Combinatorial stochastic processes for variable-dimension models

Jeffrey W. Miller

Joint work with Matt Harrison

Brown University
Division of Applied Mathematics

Duke Statistical Science Seminar
Feb 7, 2014

Nonparametric Bayesian models have found many applications . . .

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

. . . . . .

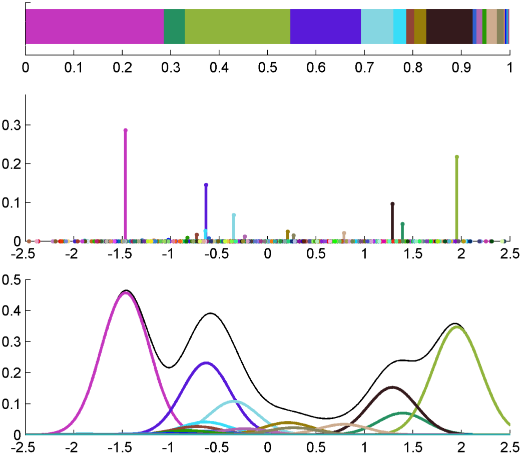
- econometrics
- phylogenetics
- species delimitation
- computer vision
- classification
- document modeling
- cognitive science
- natural language processing

. . . . . .

Variable-dimension models

- Many nonparametric models arise as the 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, put a prior on the number of components in a finite mixture.

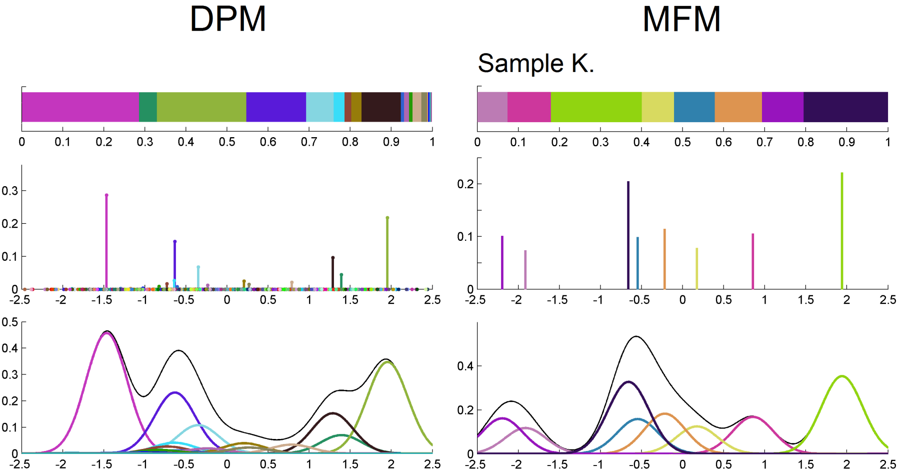
Dirichlet process mixture (DPM)



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

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

Mixture of finite mixtures (MFM)



DPM

MFM

Sample K.

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

Why use a variable-dimension model?

- Control over the distribution of the number of clusters/topics/features
- Control over the distribution of the relative sizes of clusters/topics
- Cleaner clusters/topics/features (no tendency to make tiny superfluous groups)
- Interpretability and conceptual simplicity
- Natural Bayesian approach for a data distribution of unknown complexity (if something is unknown, put a prior on it)
- Theory can be much simpler, since the parameter space is a countable union of finite-dimensional spaces, rather than an infinite-dimensional space.
    - For example, consistency typically holds "automatically" at Lebesgue almost-all parameters, by Doob's theorem (assuming identifiability).

There are some disadvantages also, as we will see.

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 DMA 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 DMA models . . .* "

- The key to such samplers is that the model can be characterized by a nice distribution on combinatorial structures (e.g., the CRP).

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 (compare with DPM)
  - hierarchical mixture of finite mixtures (compare with HDP)
  - mixture of finite feature models (compare with 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)
   - Basic properties
   - Asymptotics
   - Inference algorithms (conjugate and non-conjugate cases)
2. Empirical comparison of MFMs with DPMs
   - Density estimation
   - Clustering
   - Mixing issues
   - Bayes factors
   - Outliers
3. Hierarchical mixture of finite mixtures (compare with HDP)
4. Mixture of finite feature models (compare with IBP)

Outline

1. Mixture of finite mixtures (MFM)
   - Basic properties
   - Asymptotics
   - Inference algorithms (conjugate and non-conjugate cases)

2. Empirical comparison of MFMs with DPMs
   - Density estimation
   - Clustering
   - Mixing issues
   - Bayes factors
   - Outliers

3. Hierarchical mixture of finite mixtures (compare with HDP)

4. Mixture of finite feature models (compare with IBP)

Mixture of finite mixtures (MFM)

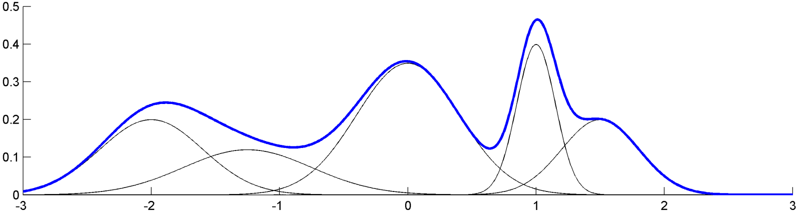
$K \sim p(k)$, a p.m.f. on $\{1, 2, \ldots\}$

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

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

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

$X_j \sim p_{\theta_{Z_j}}$ for $j = 1, \ldots, n$, given θ, Z.



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

$$p(\mathcal{C}) = v_n(t) \prod_{i=1}^{t} \gamma^{(n_i)}$$

where

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

$n_1, \ldots, n_t$ are the sizes of the parts,

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

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

Derivation of the partition distribution

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

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

$$p(\mathcal{C}) = \sum_{k=1}^{\infty} p(\mathcal{C}|k)p(k) = \Big(\prod_{i=1}^{t} \gamma^{(n_i)}\Big) \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} p(k) = v_n(t) \prod_{i=1}^{t} \gamma^{(n_i)}$$

Some properties of $v_n(t)$

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

- For any $1 \leq t \leq n$, these numbers satisfy the recursion:

$$v_{n+1}(t+1) = v_n(t)/\gamma - (n/\gamma + t) v_{n+1}(t).$$

- If $p(k) = \mathrm{Poisson}(k-1|\lambda)$ then $v_n(0) = \lambda^{-n}(1 - \sum_{k=1}^{n} p(k))$.
- Note that $k_{(t)}/(\gamma k)^{(n)} \leq k^t/(\gamma k)^n$, so the infinite series usually converges rapidly. (It always converges for $1 \leq t \leq n$.)

See Gnedin & Pitman (2006) for more general results.

