Mixture models with a prior on the number of components

Miller and Harrison, JASA, to be appear

Presented by Ilsang Ohn

October 21, 2017

Mixture of finite mixtures

★ A mixture of finite mixtures (MFM) model:

$$
\begin{aligned}
K &\sim p_K \text{ where } p_K \text{ is a pmf on } \{1, 2, \dots\} \\
\boldsymbol{\pi}_{1:K} = (\pi_1, \dots, \pi_K) | K &\sim \mathrm{Dir}_K(\gamma, \dots, \gamma) \\
z_1, \dots, z_n | \boldsymbol{\pi}_{1:K} &\overset{\text{iid}}{\sim} \boldsymbol{\pi}_{1:K} \\
\theta_1, \dots, \theta_K | K &\overset{\text{iid}}{\sim} H \\
x_i | z_i, \boldsymbol{\theta}_{1:K} &\overset{\text{ind}}{\sim} f_{\theta_{z_i}} \text{ for } i = 1, \dots, n
\end{aligned}
\tag{1}
$$

* The Dirichlet process mixture (DPM) model:

$$
\begin{aligned}
B_1, B_2, \dots &\overset{\text{iid}}{\sim} \text{Beta}(1, \alpha) \\
z_1, \dots, z_n | \boldsymbol{\pi}_{1:\infty} &\overset{\text{iid}}{\sim} \boldsymbol{\pi}_{1:\infty}, \text{ where } \pi_k = B_k \textstyle\prod_{l=1}^{k} (1 - B_l) \\
\theta_1, \theta_2, \dots &\overset{\text{iid}}{\sim} H \\
x_i | z_i, \boldsymbol{\theta}_{1:\infty} &\overset{\text{ind}}{\sim} f_{\theta_{z_i}} \text{ for } i = 1, \dots, n
\end{aligned}
\tag{2}
$$

Exchangeable partition distribution

- Let $\mathcal{C} = \{E_k : |E_k| > 0\}$ where $E_k = \{i \in [n] : z_i = k\}$ for $k = 1, 2, \ldots$.

★ Under the MFM, the pmf of $\mathcal{C}$ is

$$p(\mathcal{C}) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \tag{3}$$

where $t = |\mathcal{C}|$ is the number of blocks in the partition and

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

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

- $\mathcal{C}$ is an exchangeable random partition of $[n]$.
- $V_n(t)$ always converges to a finite value. If $t \ll n$ it converges rapidly.

Exchangeable partition distribution

★ Under the MFM, the pmf of $\mathcal{C}$ is

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

* Under the DPM, the pmf of $\mathcal{C}$ is

$$p_{\mathrm{DP}}(\mathcal{C}) = V_n^{\mathrm{DP}}(t) \prod_{c \in \mathcal{C}} (|\mathcal{C}| - 1)!$$

where $V_n^{\mathrm{DP}}(t) = \int \{\alpha^t / \alpha^{(n)}\} p(\alpha) \mathrm{d}\alpha$

Polya urn scheme

★ Polya urn scheme of the MFM:

1. Initialize with a single cluster consisting of element 1 alone: $\mathcal{C}_1 = \{\{1\}\}$.
2. For $n = 2, 3, \ldots$, element n is placed in
   - an existing cluster $c \in \mathcal{C}_{n-1}$ with probability
   $$\propto |c| + \gamma$$
   - a new cluster with probability
   $$\propto \frac{V_n(t+1)}{V_n(t)}\gamma$$

   where $t = |\mathcal{C}_{n-1}|$

* CRP: $\propto |c|$, $\propto \alpha$

Random discrete measures

- With K, $\boldsymbol{\pi}_{1:K}$ and $\boldsymbol{\theta}_{1:K}$ in the MFM model, let

$$\beta_i (= \theta_{z_i}) \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \pi_k \delta_{\theta_k}$$

★ If H is continuous, then $\beta_1 \sim H$ and

$$p(\beta_n | \beta_1, \ldots, \beta_{n-1}) \propto \frac{V_n(t+1)}{V_n(t)} \gamma H + \sum_{j=1}^{t} (n_j + \gamma) \delta_{\beta_j^*}$$

where $\beta_1^*, \ldots, \beta_t^*$ are the distinct values taken by $\beta_1, \ldots, \beta_{n-1}$ and $n_j = |\{i \in [n-1] : \beta_i = \beta_j^*\}|$

