures (similar to the HDP), and
- mixture of finite feature models (similar to the IBP).

This enables many of the inference algorithms developed for the infinite-dimensional models to be directly applied to their variable-dimension counterparts.

Outline

1. Mixture of finite mixtures (MFM)
2. Hierarchical mixture of finite mixtures (HMFM)
3. Mixture of finite feature models (MFFM)

Outline

Mixture of finite mixtures (MFM) model

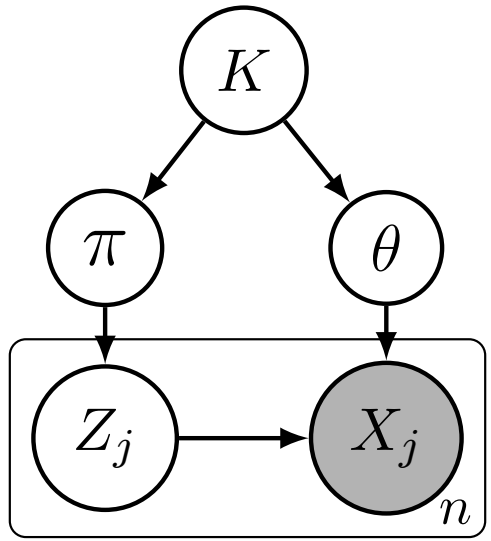
$K \sim p(k)$, a p.m.f. on $\{1, 2, \dots\}$

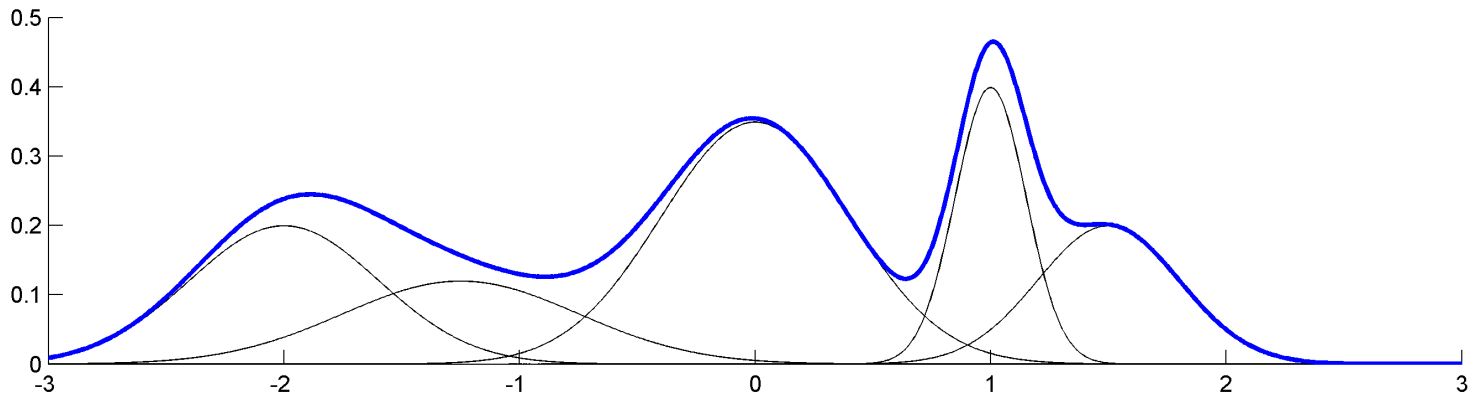
$(\pi_1, \dots, \pi_k) \sim \text{Dirichlet}(\gamma, \dots, \gamma)$, given $K = k$

$\theta_1, \dots, \theta_k \overset{\text{iid}}{\sim} H$, given $K = k$

$Z_1, \dots, Z_n \overset{\text{iid}}{\sim} \pi$, given π

$X_j \sim f_{\theta_{Z_j}}$ for $j = 1, \dots, n$, given $\theta_{1:K}, Z_{1:n}$.





Nobile (1994, 2007), Richardson & Green (1997, 2001), Stephens (2000), ...

Partition distribution

- Letting $\mathcal{C}$ denote the partition of $[n] := \{1, \ldots, n\}$ induced by $Z_1, \ldots, Z_n$, we have

$$p(\mathcal{C}) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)}$$

where

$t = |\mathcal{C}|$ is the number of parts in the partition,

$$V_n(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} \, p(k),$$

$x^{(m)} = x(x+1)\cdots(x+m-1)$, and $x_{(m)} = x(x-1)\cdots(x-m+1)$.

- This is a special case of the family of Gibbs partition distributions studied by Gnedin & Pitman (2006).
- For comparison, in the DPM,

$$p_{\text{DPM}}(\mathcal{C}) = \frac{\alpha^t}{\alpha^{(n)}} \prod_{c \in \mathcal{C}} (|c| - 1)!.$$

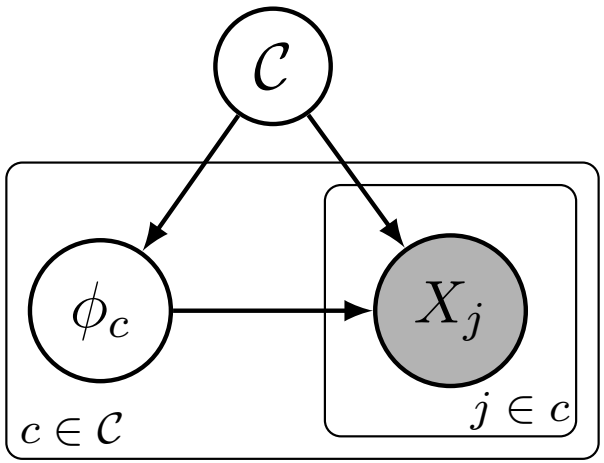
Equivalent representation of the MFM model

$\mathcal{C} \sim p(\mathcal{C})$

For $c \in \mathcal{C}$ independently, given $\mathcal{C}$,

$\phi_c \sim H$

$X_j \sim f_{\phi_c}$ independently for $j \in c$, given ϕ



This representation is useful for doing inference, since one does not have to deal with cluster labels or empty components.

Properties of $V_n(t)$

Recall that $V_n(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} p(k)$.

- For $0 \leq t \leq n$, these numbers satisfy the recursion:

$$V_{n+1}(t+1) = V_n(t)/\gamma - (n/\gamma + t)V_{n+1}(t).$$

  This is a special case of a more general recursion for Gibbs partitions (Gnedin & Pitman, 2006).

- The infinite series usually converges rapidly, since $k_{(t)}/(\gamma k)^{(n)} \leq k^t/(\gamma k)^n$. (It always converges for $0 \leq t \leq n$.)

- If $p(k) = \text{Poisson}(k-1|\lambda)$ and $\gamma = 1$, then

$$V_n(0) = \frac{1}{\lambda^n}\Big(1 - \sum_{k=1}^{n} p(k)\Big).$$

Computing $V_n(t)$

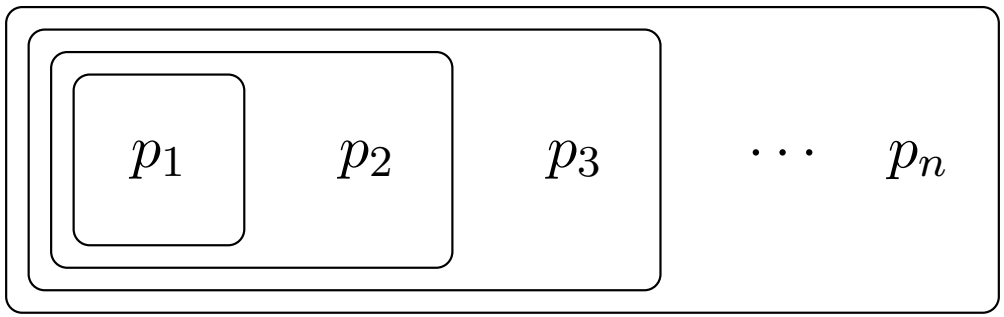
To compute $p(\mathcal{C})$, we need to compute $V_n(t)$. Typically, n will be fixed and we only need it for $t = 1, \ldots, t_{\max}$ for some relatively small $t_{\max}$.

Several possible methods:

1. Numerically approximate $V_n(1), \ldots, V_n(t_{\max})$.
   - Easy, fast, and generally applicable.
2. An analytical expression is available in at least one special case (Gnedin (2010)).
3. Use the recursion along with numerical approximations or analytical expressions for the $t = 0$ cases.

Self-consistent marginals

- For each $n = 1, 2, \ldots,$ let $p_n(\mathcal{C})$ denote the distribution on partitions of $[n]$ defined above.
- Then p_{n-1} coincides with the "marginal" distribution on partitions of $[n-1]$ induced by p_n.
  - In other words, sampling $\mathcal{C} \sim p_n$ and removing n from $\mathcal{C}$ yields a sample from p_{n-1}.
- This is because $Z_{1:m}$ has the same distribution when the model is defined for any $n \geq m$.
  - This can also be derived from the recursion for $V_n(t)$.
- (This is also true in the DPM.)



Restaurant process / Pólya urn process

Further, the sequence of partition distributions $p_1, p_2, \ldots$ can be described by a simple restaurant process.

Restaurant process for MFM & DPM

The first customer sits at a table: $\mathcal{C} = \{\{1\}\}$.

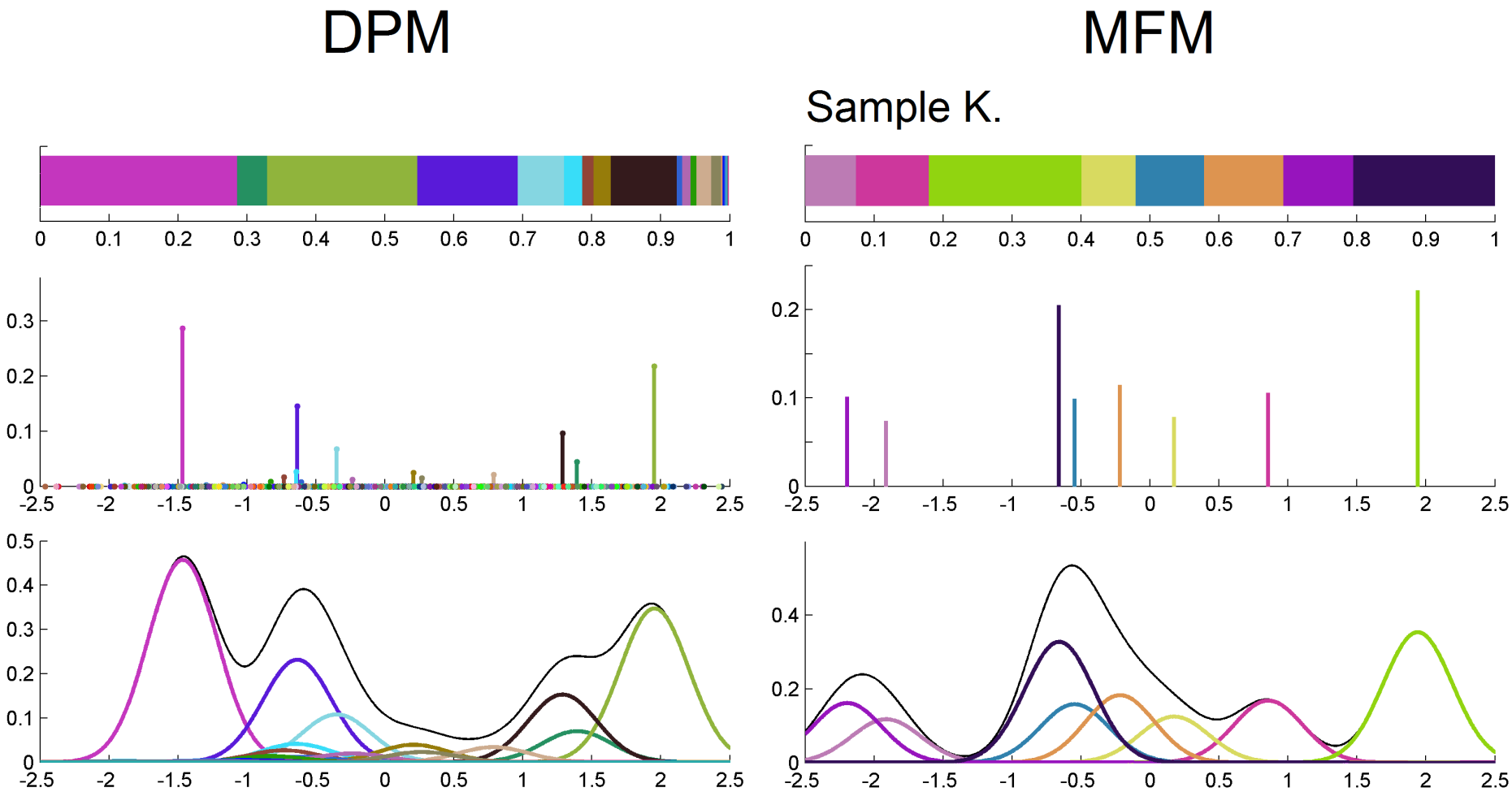
The nth customer sits ...