Computing $v_n(t)$

To compute $p(\mathcal{C})$, we need to evaluate $v_n(t)$.

Two methods:

1. Numerical approximation (my preferred method)
   - Easy, fast, and generally applicable.
   - In practice, the series can quickly be computed to within machine precision.
   - Note: The log-sum-exp trick must be used to avoid numerical underflow.
2. If there is a simple expression for $v_n(0)$ (e.g. in the Poisson case), use the recursion above.

Using either method, we can precompute $v_n(t)$ for each n, t that will be needed — typically, n will be fixed and we only need $t = 1, \ldots, t_{\max}$ for some relatively small $t_{\max}$.

Marginalization/self-consistency property

- For each $n = 1, 2, \ldots$, let $p_n(\mathcal{C})$ denote the distribution on partitions of $[n]$ defined above.
- This family of partition distributions has the property that 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 easily derived from the recursion for $v_n(t)$.

Restaurant process

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

Restaurant process for MFM

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

The nth customer sits …

at table $c \in \mathcal{C}$ with probability $\propto |c| + \gamma$

at a new table with probability $\propto \gamma\, v_n(t+1)/v_n(t)$

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

Clearly, this bears a close resemblance to the Chinese restaurant process.

Random discrete measures

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

Let

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

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

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

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

Stick-breaking representation in a special case

When $p(k) = \text{Poisson}(k-1|\lambda)$ and $\gamma = 1$, there is a nice stick-breaking representation for the weights $\pi_1, \dots, \pi_K$:

> **Start with a unit-length stick, and**
> **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.

This suggests a stick-breaking approach to constructing a variety of variable-dimension mixtures: break off pieces according to any sequence of random variables with sum almost surely greater than 1. It might be interesting to see if the stick-based inference methods for infinite-dimensional models can be applied to such models.

Asymptotics — Density estimation

- Supposing the data is $X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_0$ from some true density f_0, it is desirable to establish posterior consistency and rates of concentration for density estimation.
- For DPMs, it has been shown that in many cases the posterior on f concentrates at the true density f_0, often at the minimax-optimal rate (up to a logarithmic factor), for any sufficiently regular f_0.
  (Contributions by: Ghosal, van der Vaart, Scricciolo, Tokdar, Dunson, Bhattacharya, Lijoi, Prünster, Walker, James, Wu, Ghosh, Ramamoorthi, Ishwaran, and others.)
- For variable-dimension mixtures, similar results have been established by Kruijer (2008) and Kruijer, Rousseau, & van der Vaart (2010).

Asymptotics — Mixing distribution

- It is also desirable to establish posterior consistency (and rates of concentration, perhaps) for the mixing distribution, if the true density f_0 is, in fact, a mixture from the assumed family $\{p_\theta : \theta \in \Theta\}$.
  - Note: The mixture parameters are always non-identifiable in the usual sense, however, in many cases the mixing distribution is identifiable.
  - Note: In practice, of course, we cannot expect $\{p_\theta : \theta \in \Theta\}$ to be exactly correctly specified. The point of such results is that lack of consistency under ideal conditions would be a red flag.
- For DPMs, Nguyen (2013) has shown, in certain cases, that the posterior on the mixing distribution concentrates in Wasserstein distance at the true mixing distribution.
- For variable-dimension mixtures, if f_0 is a finite mixture from the assumed family, then consistency for the mixing distribution holds under very general conditions, for Lebesgue almost-all values of the true parameters, by Doob's theorem (Nobile, 1994).

Asymptotics — Number of components

- For DPMs, it is fairly common to see the posterior on the number of clusters used for inference about the number of components. However, we have shown that this is inconsistent under quite general conditions (M. & Harrison, 2014).

- For variable-dimension mixtures, Doob's theorem also yields consistency for the number of components in the mixture (Nobile, 1994), **if f_0 is a finite mixture from the assumed family**. However, misspecification of the family of component distributions will often be inevitable in practice, and the estimated number of components can be highly sensitive to such misspecification. One must be wary of this issue when using the number of components (or clusters) to assess the heterogeneity of the data.

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.

- One advantage of this is that these samplers tend to be more generally applicable.

- We will show how this works for two incremental Gibbs sampler algorithms:
    1. "Algorithm 3" for conjugate priors (Neal (1992, 2000), MacEachern (1994)), and
    2. "Algorithm 8" for non-conjugate priors (Neal (2000)).

Incremental Gibbs with a conjugate prior

- For $c \subset \{1, \ldots, n\}$, let $m(x_c) = \displaystyle\int \prod_{j \in c} p_\theta(x_j) H(d\theta)$.
- $m(x_c)$ can be computed analytically when H is a conjugate prior.
- Sampling from $p(\mathcal{C}|x_{1:n}) \propto p(x_{1:n}|\mathcal{C})p(\mathcal{C})$ proceeds as follows.
- Write $\mathcal{C}_{(j)}$ for the current partition, excluding j.

"Algorithm 3" for MFM and DPM

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

	MFM	DPM
at table $c \in \mathcal{C}_{(j)}$ 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 $\propto$	$\gamma \, \frac{v_n(t+1)}{v_n(t)} \, m(x_j)$	$\alpha \, m(x_j)$

where $t = |\mathcal{C}_{(j)}|$ is the number of occupied tables, excluding customer j.

Non-conjugate priors

- Often, the selected family $\{p_\theta\}$ 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 $(\mathcal{C}, (\varphi_c : c \in \mathcal{C}))$, where $\varphi_c \in \Theta$.

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

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

	MFM	DPM
at table $c \in \mathcal{C}_{(j)}$ with probability $\propto$	$(\lvert c\rvert + \gamma)\, p_{\varphi_c}(x_j)$	$\lvert c\rvert\, p_{\varphi_c}(x_j)$
at a new table with probability $\propto$	$\gamma\, \frac{v_n(t+1)}{v_n(t)}\, p_{\varphi_*}(x_j)$	$\alpha\, p_{\varphi_*}(x_j)$

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

2. For each $c \in \mathcal{C}$, sample $\varphi_c \sim p(\varphi_c \mid x_c, \mathcal{C})$, or make a move for which this distribution is invariant.