* DPM: $\beta_i \overset{\text{iid}}{\sim} \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \sim \mathrm{DP}(\alpha)$

$$p(\beta_n | \beta_1, \ldots, \beta_{n-1}) \propto \alpha H + \sum_{j=1}^{t} n_j \delta_{\beta_j^*}$$

Relationship between the number of clusters and number of components

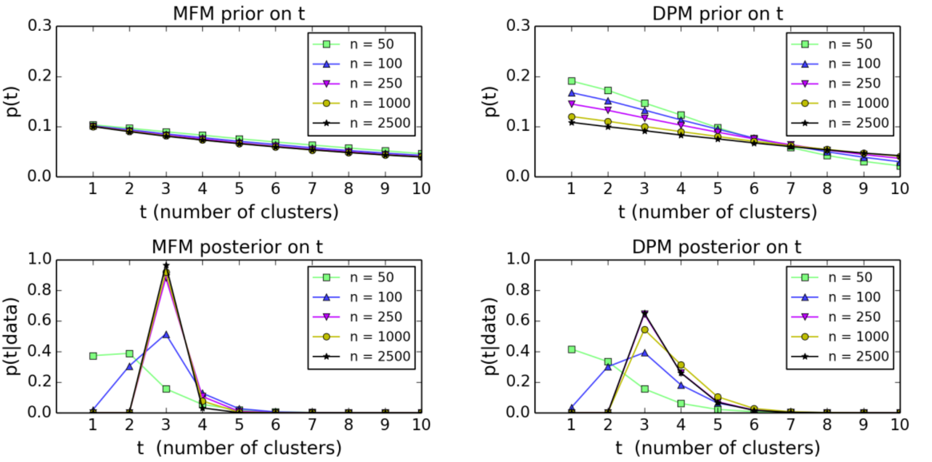
★ For given $\mathbf{x}_{1:n}$, if $p_K(1), \ldots, p_K(k) > 0$ then as $n \to \infty$

$$|p(|\mathcal{C}| = k|\mathbf{x}_{1:n}) - p(K = k|\mathbf{x}_{1:n})| \to 0$$

$$i.e., p(|\mathcal{C}| < k|K = k, \mathbf{x}_{1:n}) \to 0$$

which means that the MFM prior on $t = |\mathcal{C}|$ converges to the prior on K as n grows.

* In a Dirichlet process, the prior on t takes a particular parametric form and diverges at $\log n$ rate.


MFM prior on t
p(t)
0.0
0.1
0.2
0.3
t (number of clusters)
1 2 3 4 5 6 7 8 9 10
n = 50
n = 100
n = 250
n = 1000
n = 2500
DPM prior on t
p(t)
0.0
0.1
0.2
0.3
t (number of clusters)
1 2 3 4 5 6 7 8 9 10
n = 50
n = 100
n = 250
n = 1000
n = 2500
MFM posterior on t
p(t|data)
0.0
0.2
0.4
0.6
0.8
1.0
t (number of clusters)
1 2 3 4 5 6 7 8 9 10
n = 50
n = 100
n = 250
n = 1000
n = 2500
DPM posterior on t
p(t|data)
0.0
0.2
0.4
0.6
0.8
1.0
t (number of clusters)
1 2 3 4 5 6 7 8 9 10
n = 50
n = 100
n = 250
n = 1000
n = 2500

Distribution of the cluster sizes

- Let $A = (A_1, \ldots, A_T)$ be the ordered partition of $[n]$ obtained by randomly ordering the clusters of $\mathcal{C}$ and let $S = (S_1, \ldots, S_T)$ be the vector of block sizes of A, i.e., $S_k = |A_k|$.

★ MFM: Sizes of the clusters are similar order.