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

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

Random discrete measures / Species sampling

The MFM can also be formulated starting from a distribution on discrete mixing measures, analogous to the Dirichlet process.



Random discrete measures / Species sampling

Let

$$K \sim p_K(k)$$
$$(\pi_1, \ldots, \pi_k) \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \text{ given } K = k$$
$$\theta_1, \ldots, \theta_k \overset{\text{iid}}{\sim} H, \text{ given } K = k$$
$$G = \sum_{i=1}^{K} \pi_i \delta_{\theta_i}$$

and denote the distribution of G by $\mathcal{M}(p_K, \gamma, H)$.

Then the MFM is obtained by taking $X_1, X_2, \ldots | G$ i.i.d. from the resulting mixture, namely,

$$f_G(x) := \int f_\theta(x) G(d\theta) = \sum_{i=1}^{K} \pi_i f_{\theta_i}(x).$$

Species sampling posterior predictive

- Suppose the base measure H is continuous.
- Then G belongs to the family of "species sampling" models studied by Pitman (1996).
- If $G \sim \mathcal{M}(p_K, \gamma, H)$, and $\beta_1, \beta_2, \ldots \overset{\text{iid}}{\sim} G$ (given G), then the distribution of β_n given $\beta_1, \ldots, \beta_{n-1}$ is proportional to

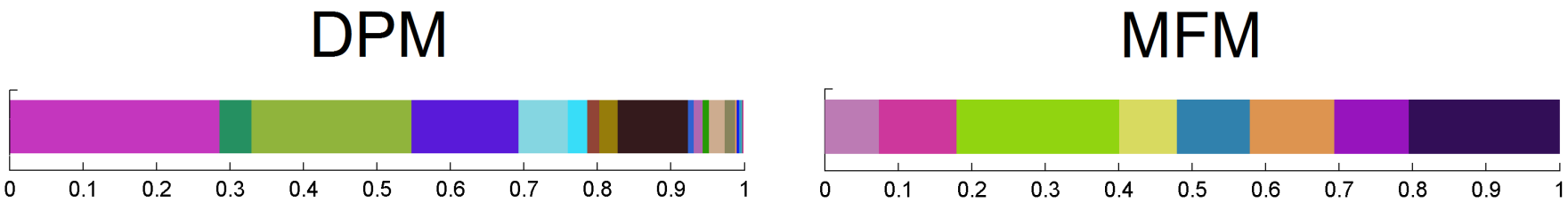
$$\gamma \frac{V_n(t+1)}{V_n(t)} H + \sum_{i=1}^{n-1} \delta_{\beta_i} + \gamma \sum_{i=1}^{t} \delta_{\beta_i^*}$$

  where $\beta_1^*, \ldots, \beta_t^*$ are the distinct values taken by $\beta_1, \ldots, \beta_{n-1}$.
- For comparison, if $G \sim \mathrm{DP}(\alpha, H)$ instead, it is proportional to

$$\alpha H \; + \; \sum_{i=1}^{n-1} \delta_{\beta_i}$$

  (Blackwell & MacQueen, 1973).

Stick-breaking representation in a special case



- The Dirichlet process has an elegant stick-breaking representation for $\pi_1, \pi_2, \ldots$, due to Sethuraman (1994).
- When $p(k) = \text{Poisson}(k-1|\lambda)$ and $\gamma = 1$, the MFM also has a nice stick-breaking representation for $\pi_1, \ldots, \pi_K$:

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

- This corresponds to a Poisson process on the unit interval.
- This suggests a stick-breaking approach to constructing a variety of variable-dimension mixtures, using renewal processes.

Posterior asymptotics

Is the posterior consistent (in cases examined so far) . . .

	MFMs	DPMs
. . . for the density?	**Yes**	**Yes**

MFMs: Kruijer, Rousseau, & Van der Vaart (2010), Nobile (1994)
DPMs: Contributions by Ghosal, Van der Vaart, Tokdar, Lijoi, Prünster, Walker, James, Dunson, Bhattacharya, Ghosh, Ramamoorthi, Wu, Khazaei, Rousseau, Balabdaoui, Tang, and others.

. . . for the mixing measure?	**Yes**	**Yes**

MFMs: Nobile (1994)
DPMs: Nguyen (2013)

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

MFMs: Nobile (1994)
DPMs: M. & Harrison (2014)

Inference algorithms

- Reversible jump MCMC is the usual approach for inference in variable-dimension mixtures (Richardson & Green, 1997).
- However, now that we have established that MFMs have many of the same attractive properties as DPMs, much of the extensive body of work on DPM samplers can be directly applied to them.
- We have applied a number of DPM samplers to the MFM: Algorithm 3 and Algorithm 8 in the terminology of Neal (2000), as well as the split-merge samplers of Jain & Neal (2004, 2007).
- Let's see how this works for two incremental Gibbs samplers:
  1. Algorithm 3 (for conjugate priors), and
  2. Algorithm 8 (for non-conjugate priors).

Incremental Gibbs with a conjugate prior

- For $c \subset \{1, \ldots, n\}$, let $m(x_c) = \int \Big(\prod_{j \in c} f_\theta(x_j) \Big) H(d\theta)$.
- $m(x_c)$ can be computed analytically when H is a conjugate prior.
- The following algorithm for sampling from $p(\mathcal{C}|x_{1:n}) \propto p(x_{1:n}|\mathcal{C})p(\mathcal{C})$ is due to MacEachern (1994) and Neal (1992, 2000) in the DPM case.
- Write $\mathcal{C} \setminus j$ for the current partition, excluding j.

"Algorithm 3" for MFM and DPM

For $j = 1, \ldots, n$: Reseat customer j...

	MFM	DPM
at table $c \in \mathcal{C} \setminus j$ with probability $\propto$	$(\lvert c\rvert + \gamma) \dfrac{m(x_{c \cup j})}{m(x_c)}$	$\lvert c\rvert \dfrac{m(x_{c \cup j})}{m(x_c)}$
at a new table with probability $\propto$	$\gamma \, \dfrac{V_n(t+1)}{V_n(t)} \, m(x_j)$	$\alpha \, m(x_j)$

where $t = \lvert \mathcal{C} \setminus j \rvert$ is the number of occupied tables, excluding customer j.

Incremental Gibbs with a non-conjugate prior

- Often, the selected family $\{f_\theta\}$ will not have a conjugate prior.
- Neal's (2000) Algorithm 8, inspired by MacEachern & Müller (1998), is a clever auxiliary variable method for non-conjugate H.
- The state of the chain is $(\mathcal{C}, \phi)$ where $\phi = (\phi_c : c \in \mathcal{C})$, $\phi_c \in \Theta$.

"Algorithm 8" (with one auxiliary variable) for MFM and DPM

1. For $j = 1, \ldots, n$: If j is seated alone, set $\phi_* = \phi_{\{j\}}$; otherwise, sample $\phi_* \sim H$. Reseat $j \ldots$

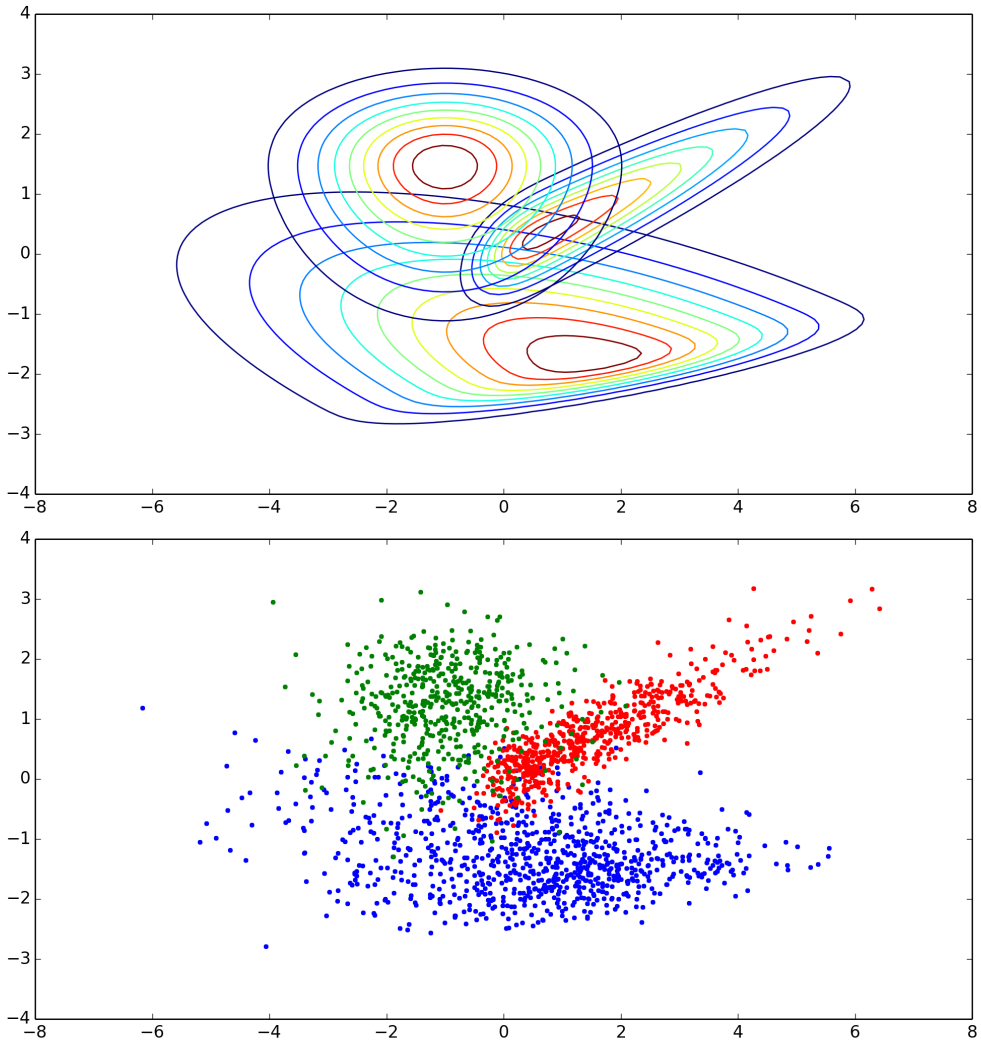
	MFM	DPM
at table $c \in \mathcal{C} \setminus j$ with probability $\propto$	$(\lvert c\rvert + \gamma)\, f_{\phi_c}(x_j)$	$\lvert c\rvert\, f_{\phi_c}(x_j)$
at a new table with probability $\propto$	$\gamma \frac{V_n(t+1)}{V_n(t)}\, f_{\phi_*}(x_j)$	$\alpha\, f_{\phi_*}(x_j)$

   where $t = \lvert\mathcal{C} \setminus j\rvert$ is the number of occupied tables, excluding j.

