EM Algorithm, Mixture of Multinomial, Latent Dirichlet Allocation

Course Webpage:
http://www.cs.cornell.edu/Courses/cs4786/2017fa/
Set of models $\Theta$ consists of parameters s.t. $P_\theta$ for each $\theta \in \Theta$ is a distribution over data.

Learning: Estimate $\theta^* \in \Theta$ that best models given data.
Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

$$\theta_{MLE} = \arg\max_{\theta \in \Theta} \log P_\theta(x_1, \ldots, x_n)$$

- A priori all models are equally good, data could have been generated by any one of them
Maximum A Posteriori

Pick $\theta \in \Theta$ that is most likely given data

Maximize a posteriori probability of model given data

$$\theta_{MAP} = \arg\max_{\theta \in \Theta} P(\theta | x_1, \ldots, x_n)$$

$$= \arg\max_{\theta \in \Theta} \log P(x_1, \ldots, x_n | \theta) + \log P(\theta)$$
EM Algorithm
We only observe $x_1, \ldots, x_n$, cluster assignments $c_1, \ldots, c_n$ are not observed.

Finding $\theta \in \Theta$ (even for 1-d GMM) that directly maximizes Likelihood or A Posteriori given $x_1, \ldots, x_n$ is hard!

Given latent variables $c_1, \ldots, c_n$, the problem of maximizing likelihood (or a posteriori) became easy.

Can we use latent variables to device an algorithm?
Say $c_1, \ldots, c_n$ are Latent variables. Eg. cluster assignments
Say $c_1, \ldots, c_n$ are Latent variables. Eg. cluster assignments

- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

  (E step) For every $t$, define distribution $Q_t$ over the latent variable $c_t$ as:

  $$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

  (M step)

  $$\theta^{(i)} = \arg\max_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t|\theta) \quad \text{if MLE}$$
Say $c_1, \ldots, c_n$ are Latent variables. Eg. cluster assignments

- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

**E step** For every $t$, define distribution $Q_t$ over the latent variable $c_t$ as:

$$Q_t^{(i)}(c_t) = P(c_t | x_t, \theta^{(i-1)})$$

$$\propto P(x_t | c_t, \theta^{(i-1)}) P(c_t | \theta^{(i-1)})$$

**M step**

$$\theta^{(i)} = \arg\max_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t | \theta) \quad \text{if MLE}$$
Say $c_1, \ldots, c_n$ are Latent variables. Eg. cluster assignments

- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

  (E step) For every $t$, define distribution $Q_t$ over the latent variable $c_t$ as:

  $$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

  $$\propto P(x_t|c_t, \theta^{(i-1)}) P(c_t|\theta^{(i-1)})$$

  (M step)

  $$\theta^{(i)} = \text{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t|\theta)$$

  if MLE

  $$\theta^{(i)} = \text{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q_t^{(i)}(c_t) \log P(x_t, c_t|\theta) + \log P(\theta)$$

  if MAP
Why EM works?

• Every iteration of EM only improves log-likelihood (log a posteriori)
Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$
\log P_{\theta^{(i)}}(x_1, \ldots, x_n)
$$
Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) = \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)$$
**Why should EM work?**

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) = \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t) \\
= \sum_{t=1}^{n} \log \left( \sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right)
$$
Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) = \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)$$

$$= \sum_{t=1}^{n} \log \left( \sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right)$$

$$= \sum_{t=1}^{n} \log \left( \sum_{c_t=1}^{K} Q^{(i)}(c_t) \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \right)$$
Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) = \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)
$$

$$
= \sum_{t=1}^{n} \log \left( \sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right)
$$

$$
= \sum_{t=1}^{n} \log \left( \sum_{c_t=1}^{K} Q^{(i)}(c_t) \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \right)
$$

$$
\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)
$$
Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) = \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)
= \sum_{t=1}^{n} \log \left( \sum_{c_t=1}^{K} P_{\theta^{(i)}}(x_t, c_t) \right)
= \sum_{t=1}^{n} \log \left( \sum_{c_t=1}^{K} Q^{(i)}(c_t) \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \right)
\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)
$$

Log(average) > average of Log
Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)
$$
Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

\[
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)
\]

\[
\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \quad \text{M-step}
\]
Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \quad \text{M-step}$$

$$= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t|x_t)} \right) \quad \text{E-step}$$
Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

$$
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \\
\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right) \quad \text{M-step} \\
= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t|x_t)} \right) \quad \text{E-step} \\
= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log P_{\theta^{(i)}}(x_t)
$$
Why should EM work?