Outline

1. Mixture of finite mixtures (MFM)
   - Basic properties
   - Asymptotics
   - Inference algorithms (conjugate and non-conjugate cases)

2. Empirical comparison of MFMs with DPMs
   - Density estimation
   - Clustering
   - Mixing issues
   - Bayes factors
   - Outliers

3. Hierarchical mixture of finite mixtures (compare with HDP)

4. Mixture of finite feature models (compare with IBP)

Skew-Normal distribution

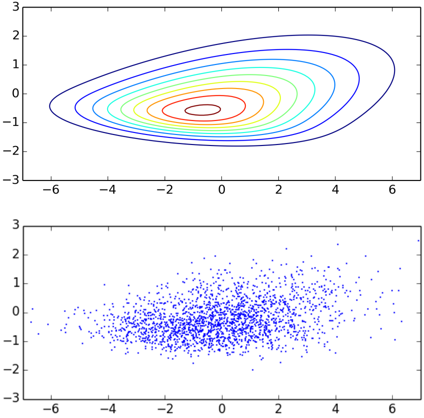
- To make things interesting, we will use multivariate Skew-Normal mixtures. (Experiments suggest that the results would be similar for other families.)
- Azzalini & Dalla Valle (1996) (see also Azzalini & Capitanio (1999)) introduced the multivariate Skew-Normal distribution, with density

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

  for $x \in \mathbb{R}^d$, where S is diagonal with $S_{ii}^2 = Q_{ii}$, and Φ is the univariate standard normal CDF. The parameters are:
  - $\xi \in \mathbb{R}^d$ (location),
  - Q positive definite (scale and correlation),
  - $\alpha \in \mathbb{R}^d$ (skew).
- This family has some nice properties (e.g. preserved under linear maps).

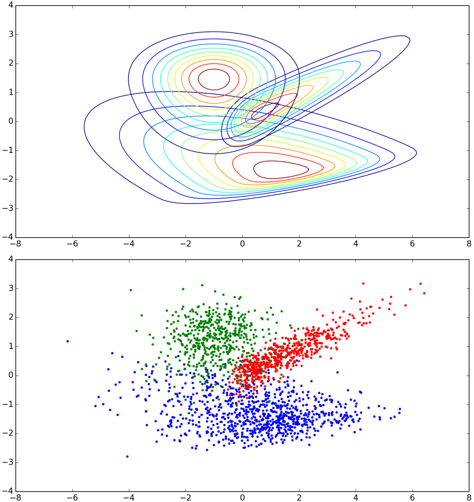
Skew-Normal distribution

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



Bivariate Skew-Normal mixture

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

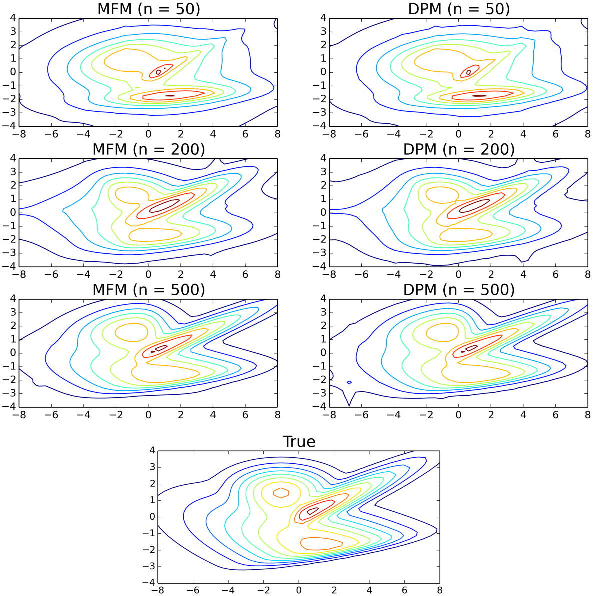


- Since we do not have a conjugate prior, Algorithm 8 as described above was used for inference.
- To define H, for convenience, we chose a parameterization of $\theta = (\xi, Q, \alpha)$ covering all of $\mathbb{R}^7$, and then used independent normal priors.
    - Note: Sensitivity to these choices should be investigated.
- For the MFM parameters, we used $\gamma = 1$ and

$$p(k) \propto \begin{cases} 1 & \text{if } k \in \{1, ..., 30\} \\ 1/(k-30)^2 & \text{if } k > 30. \end{cases}$$

- For the DPM, we used $\alpha = 1$.

Estimated densities



MFM (n = 50)

DPM (n = 50)

MFM (n = 200)

DPM (n = 200)

MFM (n = 500)

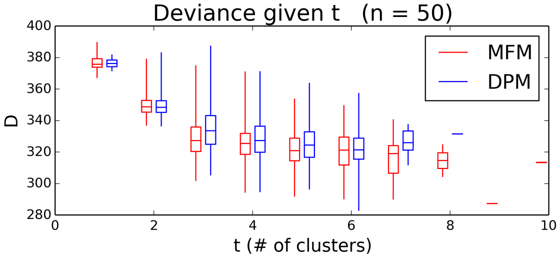
DPM (n = 500)

True

Deviance

The deviance of an estimated density $\hat{f}$ on data $x_1, \ldots, x_n$ is

$$D(\hat{f}) = -2 \sum_{i=1}^{n} \log \hat{f}(x_i) + \text{const.}$$

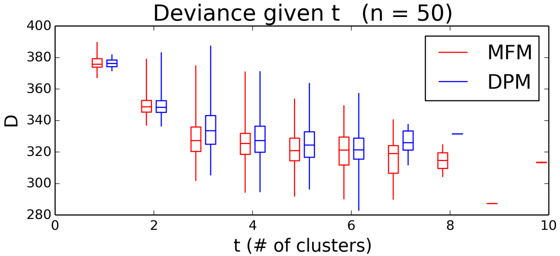


As reported by Green and Richardson (2001), the MFM deviance tends to be slightly lower, given t.

Deviance

The deviance of an estimated density $\hat{f}$ on data $x_1, \ldots, x_n$ is

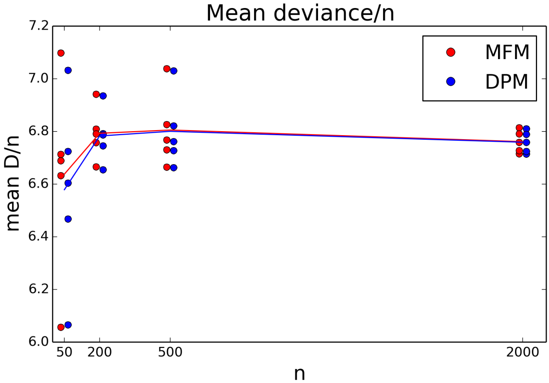
$$D(\hat{f}) = -2 \sum_{i=1}^{n} \log \hat{f}(x_i) + \text{const.}$$



As reported by Green and Richardson (2001), the MFM deviance tends to be slightly lower, given t.

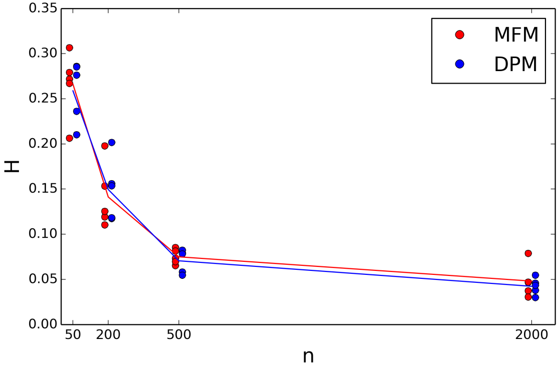
However, when averaged over t, this apparent advantage disappears.