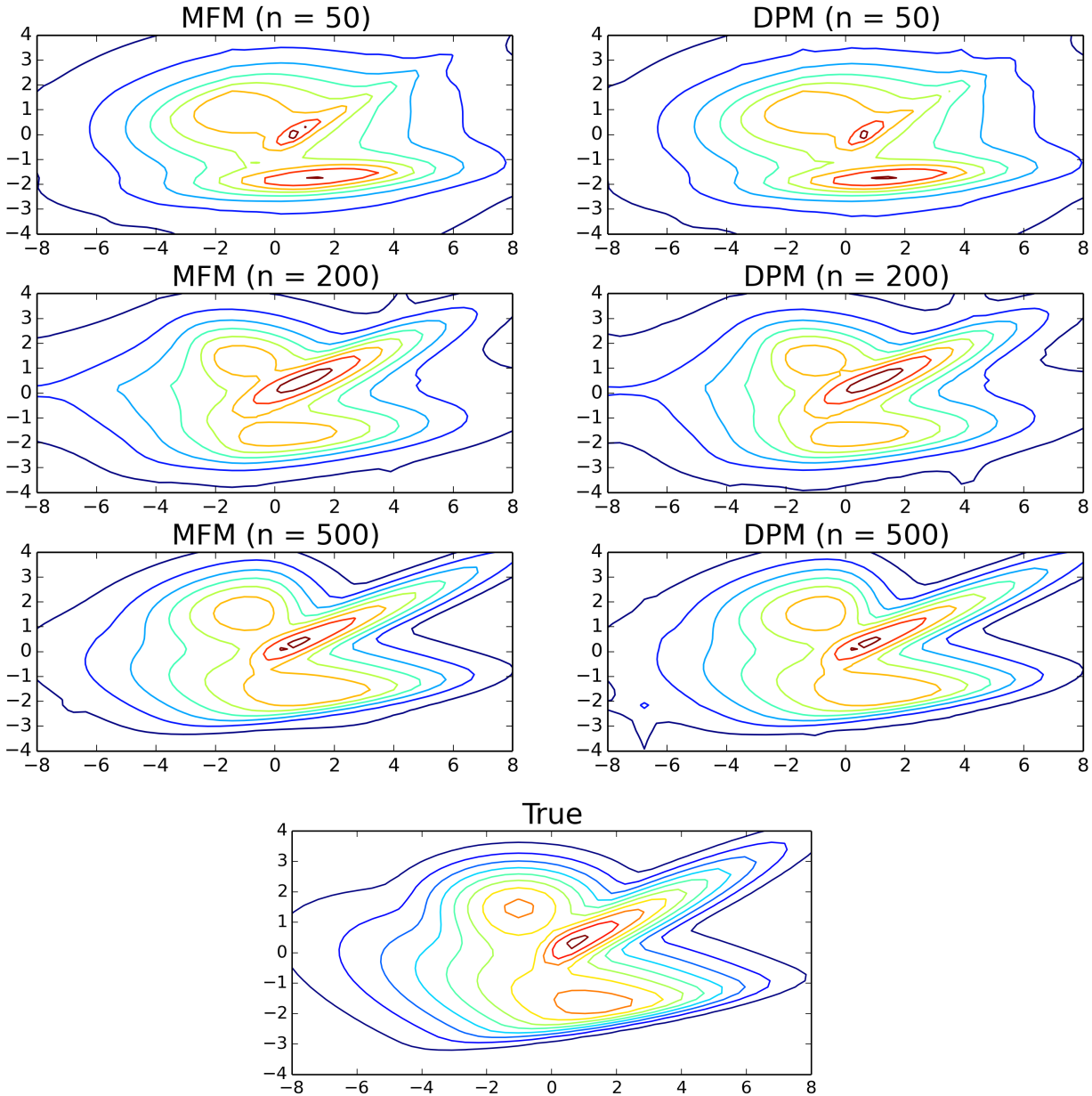
2. For each $c \in \mathcal{C}$, sample $\phi_c \sim p(\phi_c | x_c, \mathcal{C})$, or move ϕ_c according to a Markov chain that converges to this distribution.

Simulation example

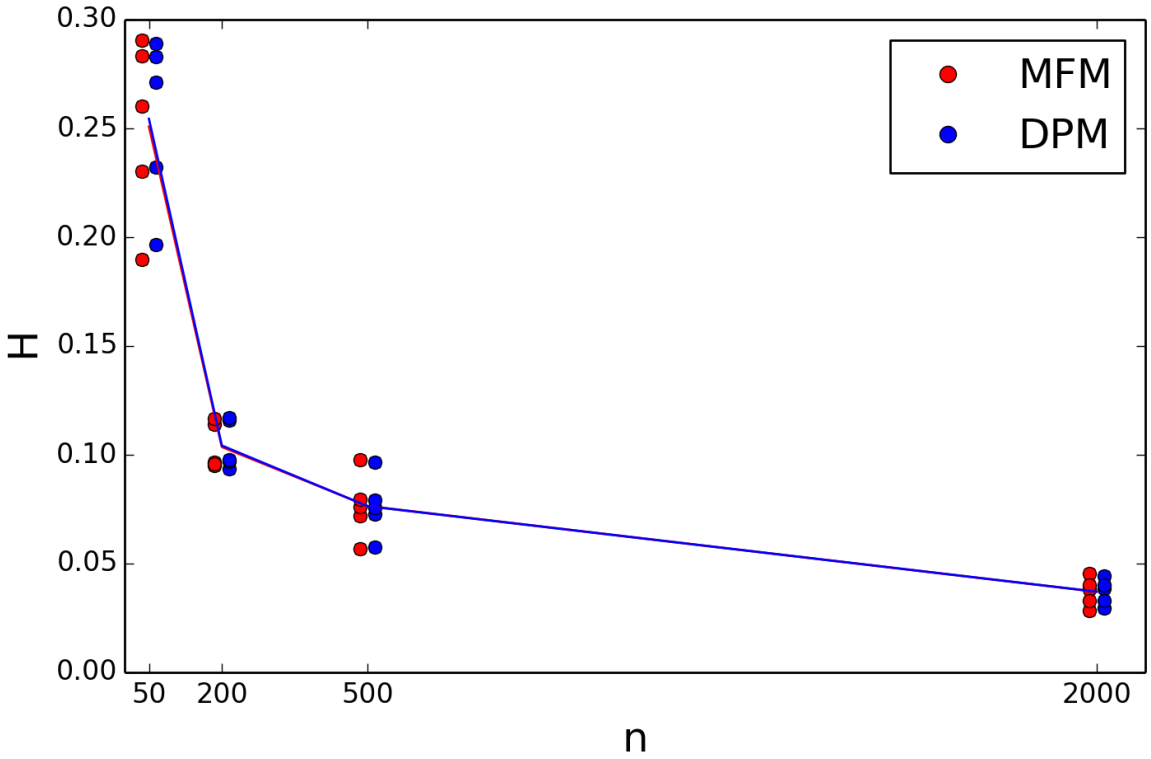
We compare the MFM and the DPM (mixing over the skew-normal family) on data from a bivariate skew-normal mixture with three components:



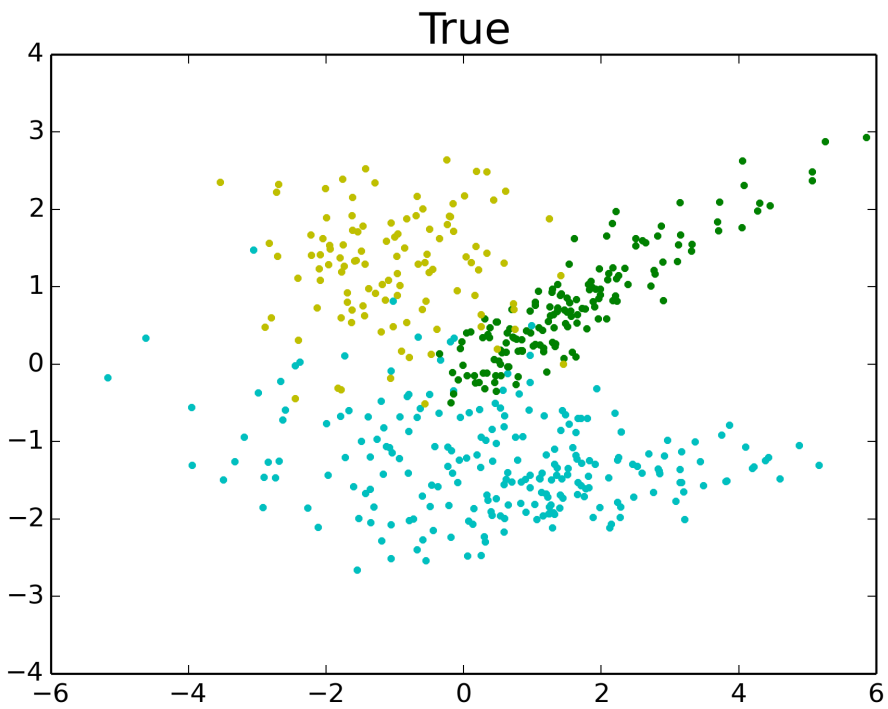
Estimated densities

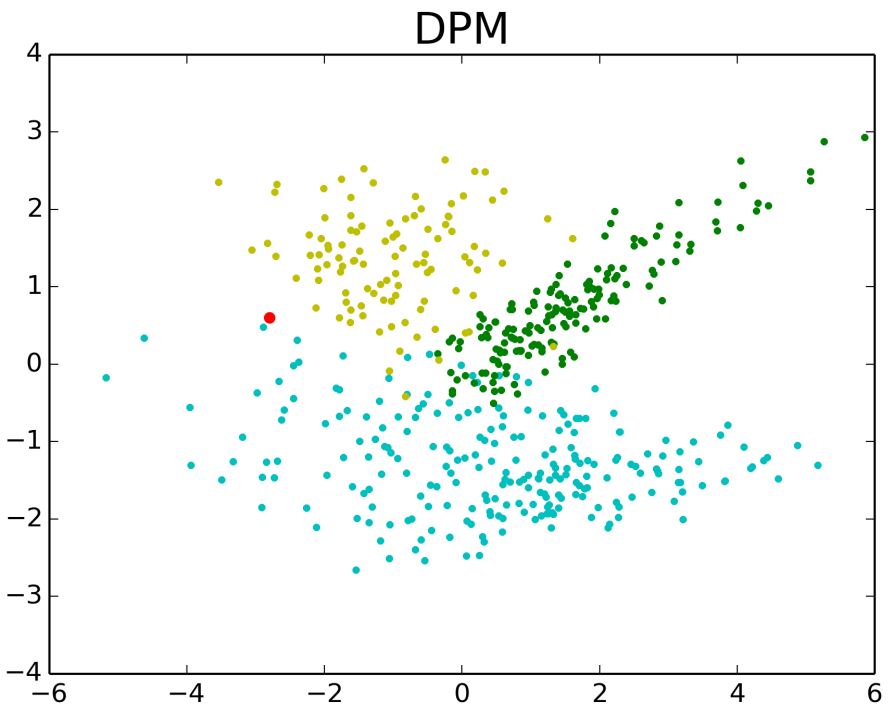


Hellinger distance to the true density

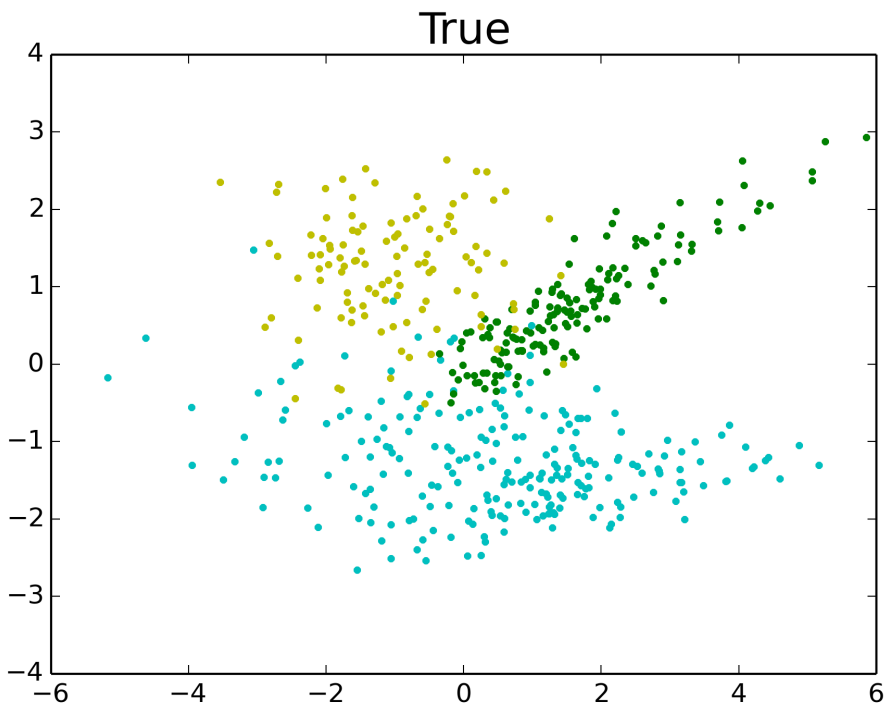


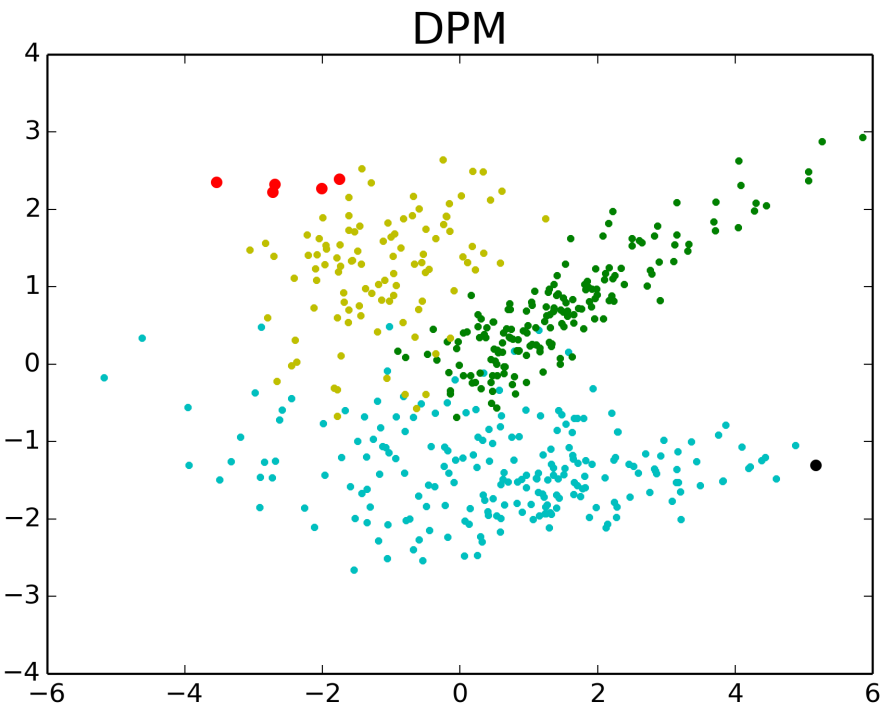
Sample clusterings from the posterior



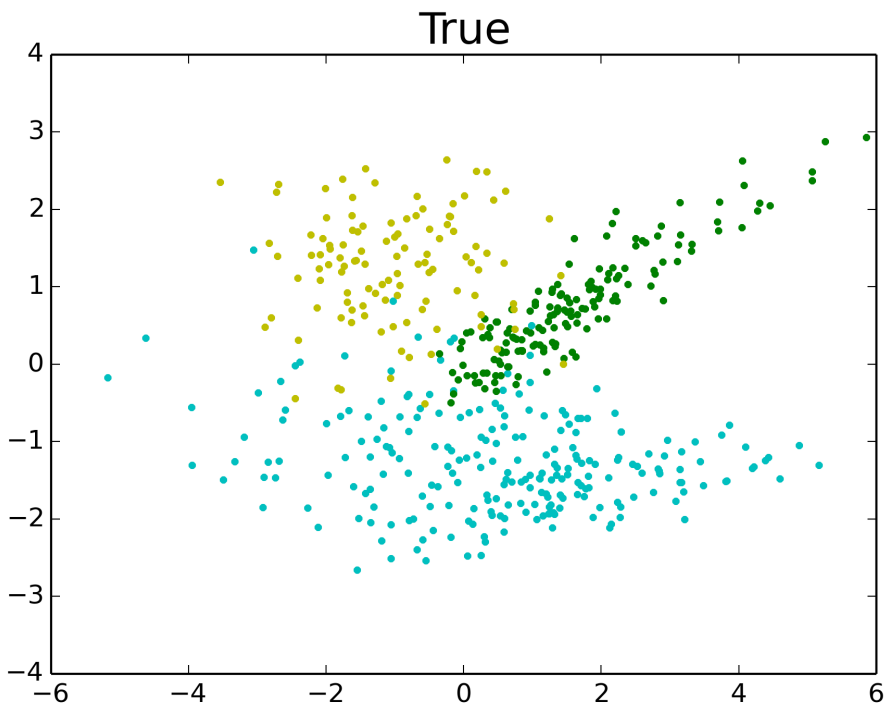


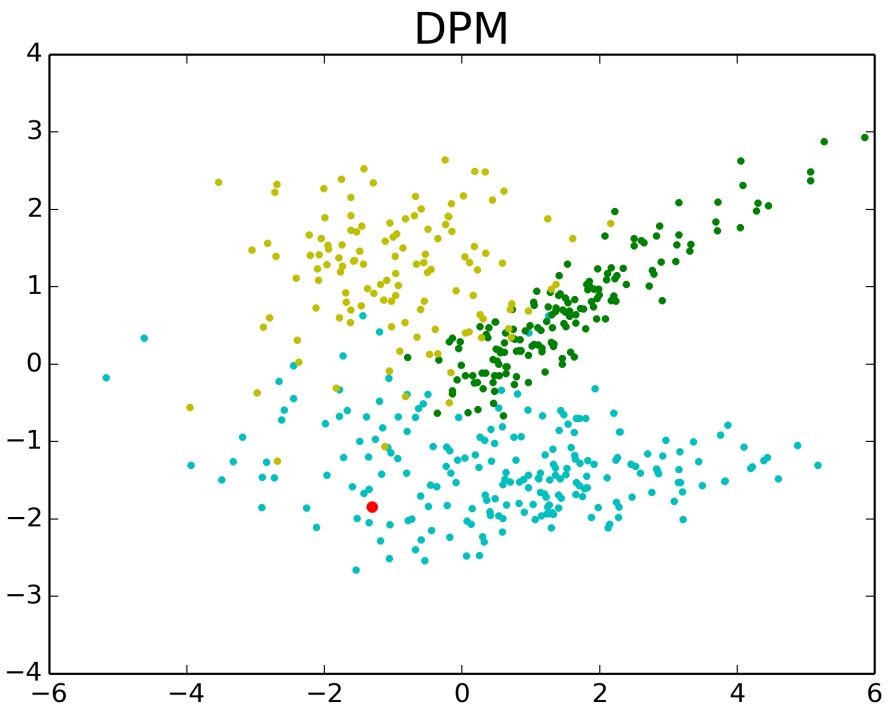
Sample clusterings from the posterior



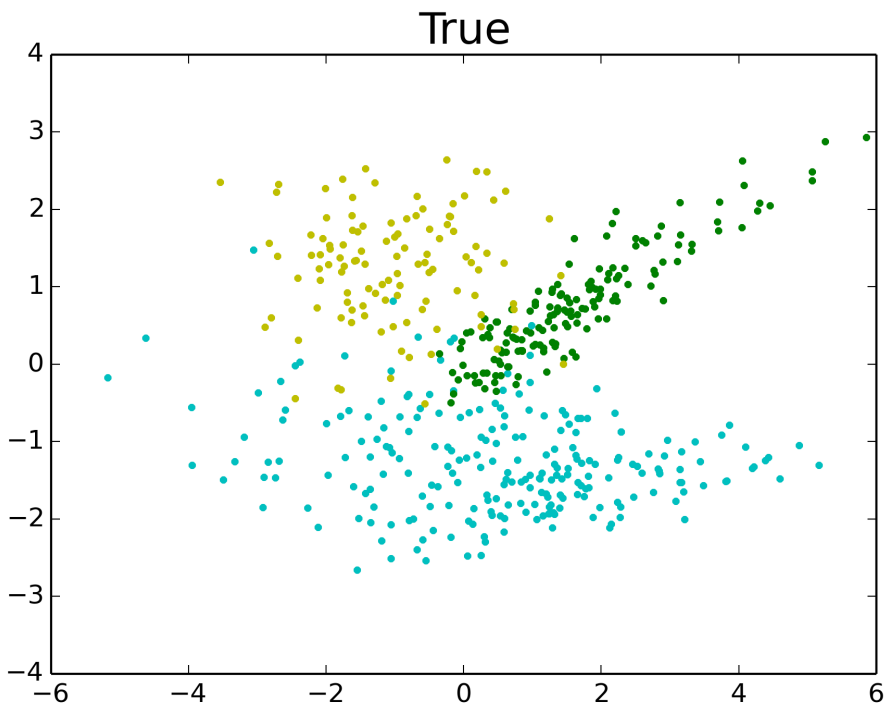


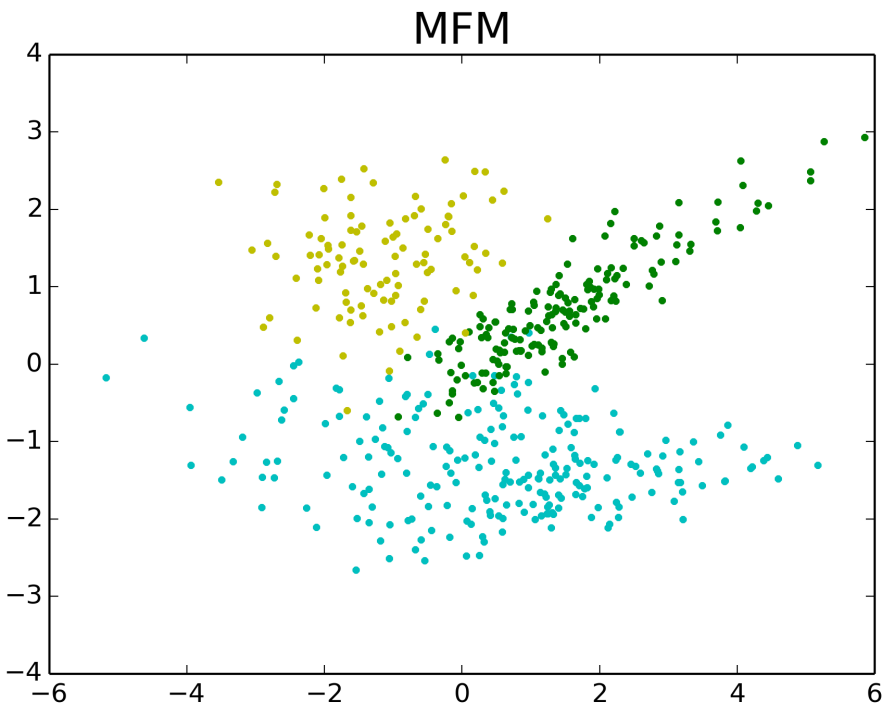
Sample clusterings from the posterior



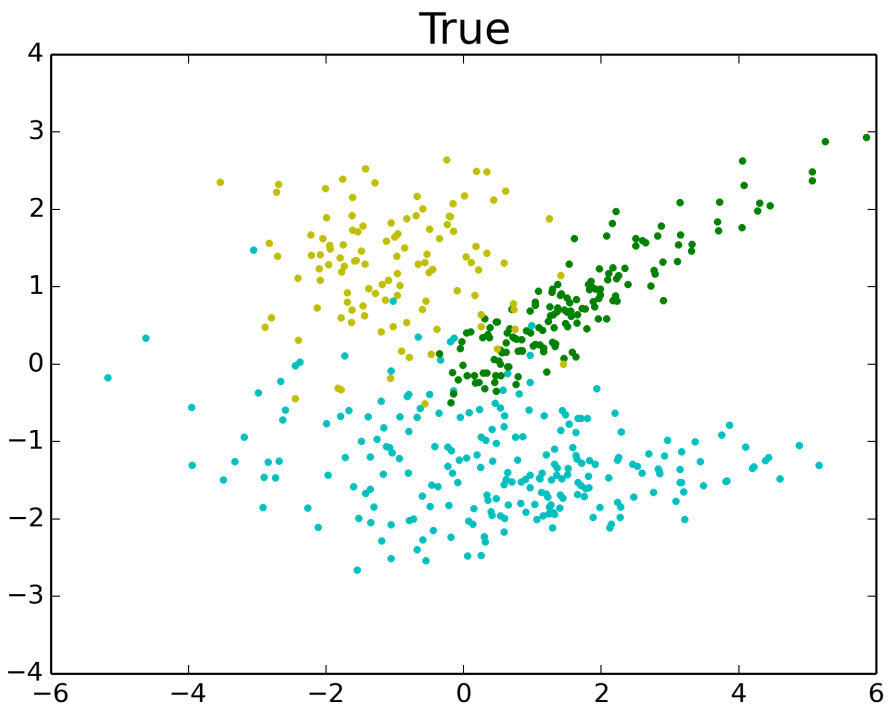


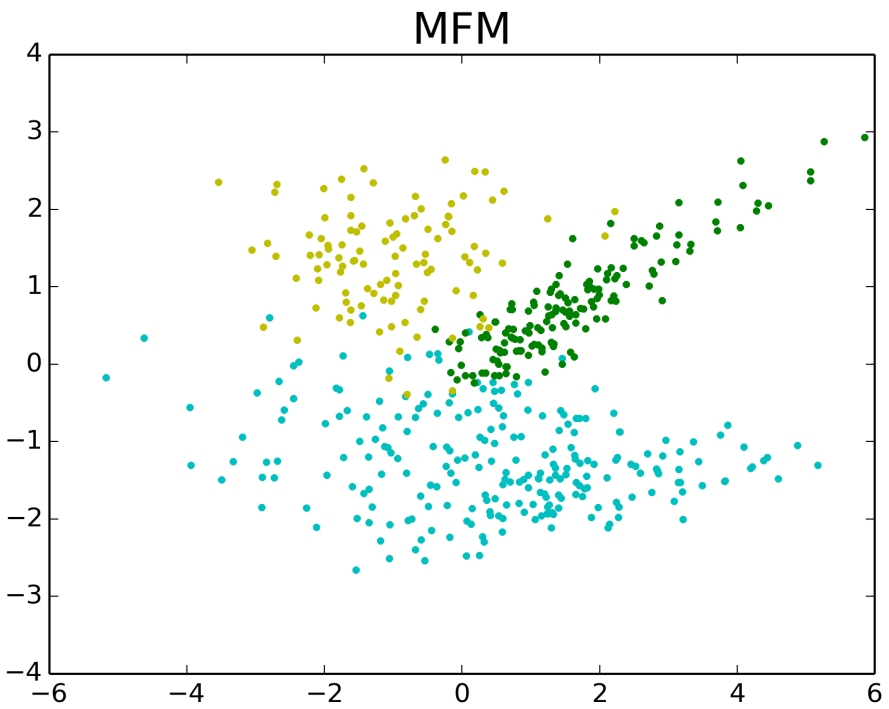