Steps to show that $\log \text{Lik}(\theta^{(i)}) \geq \log \text{Lik}(\theta^{(i-1)})$:

\[
\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)
\]

\[
\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i-1)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)
\]

\[
E\text{-step}
\]

\[
= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i-1)}}(x_t, c_t)}{P_{\theta^{(i-1)}}(c_t|x_t)} \right)
\]

\[
= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log P_{\theta^{(i)}}(x_t)
\]

\[
= \sum_{t=1}^{n} \log P_{\theta^{(i)}}(x_t)
\]

M-step
Mixture of Multinomials
Mixture of Multinomials

\[
\begin{pmatrix}
10 & 10 & 5 & 2 & 0 & 0 & 0 & 0 & 0 & 5 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 10 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
20 & 15 & 10 & 5 & 0 & 0 & 0 & 0 & 0 \\
10 & 5 & 5 & 2 & 1 & 1 & 1 & 1 & 1 & 5 \\
\end{pmatrix}
\]
Mixture of Multinomials

K buyer types
Each type: distribution over products

```
10 10 5 2 0 0 0 0 0 5
1 0 0 1 0 0 0 1 10
0 0 0 0 1 1 0 0 0
20 15 10 5 0 0 0 0 0
10 5 5 2 1 1 1 1 1 5
```
Mixture of Multinomials

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>15</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mixture of K multinomials
Mixture of Multinomials

10 10 5 2 0 0 0 0 5
1 0 0 1 0 0 0 0 1 10
0 0 0 0 1 1 0 0 0
20 15 10 5 0 0 0 0 0
10 5 5 2 1 1 1 1 1 5
Mixture of Multinomials

\[ \pi = \begin{array}{ccc}
\text{Party!} & \text{HOME} & \text{work} \\
10 & 10 & 5 \\
1 & 0 & 0 \\
0 & 0 & 0 \\
20 & 15 & 10 \\
10 & 5 & 5 \\
\end{array} \]
Mixture of Multinomials

\[ \pi = \begin{array}{c}
\text{Party!} & \text{HOME} & \text{work} \\
10 & 10 & 5 & 2 & 0 & 0 & 0 & 0 & 0 & 5 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 10 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
20 & 15 & 10 & 5 & 0 & 0 & 0 & 0 & 0 \\
10 & 5 & 5 & 2 & 1 & 1 & 1 & 1 & 1 & 5
\end{array} \]
Mixture of Multinomials

\[ \pi = \begin{array}{c}
\text{Party!} & \text{HOME} & \text{work} \\
10 & 10 & 5 \\
2 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 5 \\
\end{array} \]
Mixture of Multinomials

\[ \pi = \frac{10}{10} \frac{5}{5} \frac{2}{2} \]
Mixture of Multinomials

- Eg. Model purchases of each customer

- $K$-types of customers, each designated with distribution over the $d$ items to buy

- Generative model:
  - $\pi$ is mixture distribution over the $K$-types of buyers
  - $p_1, \ldots, p_K$ are the $K$ distributions over the $d$ items, one for each customer type
  - Generative process, each round draw customer type $c_t \sim \pi$
  - Next given $c_t$ draw list of purchases as $x_t \sim \text{multinomial}(p_{c_t})$
Multinomial Distribution

\[ P(x|p) = \frac{m!}{x[1]! \cdots x[d]!} p[1]^{x[1]} \cdots p[d]^{x[d]} \]

Probability of purchase vector \( x \) while drawing products independently \( m \) times from \( p \)
E-step

\[ Q_t^{(i)}(c_t) \propto P(x_t | c_t, \theta^{(i-1)}) P(c_t | \theta^{(i-1)}) \]
E-step

\[
Q_t^{(i)}(c_t) \propto P(x_t|c_t, \theta^{(i-1)})P(c_t|\theta^{(i-1)})
\]

\[
= \frac{P(x_t|p_{ct}^{(i-1)})\pi^{(i-1)}(c_t)}{\sum_{k=1}^{K} P(x_t|p_{k}^{(i-1)})\pi^{(i-1)}(k)}
\]
E-step

\[ Q_t^{(i)}(c_t) \propto P(x_t|c_t, \theta^{(i-1)}) P(c_t|\theta^{(i-1)}) \]