Deviance



- When averaged over t, the deviances are very similar. This is possible since the DPM puts more weight on higher t values, for which the deviance tends to be smaller.
- Interesting example of **Simpson's paradox.**

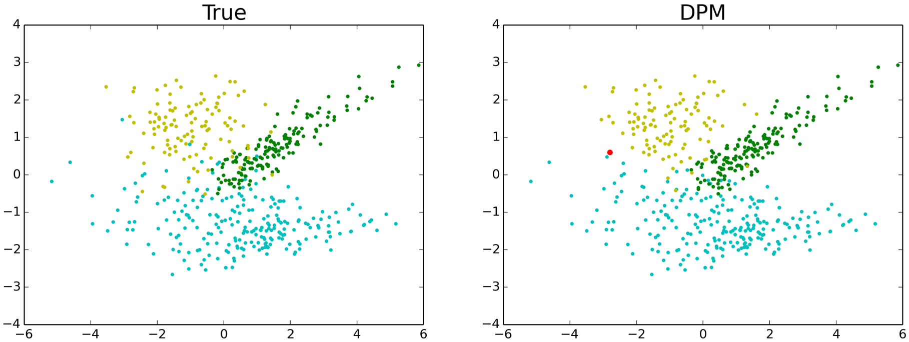
Hellinger distance to the true density



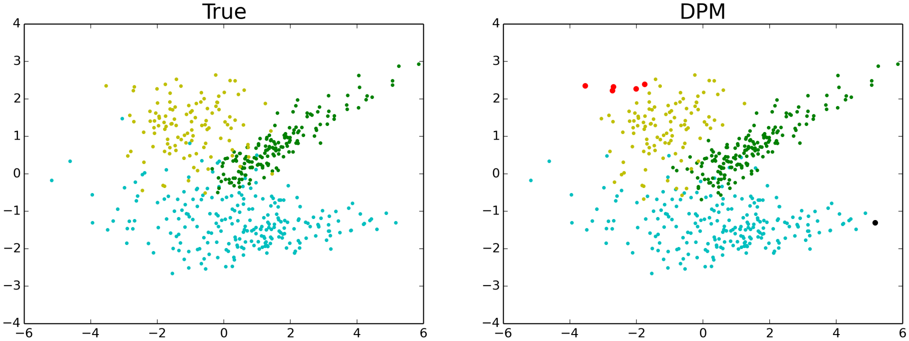
$$H(f,g)^2 = \frac{1}{2}\int\left(\sqrt{f(x)}-\sqrt{g(x)}\right)^2 dx \approx \frac{1}{2N}\sum_{i=1}^{N}\left(\sqrt{\frac{f(Y_i)}{g(Y_i)}}-1\right)^2$$

where $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} g$. Each distance was estimated with $N = 10^3$ samples from the true density.