Sample clusterings from the posterior



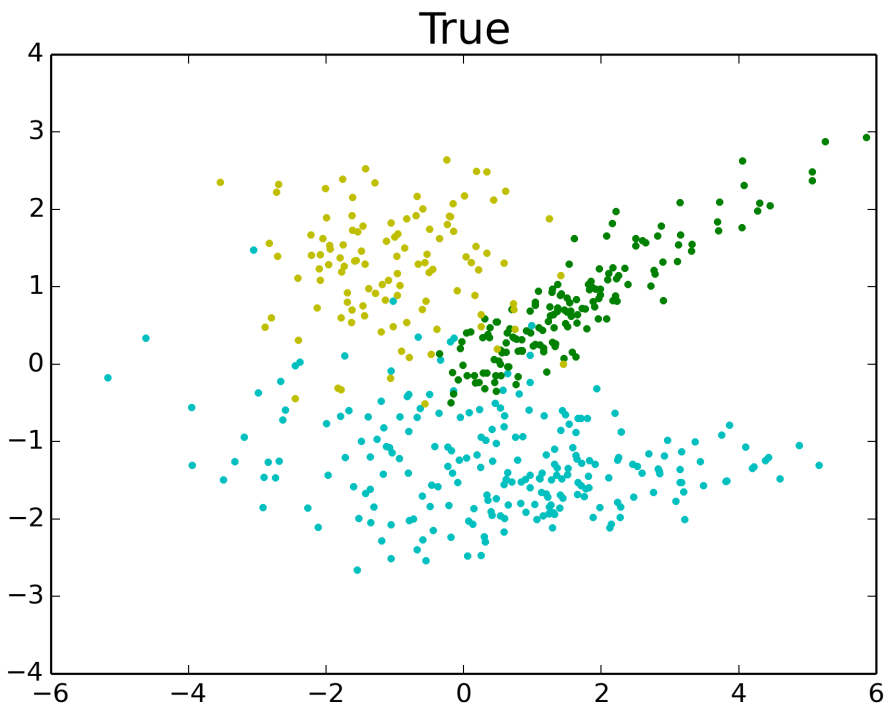


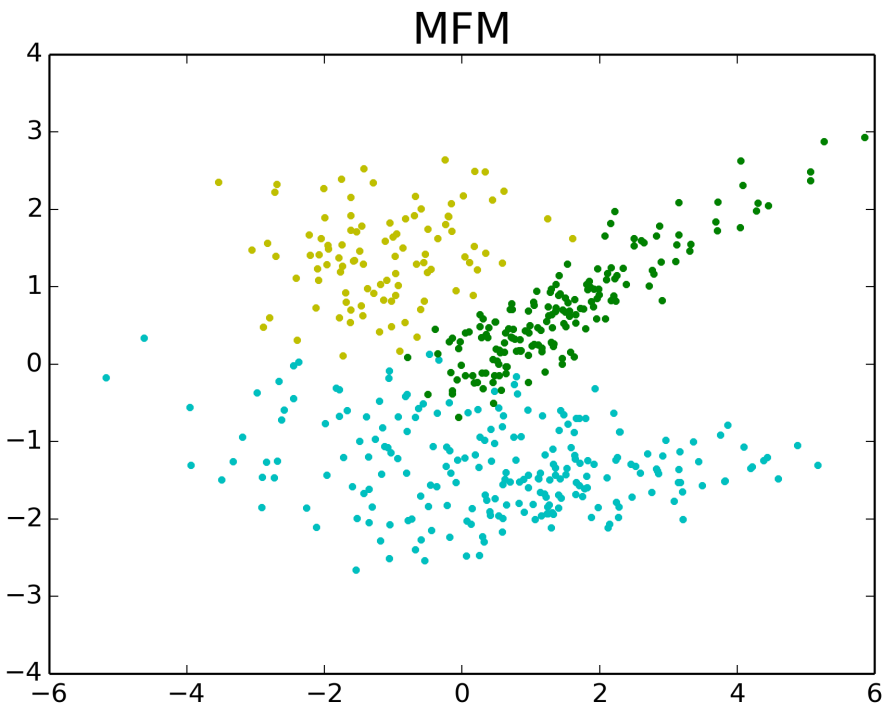
Sample clusterings from the posterior



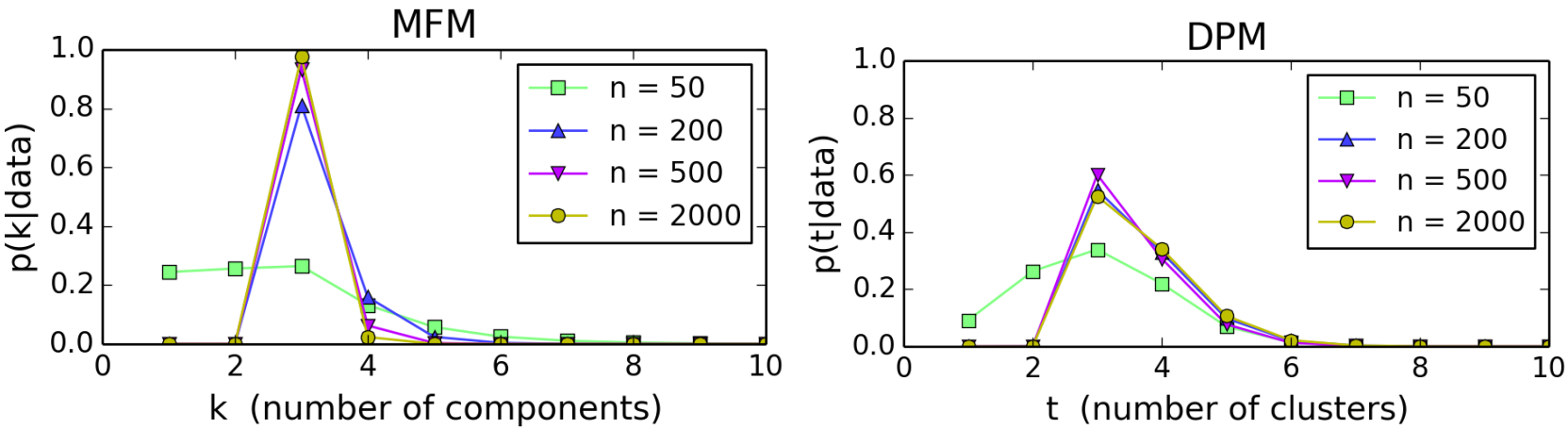


Sample clusterings from the posterior





Posteriors on the number of components and clusters



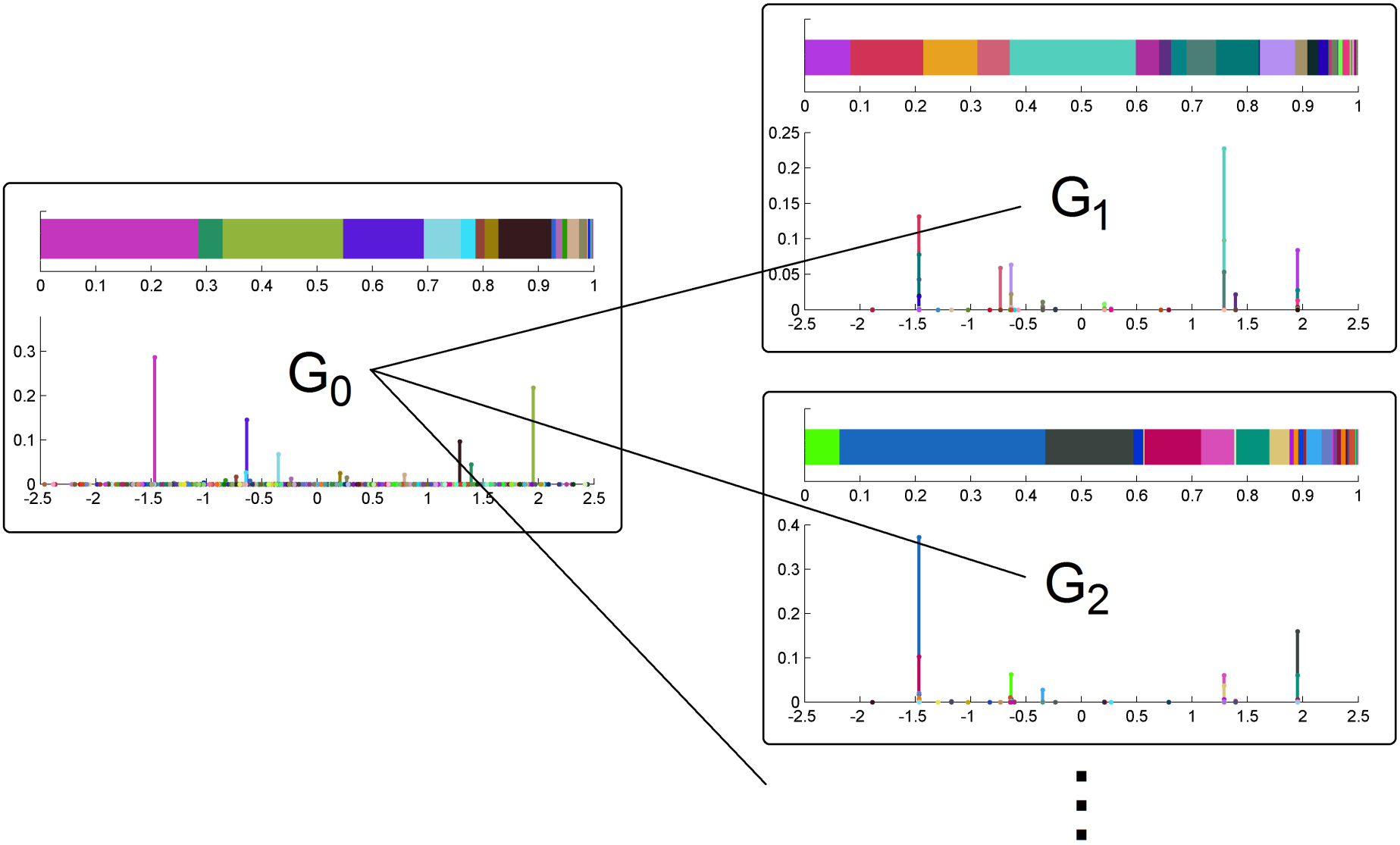
- The MFM appears to be concentrating at the true number of components, and the DPM does not.
- These results are the average over 5 runs.
- Note: These posteriors can be sensitive to H and sensitive to misspecification.