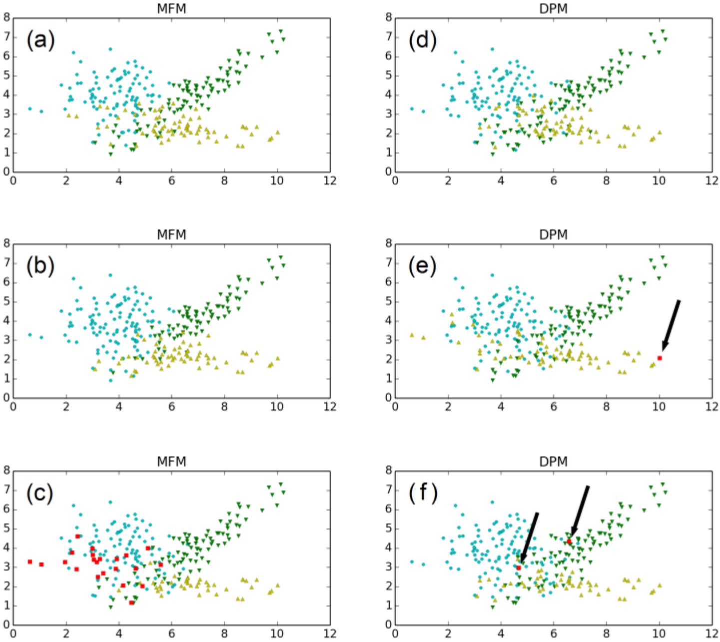
$$p(S = s | T = t) \approx \kappa \prod_{k=1}^{t} s_i^{\gamma - 1}$$

where κ is a normalization constant.

* DPM: Sizes of the clusters are vary widely, with a few large clusters and many very small clusters.

$$p_{\mathrm{DP}}(S = s | T = t) = \kappa_{\mathrm{DP}} \prod_{k=1}^{t} s_i^{-1}$$

where $\kappa_{\mathrm{DP}} = V_n^{\mathrm{DP}}(t) n! / t!$.


MFM
(a)
DPM
(d)
MFM
(b)
DPM
(e)
MFM
(c)
DPM
(f)

Inference algorithms

★ Inference algorithm for the MFM:

1. Initialize $\mathcal{C} = \{[n]\}$
2. For $i = 1, \ldots, n$: Remove element i from $\mathcal{C}$ and place it
   - in $c \in \mathcal{C} \setminus i$ with probability

$$\propto (|c| + \gamma) \frac{m(\mathbf{x}_{c \cup i})}{m(\mathbf{x}_c)}$$

     where $m(\mathbf{x}_c) = \int \left[\prod_{i \in c} f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right] H(\mathrm{d}\boldsymbol{\theta})$
   - in a new cluster with probability

$$\propto \frac{V_n(t+1)}{V_n(t)} m(x_i)$$

     where $t = |\mathcal{C} \setminus i|$
3. Repeat the above steps N times to obtain N samples:

This is direct adaptation of "Algorithm 3" for DPMs. When $m(\mathbf{x}_c)$ can not be computed, we can apply an auxiliary variable technique such as "Algorithm 8"

* For DPMs, $|c| + \gamma$ is replaced by γ and $\gamma V_n(t+1)/V_n(t)$ is replaced by α.

Mixture of finite latent feature models

Mixture of finite latent feature models

★ A mixture of finite latent factor model (MFLFM):

$$
\begin{aligned}
K &\sim p_K \text{ where } p_K \text{ is a pmf on } \{0, 1, 2, \ldots\} \\
\pi_k | K &\overset{\text{iid}}{\sim} \text{Beta}\,(\alpha, 1) \;\text{ for } k = 1, \ldots, K \\
z_{1k}, \ldots, z_{nk} | \pi_k, K &\overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_k) \text{ for } k = 1, \ldots, K
\end{aligned} \tag{4}
$$

* Indian buffet process (IBP): A limit of the followings as $K \to \infty$:

$$
\begin{aligned}
\pi_k | K &\overset{\text{iid}}{\sim} \text{Beta}\,(\alpha/K, 1) \;\text{ for } k = 1, \ldots, K \\
z_{1k}, \ldots, z_{nk} | \pi_k &\overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_k) \text{ for } k = 1, \ldots, K
\end{aligned} \tag{5}
$$