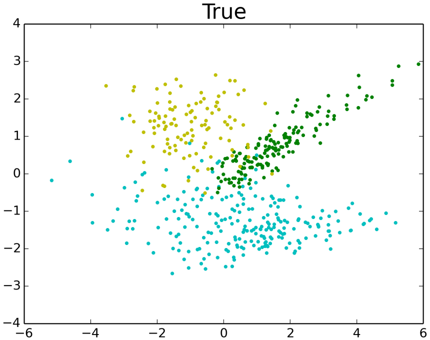
Sample clusterings from the posterior

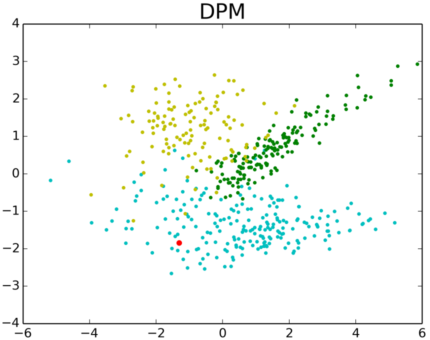


Sample clusterings from the posterior

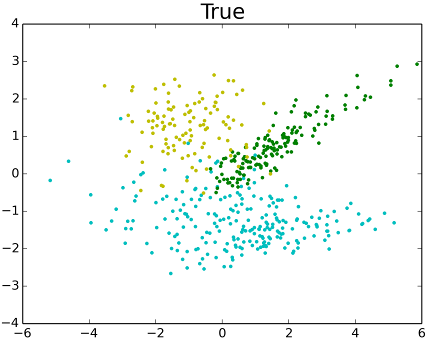


Sample clusterings from the posterior

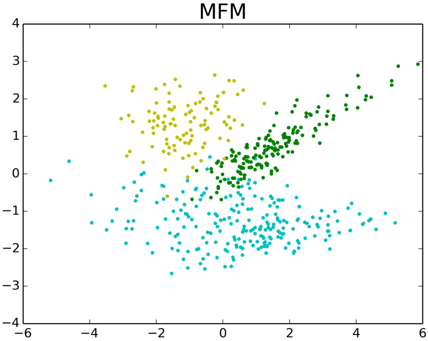




Sample clusterings from the posterior

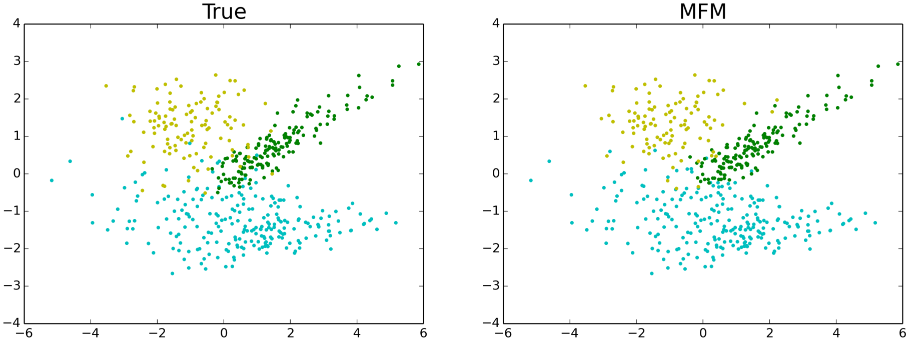


True

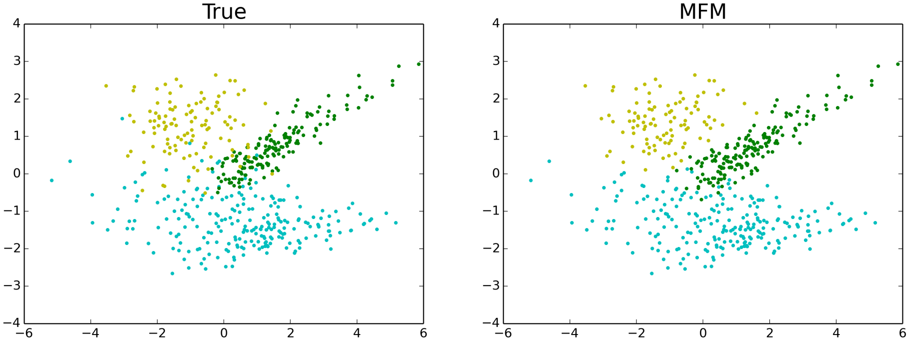


MFM

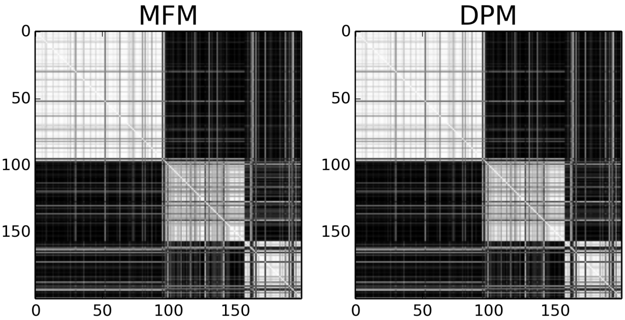
Sample clusterings from the posterior



Sample clusterings from the posterior

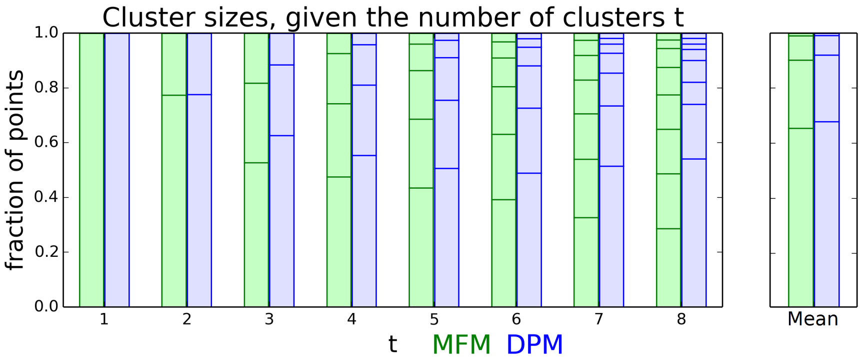


Pairwise probability matrix



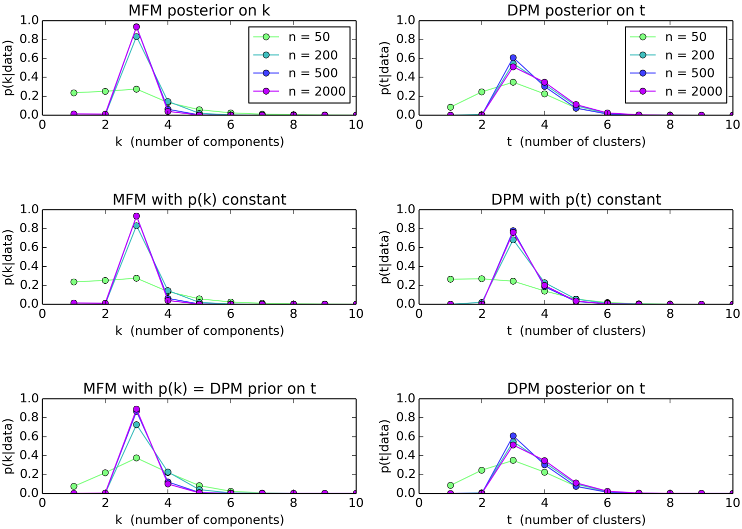
- Entry i, j is the posterior probability that data points i and j are in the same cluster.
- Very similar results:
  - Mean squared difference $= \frac{1}{n^2}\sum_{i,j}(p_{ij} - q_{ij})^2 = 0.0003$
  - Max absolute difference $= \max_{i,j} |p_{ij} - q_{ij}| = 0.08$

Cluster sizes



(Data shown is for $n = 50$.)

- As reported by Green & Richardson (2001), the MFM clusters tend to be more equally sized (i.e. higher entropy), given t.
- However, the difference is less pronounced (usually negligible) when averaging over t.



MFM posterior on k

DPM posterior on t

MFM with p(k) constant

DPM with p(t) constant

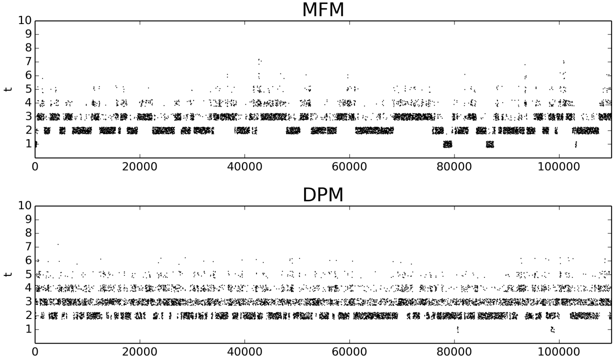
MFM with p(k) = DPM prior on t

DPM posterior on t

- These results are the average over 5 runs.
- Note: These posteriors can be sensitive to the prior, H.
- Note: $n = 2000$ for MFM is unreliable due to poor mixing.

Mixing issues with incremental Gibbs

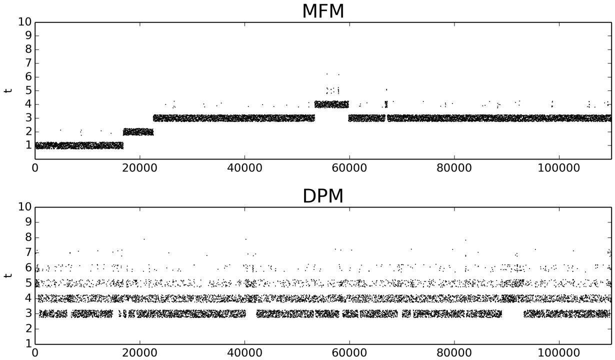
Traceplot of the number of clusters t, with $n = 50$



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 severe. 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.

Solutions?

- Richardson & Green (1997, 2001) used a split-merge sampler with reversible jump (in the univariate Gaussian case).

- Many split-merge samplers for the DPM have been developed (e.g. Dahl (2003, 2005), Jain & Neal (2004, 2007)) for both the conjugate and non-conjugate case.
    - These can now be used for MFMs also, using the partition distribution as derived above.

- Alternatively, we can use **importance sampling** to take advantage of good DPM samplers . . .