Outline

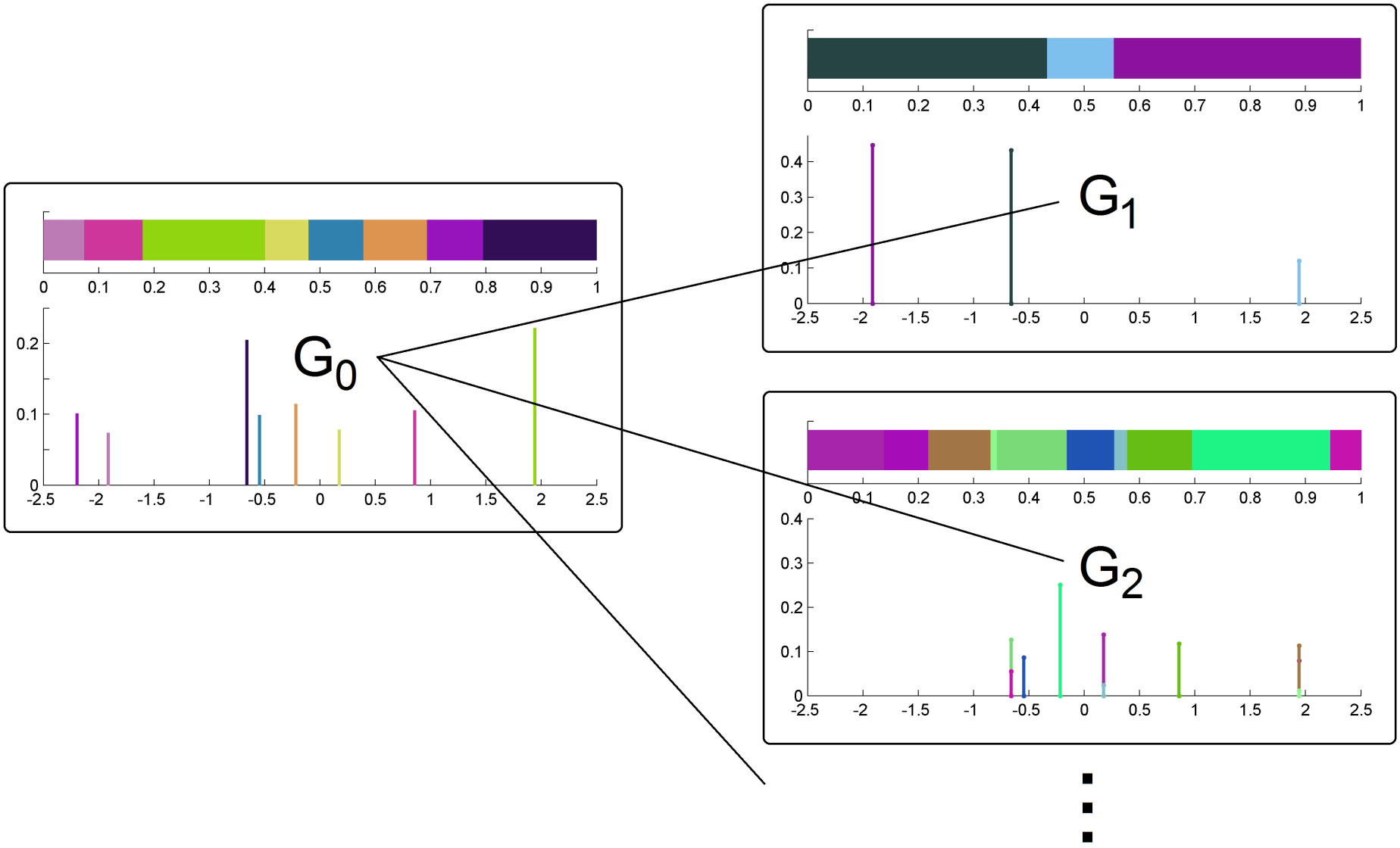
Motivation

- Suppose the data comes in m groups, e.g., words in m documents.
- Should we use an independent mixture model for each group?
- Should we use a single mixture model for all the data together?
- Idea: Allow mixture components to be shared among groups.
- This is the approach of the hierarchical Dirichlet process (HDP) of Teh et al. (2004), which has been used in many applications, including
  - document modeling (Teh et al., 2004),
  - natural language modeling (Liang et al., 2007),
  - object tracking (Fox et al., 2007),
  - haplotype inference (Xing et al., 2006),
  - natural image processing (Kivinen et al., 2007),
  - cognitive science (Griffiths et al., 2007), and
  - measuring similarity between musical pieces (Hoffman et al., 2008).

Hierarchical Dirichlet process (HDP) (Teh et al., 2004)



Hierarchical MFM (HMFM)



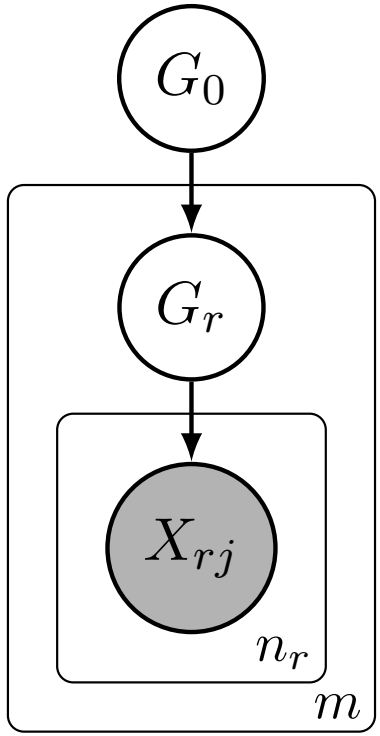
Hierarchical MFM (HMFM)

$$G_0 \sim \mathcal{M}(q_0, \gamma_0, H)$$

For $r = 1, \ldots, m$ independently, given G_0,

$$G_r \sim \mathcal{M}(q_r, \gamma_r, G_0)$$

$$X_{rj} \sim f_{G_r}(x) \text{ independently for } j = 1, \ldots, n_r.$$



- In the HDP, the $\mathcal{M}$ distributions are replaced by Dirichlet processes.
- This has the same exchangeability properties as the HDP:
  1. the datapoints within each group are exchangeable, and
  2. if $q_1 = \cdots = q_m$ and $\gamma_1 = \cdots = \gamma_m$, then the order of the groups does not matter (as long as the group sizes $n_1, \ldots, n_m$ are handled appropriately).

Franchise process

- In the same way as the HDP, this can be represented by a "franchise process".
- We envision the m groups of datapoints as groups of customers visiting m restaurants, respectively, and suppose each occupied table is served a single dish, chosen from a franchise-wide menu.
- For groups $r = 1, \ldots, m$, let $V_n^r(t)$ denote the MFM coefficient using the group-specific γ_r and q_r, and let $V_n^0(t)$ denote the MFM coefficient using γ_0 and q_0.

Franchise processes for HMFM and HDP

For each restaurant $r = 1, \ldots, m$:

For $j = 1, \ldots, n_r$:

1. The jth customer at restaurant r sits ...

	HMFM	HDP
at existing table c with probability $\propto$	$\lvert c \rvert + \gamma_r$	$\lvert c \rvert$
or at a new table with probability $\propto$	$\gamma_r \dfrac{V_j^r(t_r + 1)}{V_j^r(t_r)}$	α_r

where t_r is the number of tables occupied so far at restaurant r.

2. If a new table is chosen, serve it ...

	HMFM	HDP
an existing dish d with probability $\propto$	$\lvert d \rvert + \gamma_0$	$\lvert d \rvert$
or a new dish with probability $\propto$	$\gamma_0 \dfrac{V_L^0(t_0 + 1)}{V_L^0(t_0)}$	α_0

where $\lvert d \rvert$ is the number of tables with dish d so far, t_0 is the number of dishes tried so far in all restaurants, and L is the number of tables occupied so far at all restaurants, including the new table.

Hierarchical partition distribution

- At restaurant r, let $\mathcal{C}_r$ be the partition of the n_r customers according to table.
- Let $\mathcal{C}_0$ be the partition of all $L = \sum_{r=1}^{m} |\mathcal{C}_r|$ tables according to dish.
- Under the HMFM franchise process above,

$$p(\mathcal{C}_{0:m}) = P_L^0(\mathcal{C}_0) \prod_{r=1}^{m} P_{n_r}^r(\mathcal{C}_r)$$

  where P_n^r is the MFM partition distribution using coefficient $V_n^r(t)$, for $r = 0, 1, \ldots, m$.
- The formula for the HDP is the same, except that P_n^r is replaced by the corresponding DPM partition distribution.

Equivalent representation of the HMFM model

As in the HDP, the following partition-based model is equivalent to the discrete measure-based model:

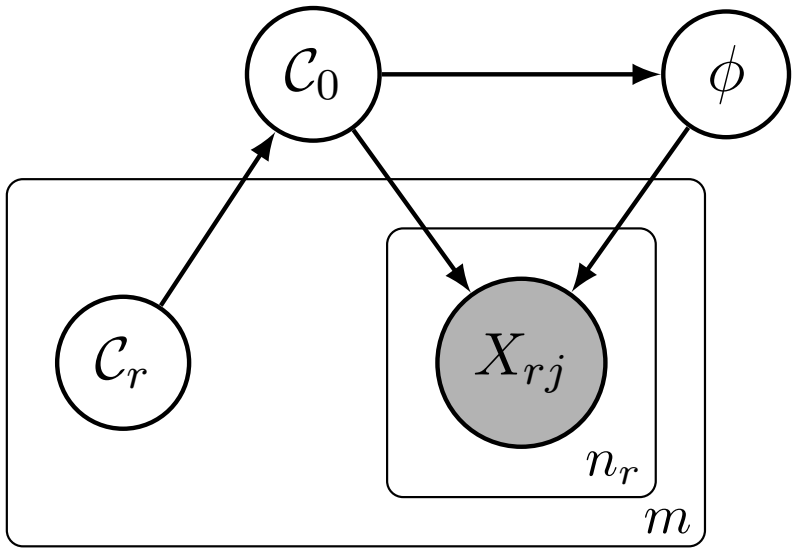
$\mathcal{C}_r \sim P^r_{n_r}$ independently for $r = 1, \ldots, m$

$\mathcal{C}_0 \sim P^0_L$ given $\mathcal{C}_{1:m}$, where $L = \sum_{r=1}^m |\mathcal{C}_r|$

For $d \in \mathcal{C}_0$ independently, given $\mathcal{C}_0$,

$\phi_d \sim H$

$X_{rj} \sim f_{\phi_d}$ independently for $(r, j) \in c$, $c \in d$.



Inference for the HMFM

- Using this representation, the Gibbs sampling algorithm for the HDP is easily adapted to the HMFM.
- It might be interesting to try to adapt other HDP inference algorithms as well, such as the augmented sampler of Teh et al. (2006).

Outline

Latent binary feature models

- A mixture model can be viewed as having a single latent feature.
- Often, it is more realistic to allow objects to have multiple features; this can be encoded in a binary matrix such that $Z_{ij} = 1$ when object i has feature j.
- Given such a matrix Z, the observed data X may be modeled in a variety of ways. A simple example would be

$$X_i = \sum_j Z_{ij}\mu_j + \varepsilon_i$$

  where the μ's and ε's are normal.
- The Indian buffet process (IBP) of Griffiths & Ghahramani (2005) is a nonparametric model for Z with infinitely many features. The IBP has seen a number of applications, including models for
  - protein complexes (Chu et al., 2006),
  - gene expression data (Knowles et al., 2007),
  - causal graph structure (Wood et al., 2006),
  - similarity judgments (Navarro and Griffiths, 2007),
  - network data (Miller et al., 2009), and
  - multiple time-series (Fox et al., 2009).

Latent binary feature models

- The IBP is an infinite-dimensional limit of a family of finite feature models.
- By instead placing a prior on the number of features, we obtain a distribution with many of the same attractive properties as the IBP:
  - an exchangeable distribution on equivalence classes of binary matrices,
  - representation via a simple buffet process, and
  - approximate inference via the same Gibbs sampling algorithms.
- Further, this model allows for more control over the distribution.
- In fact, in a certain sense the IBP (as well as the two-parameter generalization of Ghahramani et al. (2007)) can be viewed as a special case.

A distribution on binary matrices

Sample $K \sim p(k)$, a p.m.f. on $\{0, 1, 2, \dots\}$.

π_1	π_2	π_3	$\cdots$	π_K
0	1			
1	0			
1	1	$\cdots$	$Z_{ij} \sim \mathrm{B}(\pi_j)$	
$\vdots$	$\vdots$		given π_j	
0	0			
1	0			

$\sim \mathrm{Beta}(a, b)$, given K

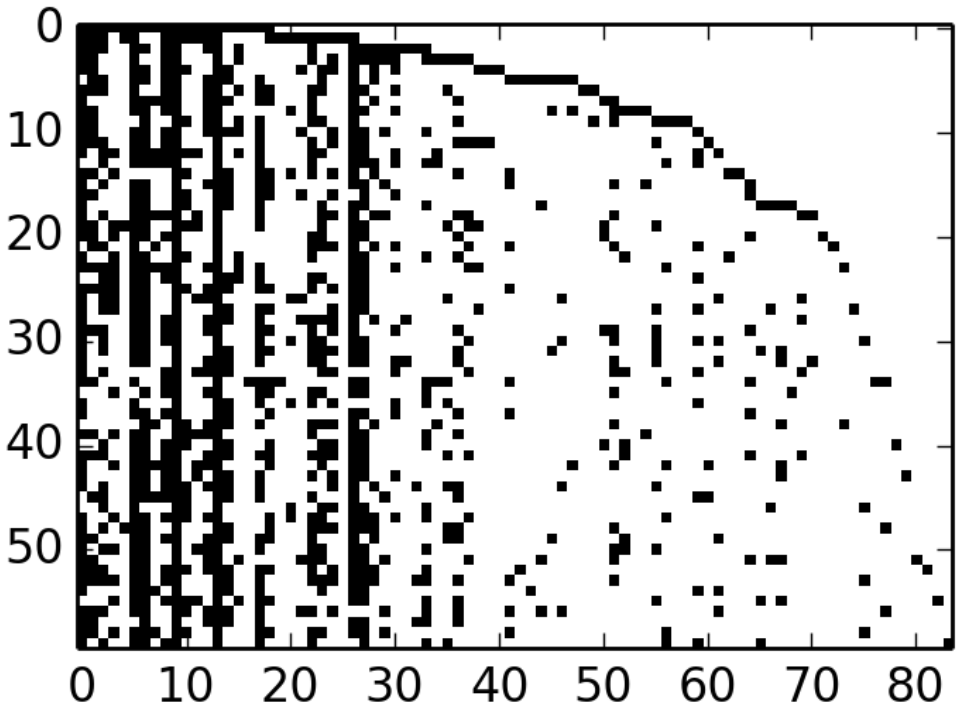
This is simply the finite feature model of Griffiths and Ghahramani (2005), with a prior on the number of features, K.