2- parameter IBP $\pi_k | K \overset{\text{iid}}{\sim} \text{Beta}\,(\alpha\beta/K, \beta)$

Probability of lof equivalence class

★ MFLFM: If $p_K = \mathrm{Poisson}(\gamma)$,

$$\mathbb{P}([\mathbf{Z}]) = \frac{(\alpha\gamma)^{K_+}}{\prod_{h=0}^{2^n-1} K_h} \exp\left(-\alpha\gamma \sum_{j=1}^{n} \frac{1}{j} \prod_{l=1}^{j} \frac{l}{l+\alpha}\right) \left[\prod_{k=1}^{K_+} B(m_k+\alpha, n-m_k+1)\right]$$

* IBP:

$$\mathbb{P}([\mathbf{Z}]) = \frac{(\alpha)^{K_+}}{\prod_{h=0}^{2^n-1} K_h} \exp\left(-\alpha \sum_{j=1}^{n} \frac{1}{j}\right) \left[\prod_{k=1}^{K_+} B(m_k, n-m_k+1)\right]$$

2-parameter IBP:

$$\mathbb{P}([\mathbf{Z}]) = \frac{(\alpha\beta)^{K_+}}{\prod_{h=0}^{2^n-1} K_h} \exp\left(-\alpha \sum_{j=1}^{n} \frac{\beta}{j+\beta-1}\right) \left[\prod_{k=1}^{K_+} B(m_k, n-m_k+\beta)\right]$$

Restaurant process

★ MFLFM:

1. The first customer tries $\mathrm{Poisson}(\alpha\gamma/(1+\alpha))$ dishes.
2. For every $j \geq 2$, the j-th customer
   - tries each previously tasted dish independently with probability
     $$\frac{m_k + \alpha}{j + \alpha}$$
     where m_k is the number of people who have tried dish k;
   - and tries
     $$\mathrm{Poisson}\left(\alpha\gamma\frac{1}{j}\prod_{l=1}^{j}\frac{l}{l+\alpha}\right)$$
     new dishes

* IBP:

$$\frac{m_k}{j};\ \mathrm{Poisson}\left(\frac{\alpha}{j}\right)$$

2-parameter IBP:

$$\frac{m_k}{\beta+j-1};\ \mathrm{Poisson}\left(\frac{\alpha\beta}{\beta+j-1}\right)$$

The number of factors

★ MFLFM:

$$K^+ \sim \text{Poisson}\left(\alpha\gamma \sum_{j=1}^{n} \frac{1}{j} \prod_{l=1}^{j} \frac{l}{l+\alpha}\right)$$

and

$$\mathbb{E}K^+ \sim \sum_{j=1}^{n} \frac{1}{n^{\alpha+1}} \overset{n\to\infty}{<} \infty$$

* IBP: $K^+ \sim \text{Poisson}\left(\alpha \sum_{j=1}^{n} \frac{1}{j}\right)$ and

$$\mathbb{E}K^+ \sim \sum_{j=1}^{n} \frac{1}{j} \sim \log n$$

2-parameter IBP: $K^+ \sim \text{Poisson}\left(\alpha \sum_{j=1}^{n} \frac{\beta}{\beta+j-1}\right)$ and $\mathbb{E}K^+ \sim \sum_{j=1}^{n} \frac{1}{j} \sim \log n$

Posterior consistency

- Posterior consistency of the number of factors

$$P(K^+ < k_{0n}|\mathbf{x}_{1:n}) \xrightarrow{\mathbb{P}_0} 0, \quad P(K^+ > k_{0n}|\mathbf{x}_{1:n}) \xrightarrow{\mathbb{P}_0} 0,$$

- For the first term

$$P(K^+ < k_{0n}|\mathbf{x}_{1:n}) = \frac{\int_{\{K^+ < k_{0n}\}} f(\mathbf{x}_{1:n}|\phi)\mathrm{d}P(\phi)}{\int f(\mathbf{x}_{1:n}|\phi)\mathrm{d}P(\phi)} \leq \frac{P(K^+ < k_{0n})}{\int f(\mathbf{x}_{1:n}|\phi)\mathrm{d}P(\phi)}$$