Importance sampling

If $f(y) = \widetilde{f}(y)/Z_f$ and $g(y) = \widetilde{g}(y)/Z_g$ then

$$\begin{aligned} E_f h(Y) &= \sum_y h(y) f(y) \\ &= \frac{Z_g}{Z_f} \sum_y h(y) \frac{\widetilde{f}(y)}{\widetilde{g}(y)} g(y) \\ &\approx \frac{Z_g}{Z_f} \frac{1}{N} \sum_{i=1}^{N} h(Y_i) \widetilde{w}(Y_i) \end{aligned}$$

and

$$\frac{Z_f}{Z_g} \approx \frac{1}{N} \sum_{i=1}^{N} \widetilde{w}(Y_i),$$

where $\widetilde{w}(y) = \widetilde{f}(y)/\widetilde{g}(y)$ and $Y_1, \ldots, Y_N \sim g$.

Importance sampling for the MFM

Denote $p =$ MFM and $q =$ DPM. Taking $y = \mathcal{C}$ (the partition) and

$$\begin{aligned}\widetilde{f}(\mathcal{C}) &= p(x_{1:n}|\mathcal{C})p(\mathcal{C}),\\ \widetilde{g}(\mathcal{C}) &= q(x_{1:n}|\mathcal{C})q(\mathcal{C}),\end{aligned}$$

we can use samples $\mathcal{C}_1, \ldots, \mathcal{C}_N$ from the DPM posterior $g(\mathcal{C}) = q(\mathcal{C}|x_{1:n})$ to estimate posterior expectations under the MFM via

$$E_p(h(\mathcal{C})|x_{1:n}) \approx \frac{\sum_i h(\mathcal{C}_i)\widetilde{w}(\mathcal{C}_i)}{\sum_i \widetilde{w}(\mathcal{C}_i)}.$$

The importance weights simplify:

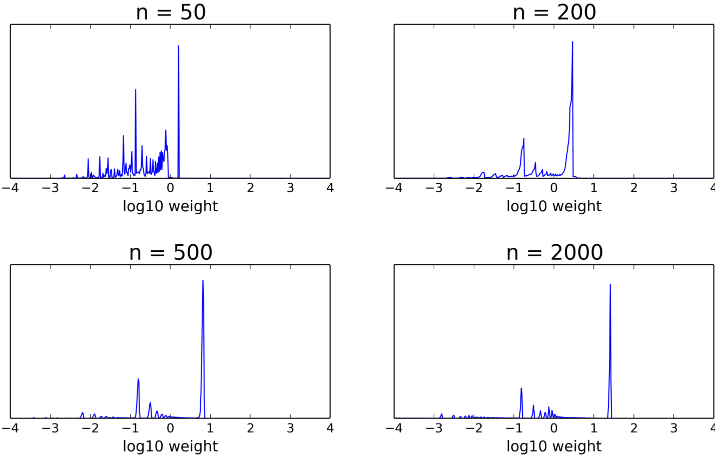
$$\widetilde{w}(\mathcal{C}) = \frac{p(x_{1:n}|\mathcal{C})p(\mathcal{C})}{q(x_{1:n}|\mathcal{C})q(\mathcal{C})} = \frac{p(\mathcal{C})}{q(\mathcal{C})},$$

making this a very easy and efficient technique.

Importance sampling for the MFM

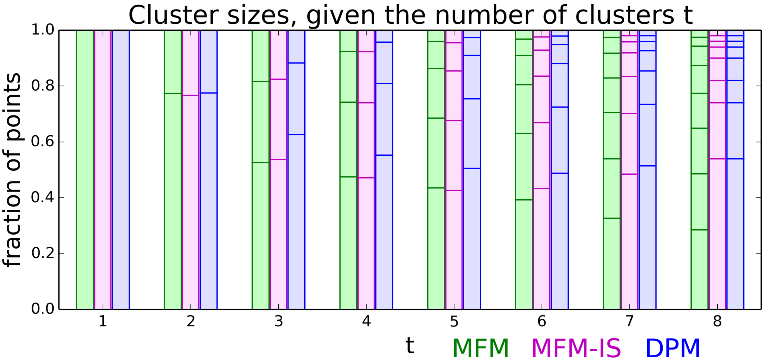
- In principle, this allows us to take the output of any DPM sampler and do inference for the corresponding MFM.
- Can take advantage of all the advances made in DPM sampler development.
- This is only possible now that we have a simple expression for $p(\mathcal{C})$.
- For this to work, the importance weights need to be well-behaved . . .

Distribution of the importance weights



The effective sample size (ESS) is usually over 50%, and ranges from 30% to 75% in simulations so far.

Example: Cluster sizes



The MFM and the DPM-based importance sampling approximation (MFM-IS) are close in the range of t values of non-negligible posterior probability $(t = 2, \ldots, 5)$.

Somewhat more generally

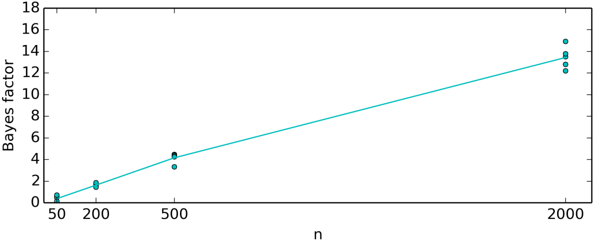
- This IS approach could also be used for other partition-based mixture models, such as product partition models (as long as the importance weights remain well-behaved).
- Inference in multiple such models could be done easily and quickly — all using the same set of samples — just by specifying the different partition distributions.
- Conveniently, the Bayes factors are already estimated as part of the procedure.

Bayes factors: MFM/DPM

The importance weights also give us an estimate of the ratio of the normalizing constants, which equals the Bayes factor:

$$\frac{p(x_{1:n})}{q(x_{1:n})} = \frac{Z_f}{Z_g} \approx \frac{1}{N}\sum_{i=1}^{N} \frac{p(\mathcal{C}_i)}{q(\mathcal{C}_i)}.$$

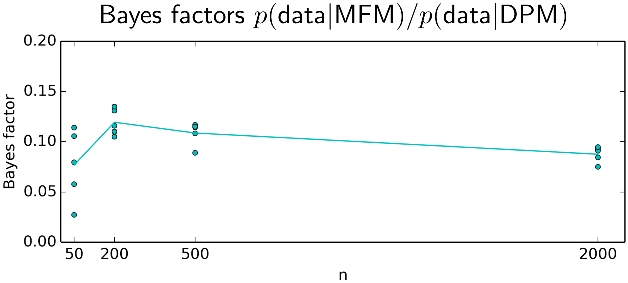
Bayes factors $p(\text{data}|\text{MFM})/p(\text{data}|\text{DPM})$



Clearly, the Bayes factors are favoring the MFM for larger n. This makes some sense, since the MFM is correctly specified in this situation.