Staircase form

For $z \in \{0,1\}^{n \times k}$, define

$$z' = \begin{bmatrix} B_1 & B_2 & \cdots & B_n \end{bmatrix}$$

where B_m is the submatrix of columns j such that $m = \min\{i : z_{ij} = 1\}$. We refer to this as *staircase form*.



Buffet processes for MFFM, IBP, and 2-param IBP

When $p(k) = \text{Poisson}(k|\lambda)$, the distribution of Z' can be described by the following buffet process, closely resembling the IBP.

For $m = 1, \ldots, n$: Customer m ...

1. tries each previously-tried dish j with probability

MFFM	IBP	2-param IBP
$\frac{a + s_j}{a + b + m - 1}$	$\frac{s_j}{m}$	$\frac{s_j}{\beta + m - 1}$

where s_j is the number of previous customers trying dish j, and

2. tries a $\text{Poisson}(\lambda_m)$ number of new dishes, where λ_m equals

MFFM	IBP	2-param IBP
$\lambda \frac{ab^{(m-1)}}{(a+b)^{(m)}}$	$\frac{\alpha}{m}$	$\frac{\alpha\beta}{\beta + m - 1}$

The IBP and two-parameter IBP can be viewed as limiting cases.

Distribution on staircase matrices

- This buffet process generates a binary matrix z' in staircase form, where $z'_{mj} = 1$ if and only if customer m tries dish j (labeling the dishes in the order they are tried).
- The distribution on matrices z' generated by this process is the same as sampling from the MFFM and reducing to staircase form:

$$\mathbb{P}(Z' = z') = \frac{t!}{t_1! \cdots t_n!} v_n(t) \prod_{j=1}^{t} \frac{a^{(s_j)} b^{(n-s_j)}}{(a+b)^{(n)}}$$

  where t_m is the number of new dishes tried by customer m, $t = \sum t_m$, s_j is the number of customers trying dish j, and

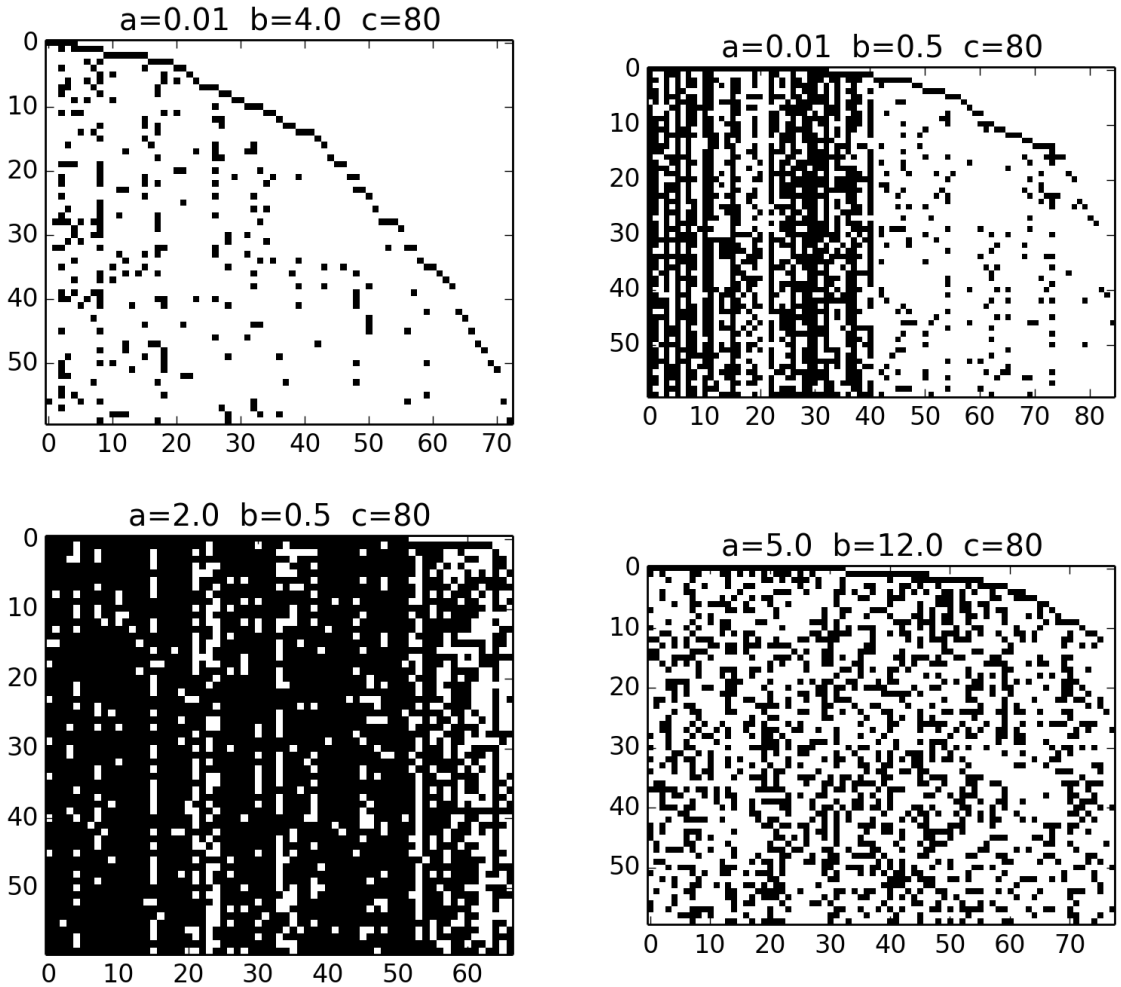
$$v_n(t) = \exp\left(\lambda \frac{b^{(n)}}{(a+b)^{(n)}}\right) \text{Poisson}(t|\lambda).$$

- For comparison, the corresponding distribution for the IBP is

$$\mathbb{P}_{\text{IBP}}(Z' = z') = \frac{t!}{t_1! \cdots t_n!} \exp(-\alpha H_n) \frac{\alpha^t}{t!} \prod_{j=1}^{t} \frac{(s_j - 1)!\,(n - s_j)!}{n!}.$$

Some sampled matrices

A diverse variety of matrices can be generated with the MFFM. We reparametrize to a, b, c where $c = \lambda\left(1 - \frac{b^{(n)}}{(a+b)^{(n)}}\right)$, so that the total number of dishes tried is a $\text{Poisson}(c)$ random variable.



Mixture of finite feature models (MFFM)

- In the same way as the IBP, exchangeability of the customers can be obtained by a slight modification.
- The Gibbs sampling algorithm for the IBP is easily adapted to the MFFM.
  - However, it is well-known that mixing can be very slow for the IBP, and the MFFM has the same issue.
- It might be interesting to try to adapt other IBP inference algorithms (e.g., Teh et al. (2007), Doshi-Velez et al. (2008)) to the MFFM.

Conclusion

- Variable-dimension models provide an interesting alternative to infinite-dimensional models.
- Many of the attractive properties of infinite-dimensional models can also be exhibited by their variable-dimension counterparts.
- This allows many of the same inference algorithms to be used.

Combinatorial stochastic processes for variable-dimension models

Jeffrey W. Miller

Joint work with Matt Harrison

(Research supported by DARPA and the NSF)

Duke University
Department of Statistical Science

Texas A&M Statistics Colloquium
Oct 31, 2014

Derivation of the MFM partition distribution

$$p(z|k) = \int p(z|\pi)p(\pi|k)d\pi = \frac{1}{(\gamma k)^{(n)}} \prod_{i=1}^{k} \gamma^{(n_i)}$$

$$\begin{aligned} p(\mathcal{C}|k) &= \sum_{z \in [k]^n : \mathcal{C}(z) = \mathcal{C}} p(z|k) \\ &= \#\Big\{ z \in [k]^n : \mathcal{C}(z) = \mathcal{C} \Big\} \frac{1}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \\ &= \frac{k_{(t)}}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \end{aligned}$$

$$p(\mathcal{C}) = \sum_{k=1}^{\infty} p(\mathcal{C}|k)p(k) = \Big(\prod_{c \in \mathcal{C}} \gamma^{(|c|)} \Big) \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} p(k) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)}$$

Density estimation

For $c \subset [n]$, define $m(x_c) = \int \Big(\prod_{j \in c} f_\theta(x_j) \Big) H(d\theta)$.

If $m(x_c)$ can be easily computed, the posterior predictive density can be estimated by

$$p(x_{n+1} \mid x_{1:n}) \approx \frac{1}{N} \sum_{i=1}^{N} p(x_{n+1} \mid \mathcal{C}^{(i)}, x_{1:n})$$

where $\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(N)}$ are samples from $\mathcal{C} \mid x_{1:n}$, and

$$p(x_{n+1} \mid \mathcal{C}, x_{1:n}) \propto \frac{V_{n+1}(t+1)}{V_{n+1}(t)} \gamma \, m(x_{n+1}) + \sum_{c \in \mathcal{C}} (|c| + \gamma) \frac{m(x_{c \cup \{n+1\}})}{m(x_c)}$$

where $t = |\mathcal{C}|$.

Density estimation

- Often, however, $m(x_c)$ cannot be easily computed (e.g., when a conjugate prior does not exist).
- In this case, the posterior predictive density can be approximated by assuming that $n+1$ is assigned to an existing cluster, and using samples from $\mathcal{C}, \phi \mid x_{1:n}$.
- This gives the density estimate

$$p(x_{n+1} \mid x_{1:n}) \approx \frac{1}{N} \sum_{i=1}^{N} p_*(x_{n+1} \mid \mathcal{C}^{(i)}, \phi^{(i)})$$

where $(\mathcal{C}^{(1)}, \phi^{(1)}), \ldots, (\mathcal{C}^{(N)}, \phi^{(N)})$ are samples from $\mathcal{C}, \phi \mid x_{1:n}$, and

$$p_*(x_{n+1} \mid \mathcal{C}, \phi) = \sum_{c \in \mathcal{C}} \frac{|c| + \gamma}{n + \gamma t} f_{\phi_c}(x_{n+1})$$

where $t = |\mathcal{C}|$.

Posterior asymptotics

Given a prior on the mixing measure G, consider the following questions.

1. *Density estimation.* Does the posterior on the density concentrate at the true density f_*? That is, does

$$\mathbb{P}_{\text{model}}(\text{dist}(f_G, f_*) < \varepsilon \mid X_{1:n}) \xrightarrow[n\to\infty]{\mathbb{P}_{\text{data}}} 1$$

   for all $\varepsilon > 0$? If so, at what rate of convergence?

2. *Mixing measure.* Does the posterior on the mixing measure concentrate at the true mixing measure G_* (assuming there is one)? That is, does

$$\mathbb{P}_{\text{model}}(\text{dist}(G, G_*) < \varepsilon \mid X_{1:n}) \xrightarrow[n\to\infty]{\mathbb{P}_{\text{data}}} 1$$

   for all $\varepsilon > 0$?

3. *Number of components.* Does the posterior on the number of components concentrate at the true number of components, for data from a finite mixture? Does the posterior on the number of *clusters* concentrate at the true number of *components*?

Skew-Normal distribution

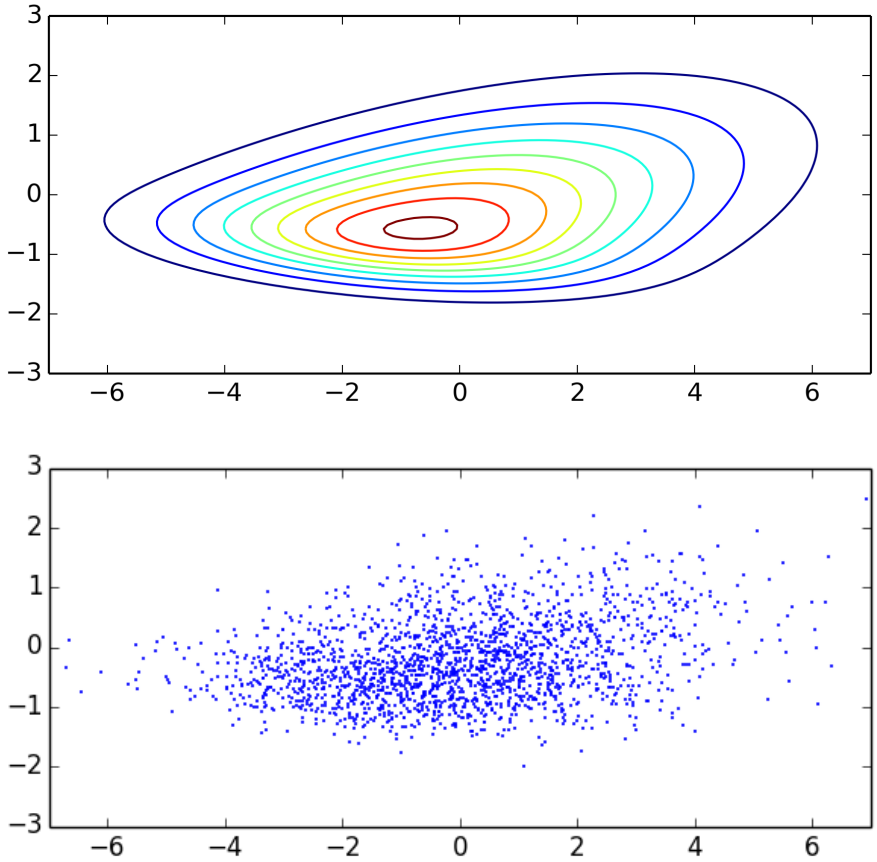
- To make things interesting, we will use multivariate Skew-Normal mixtures. (The results are similar for other families.)
- Azzalini & Dalla Valle (1996) (see also Azzalini & Capitanio (1999)) introduced the multivariate Skew-Normal distribution, with density

$$\mathcal{SN}(x \mid \xi, Q, w) = 2\,\mathcal{N}(x \mid \xi, Q)\ \Phi(w^{\mathsf{T}} S^{-1}(x - \xi))$$

  for $x \in \mathbb{R}^d$, where S is diagonal with $S_{ii} = \sqrt{Q_{ii}}$, and Φ is the univariate standard normal CDF. The parameters are:
  - $\xi \in \mathbb{R}^d$ (location),
  - Q positive definite (scale and correlation),
  - $w \in \mathbb{R}^d$ (skew).
- This family has some nice properties (e.g., preserved under linear maps).