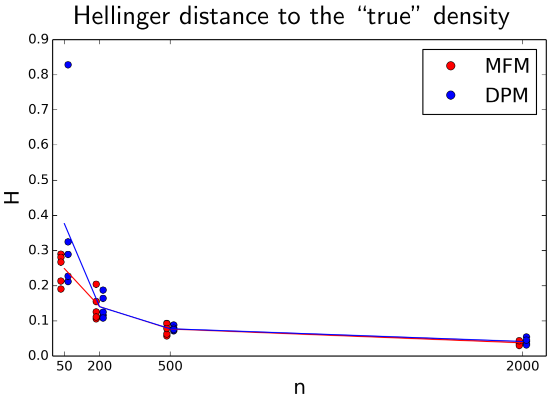
So, what if there is misspecification, or outliers?

If the experiments are repeated with a single outlier at $(100, 0)$, the Bayes factors are in favor of the DPM:



This makes sense, since the DPM likes having tiny clusters, while the MFM does not.

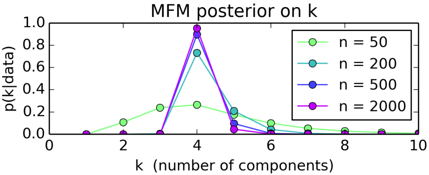
Density estimation with an outlier

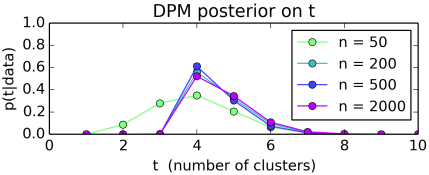


Still, as before, the density estimates appear to be very similar.

Clustering with an outlier

And both the MFM and DPM put the outlier in a separate cluster:





Outline

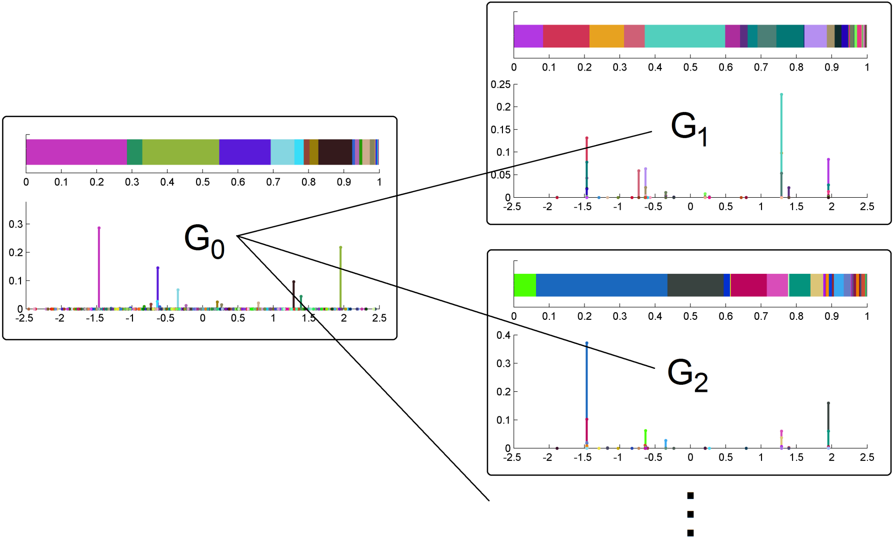
1. Mixture of finite mixtures (MFM)
   - Basic properties
   - Asymptotics
   - Inference algorithms (conjugate and non-conjugate cases)

2. Empirical comparison of MFMs with DPMs
   - Density estimation
   - Clustering
   - Mixing issues
   - Bayes factors
   - Outliers

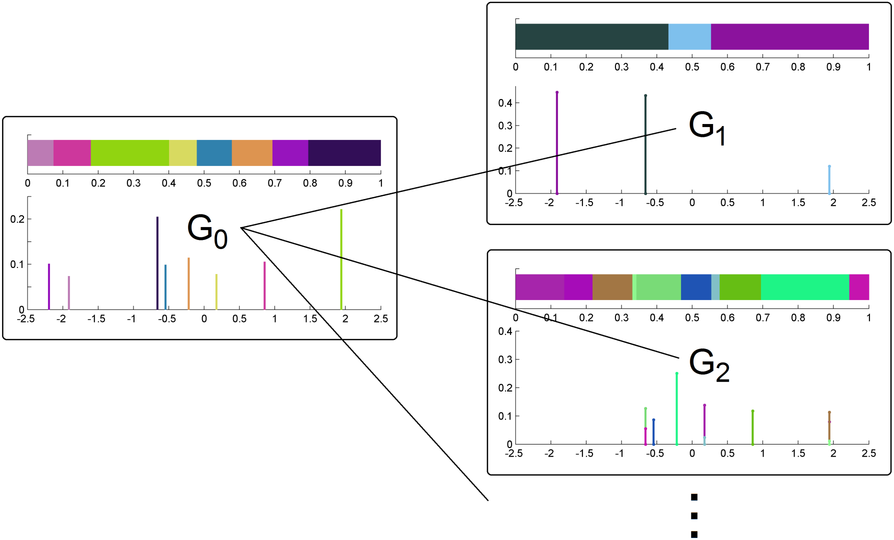
3. Hierarchical mixture of finite mixtures (compare with HDP)

4. Mixture of finite feature models (compare with IBP)

Hierarchical Dirichlet process (HDP) (Teh et al. 2006)



Hierarchical MF (HMF)

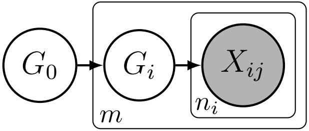


Hierarchical MFM (HMFM)

$G_0 \sim \mathrm{MF}(\gamma, H, p_K)$

$G_1, \ldots, G_m \overset{\text{iid}}{\sim} \mathrm{MF}(\gamma, G_0, p_K)$, given G_0

$X_{ij} \sim f_{G_i}(x)$ independent for $j \in \{1, \ldots, n_i\}$, $i \in \{1, \ldots, m\}$.



We refer to the distribution of the G's as a HMF, and the distribution of the X's as a HMFM.

Hierarchical partition distribution (HDP vs HMF)

For $i = 1, \ldots, m$, let $\mathcal{C}_i$ be a partition of $\{1, \ldots, n_i\}$, and let $t_i = |\mathcal{C}_i|$. Let $\mathcal{C}_0$ be a partition of $\{1, \ldots, N\}$ where $N = \sum t_i$, and let $t_0 = |\mathcal{C}_0|$. Then letting $\mathcal{C} = (\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_m)$,

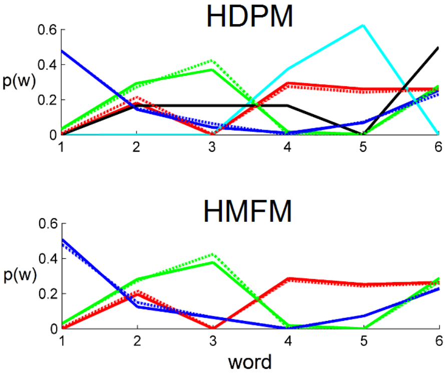
$$P_{\text{HDP}}(\mathcal{C}) = P_{\text{DP}}^{(N)}(\mathcal{C}_0) \prod_{i=1}^{m} P_{\text{DP}}^{(n_i)}(\mathcal{C}_i) \qquad P_{\text{HMF}}(\mathcal{C}) = P_{\text{MF}}^{(N)}(\mathcal{C}_0) \prod_{i=1}^{m} P_{\text{MF}}^{(n_i)}(\mathcal{C}_i)$$

where $P_{\text{DP}}^{(n)}$ and $P_{\text{MF}}^{(n)}$ are the DP and MF partition distributions on $\{1, \ldots, n\}$, respectively. (Note that $\mathcal{C}_0$ depends on $\mathcal{C}_1, \ldots, \mathcal{C}_m$ through $N = \sum t_i$.)

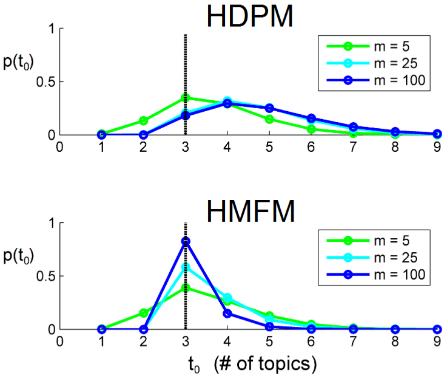
- This leads to a simple “franchise process” very similar that of the HDP.
- Gibbs sampling for HMFMs and HDPMs is nearly identical.
- Since N is not fixed, caching $v_N(t_0)$ is more efficient than precomputing.

Preliminary results with a toy topic model

Typical posterior topic distributions



Posterior on # of topics



Outline

1. Mixture of finite mixtures (MFM)
   - Basic properties
   - Asymptotics
   - Inference algorithms (conjugate and non-conjugate cases)
2. Empirical comparison of MFMs with DPMs
   - Density estimation
   - Clustering
   - Mixing issues
   - Bayes factors
   - Outliers
3. Hierarchical mixture of finite mixtures (compare with HDP)
4. Mixture of finite feature models (compare with IBP)

Mixture of finite feature models (MFFM)

An alternative to the Indian buffet process (IBP) of Griffiths & Ghahramani (2005):

Sample K.

$$\begin{array}{ccccc} \pi_1 & \pi_2 & \pi_3 & \cdots & \pi_k \\ \hline 0 & 1 & & & \\ 1 & 0 & & & \\ 1 & 1 & \cdots & \multicolumn{2}{l}{Z_{ij} \sim B(\pi_j)} \\ \vdots & \vdots & & \multicolumn{2}{l}{\text{given } \pi_1,\ldots,\pi_k} \\ 0 & 0 & & & \\ 1 & 0 & & & \end{array}$$

$\sim$ Beta(a,b)
i.i.d. given K=k

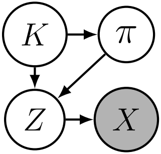
Mixture of finite feature models (MFFM)

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

$\pi_1, \dots, \pi_k \overset{\text{iid}}{\sim} \mathrm{Beta}(a, b)$ (given $K = k$)

For $j \in \{1, \dots, k\}$ (given $K = k$ and π):

$Z_{1j}, \dots, Z_{nj} \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(\pi_j)$.



We refer to the distribution on the "feature matrix" Z as a MFFM.

Equivalence class distribution (IBP vs MFFM)

Consider two binary matrices equivalent if they are the same after removing any columns containing only zeros. The probability of obtaining $\bar{Z} \in \{0,1\}^{n \times t}$ with column sums $m_1, \ldots, m_t > 0$ after removing any zero columns from Z is

$$P_{\text{IBP}}(\bar{Z}) = \frac{\alpha^t e^{-\alpha H_n}}{t!} \prod_{i=1}^{t} \frac{(m_i - 1)!\,(n - m_i)!}{n!}$$

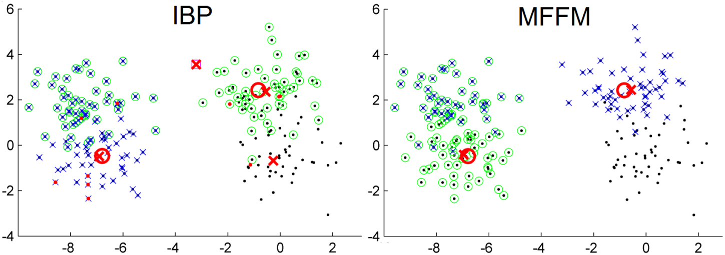
$$P_{\text{MFFM}}(\bar{Z}) = v_n'(t) \prod_{i=1}^{t} \frac{a^{(m_i)}\, b^{(n-m_i)}}{(a+b)^{(n)}}$$

where $v_n'(t) = \sum_{k=0}^{\infty} \binom{k}{t} c_n^{k-t} p(k)$, with $c_n = \dfrac{b^{(n)}}{(a+b)^{(n)}}$.

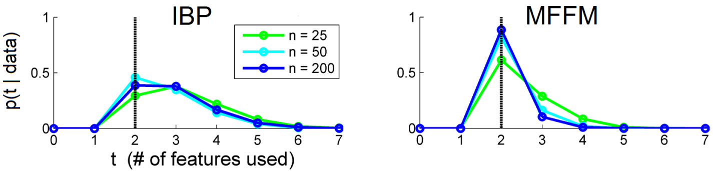
- If $p(k) = \text{Poisson}(k \mid \lambda)$ then $v_n'(t) = e^{\lambda c_n}\, \text{Poisson}(t \mid \lambda)$.
- In general, $v_n'(t)$ can be efficiently precomputed to arbitrary precision.
- Gibbs sampling for MFFMs and IBPs is nearly identical.

Preliminary results with a toy feature model

Typical posterior feature assignments



Posterior on # of features used



Possible future work

- Dealing with misspecification
- Sensitivity to the prior
- Posterior concentration rates
- Split-merge samplers, variational methods

Combinatorial stochastic processes for variable-dimension models

Jeffrey W. Miller

Joint work with Matt Harrison

Brown University
Division of Applied Mathematics

Duke Statistical Science Seminar
Feb 7, 2014

This research was supported by DARPA and the NSF.