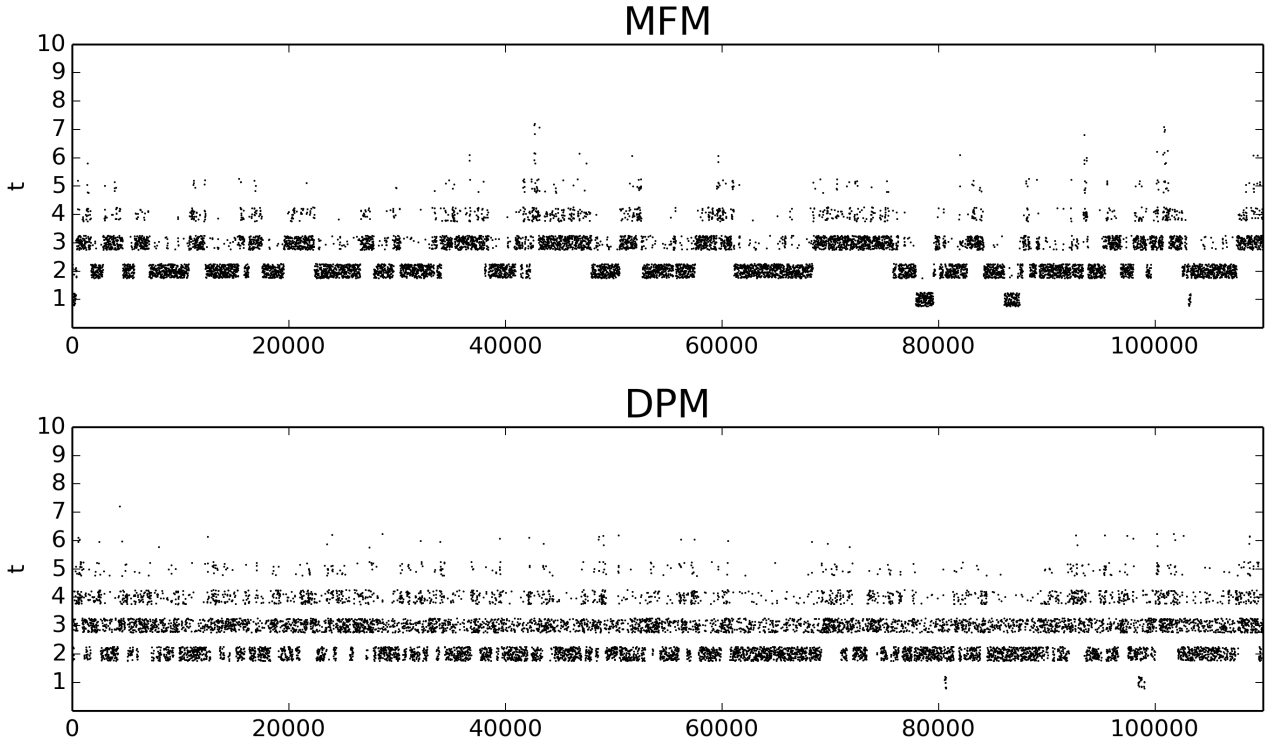
Skew-Normal distribution

Basically, it's a multivariate normal which has been skewed in the direction of $w/|w|$, by magnitude $|w|$, according to the skew parameter $w \in \mathbb{R}^d$.



Mixing issues with incremental Gibbs

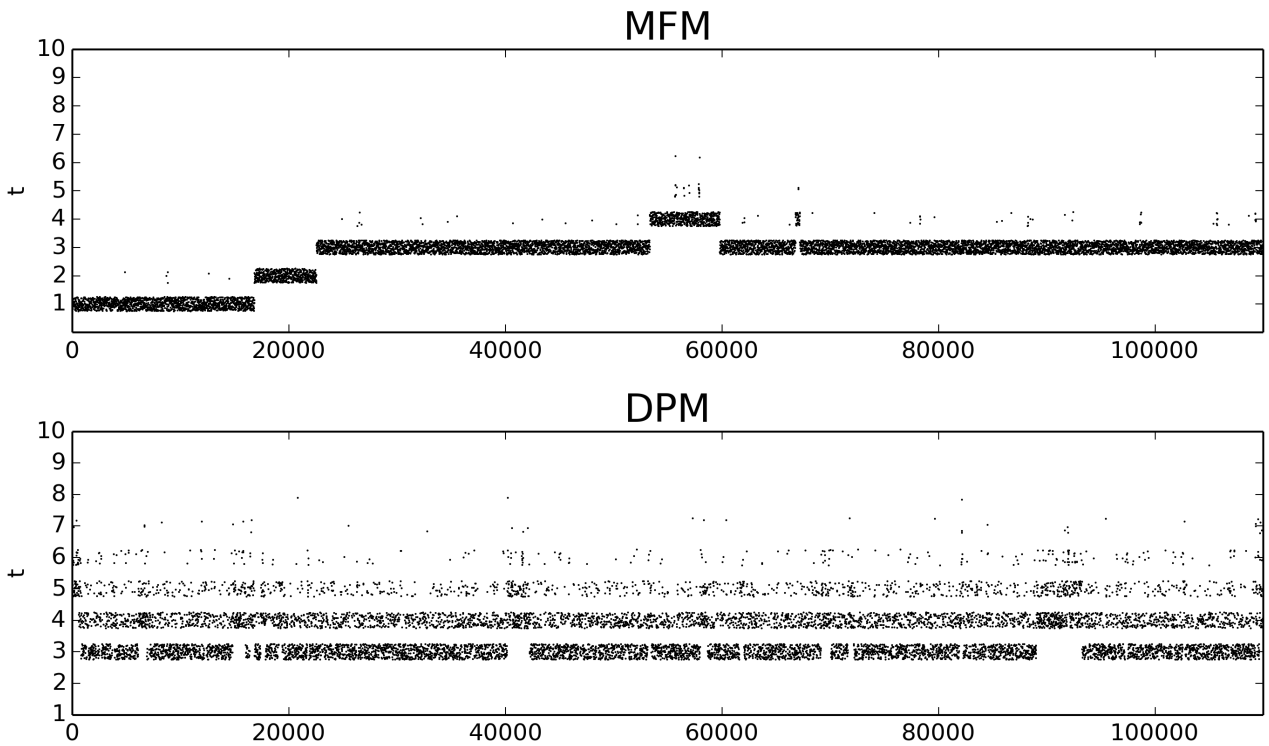
Traceplot of the number of clusters t, with $n = 50$



For smallish n, MFM mixing seems somewhat worse than the DPM.

Mixing issues with incremental Gibbs

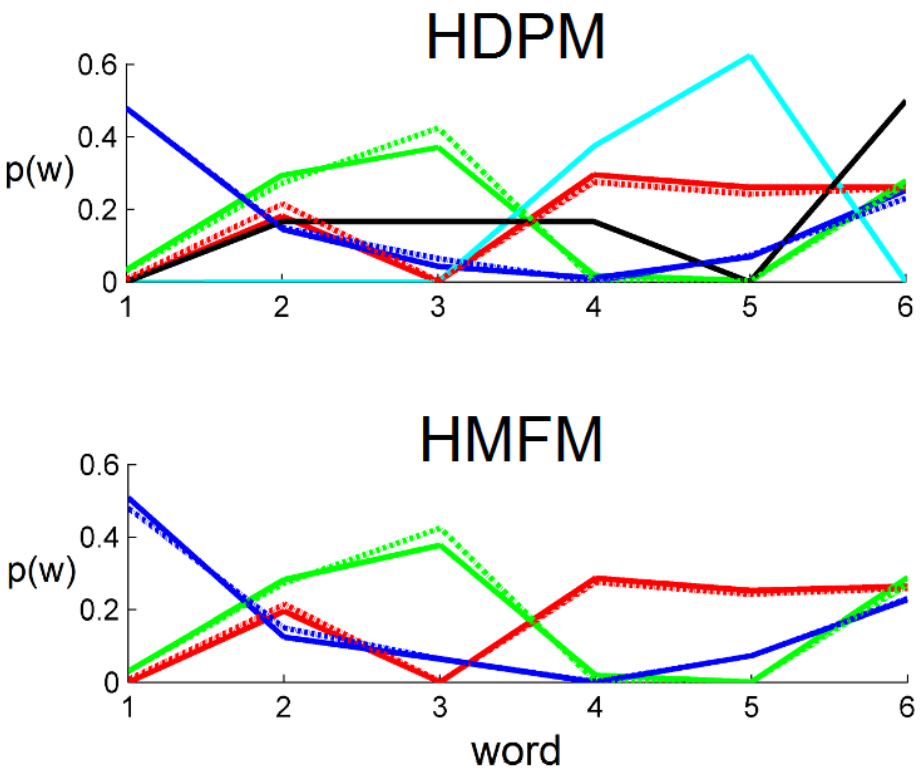
Traceplot of the number of clusters t, with $n = 2000$



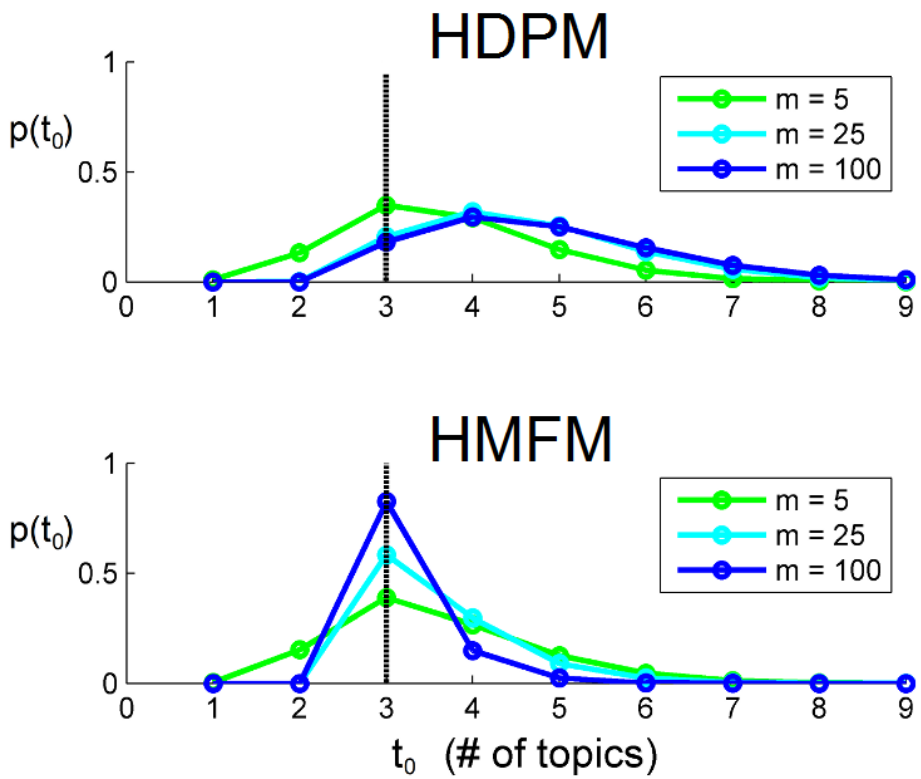
For larger n, the issue becomes worse. The MFM doesn't like having small clusters, so it's difficult to make or destroy substantial clusters by moving one point at a time.

Preliminary HMFM results with a toy topic model

Typical posterior topic distributions



Posterior on # of topics



Preliminary MFFM results with a toy feature model