\[ = \frac{P(x_t|p_{c_t}^{(i-1)}) \pi^{(i-1)}(c_t)}{\sum_{k=1}^{K} P(x_t|p_{k}^{(i-1)}) \pi^{(i-1)}(k)} \]

\[ = \frac{p_{c_t[1]} x_{t[1]} \cdots p_{c_t[d]} x_{t[d]} \cdot \pi^{(i-1)}}{\sum_{k=1}^{K} p_{k[1]} x_{t[1]} \cdots p_{c_t[d]} x_{t[d]} \cdot \pi^{(i-1)}_k} \]
M-step

\[ \theta^{(i)} = \arg\max_{\theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log (P(x_t|c_t = k, \theta)P(c_t = k|\theta)) \]
\[ \theta^{(i)} = \arg\max_{\theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log (P(x_t | c_t = k, \theta) P(c_t = k | \theta)) \]

\[ = \arg\max_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left( \frac{m!}{x_t[1]! \cdots x_t[d]!} p_k[1]^{x_t[1]} \cdots p_k[d]^{x_t[d]} \right) \right\} \]

\[ + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \]
M-step

\[ \theta^{(i)} = \arg\max_{\theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log (P(x_t|c_t = k, \theta) P(c_t = k|\theta)) \]

\[ = \arg\max_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left( \frac{m!}{x_t[1]! \cdots x_t[d]!} p_k[1]^{x_t[1]} \cdots p_k[d]^{x_t[d]} \right) \right\} \]

\[ + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \}

\[ = \arg\max_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left( p_k[1]^{x_t[1]} \cdots p_k[d]^{x_t[d]} \right) \right\} \]

\[ + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \} \]
$\theta^{(i)} = \arg\max_{\theta} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log (P(x_t|c_t = k, \theta) P(c_t = k|\theta))$

$= \arg\max_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left( \frac{m!}{x_t[1]! \cdots x_t[d]!} p_{k[1]} x_{t[1]} \cdots p_{k[d]} x_{t[d]} \right) \right\}$

$= \arg\max_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left( p_{k[1]} x_{t[1]} \cdots p_{k[d]} x_{t[d]} \right) \right\}$

$= \arg\max_{\pi, p_1, \ldots, p_K} \left\{ \sum_{t=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{d} Q_t^{(i)}(k) x_{t[j]} \log (p_{k[j]}) + \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \pi_k \right\}$
M-step

\[
\pi^{(i)}_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{n}
\]

\[
p_k[j] = \frac{\sum_{t=1}^{n} x_t[j] Q_t^{(i)}(k)}{m \sum_{t=1}^{n} Q_t^{(i)}(k)}
\]
M-step

\[ \pi_k^{(i)} = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{n} \]

proportion of weights for each type

\[ p_k[j] = \frac{\sum_{t=1}^{n} x_t[j] Q_t^{(i)}(k)}{m \sum_{t=1}^{n} Q_t^{(i)}(k)} \]

weighted number of jth product
What is missing in this story?
What is missing in this story?

```
| 10 | 10 | 5 | 2 | 0 | 0 | 0 | 0 | 5 |
| 1  | 0  | 0 | 1 | 0 | 0 | 0 | 1 | 10|
| 0  | 0  | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 20 | 15 | 10| 5 | 0 | 0 | 0 | 0 | 0 |
| 10 | 5  | 5 | 2 | 1 | 1 | 1 | 1 | 5 |
```
Mixture of Multinomials

What is missing in this story?

What is missing in this story?

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
What is missing in this story?

![Diagram with numbers and items]

Everyone is a bit of party and a bit of work!
Generative story:
For $t = 1$ to $n$
    For each customer draw mixture of types $\pi_t$
    For $i = 1$ to $m$
        For each item to purchase, first draw type $c_t[i] \sim \pi_t$
        Next, given the type draw $x_t[i] \sim p_{c_t[i]}$
    End For
End For
Its a distribution over distributions!

Parameters $\alpha_1, \ldots, \alpha_K$ s.t. $\alpha_k > 0$

The density function is given as

$$p(\pi; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k}$$

where $B(\alpha) = \prod_{k=1}^{K} \Gamma(\alpha_k)/\Gamma(\sum_{k=1}^{K} \alpha_k)$
Dirichlet Distribution

Dirichlet(.5,.5,.5)

Dirichlet(1,1,1)

Dirichlet(5,10,8)
Generative story:
For \( t = 1 \) to \( n \)
  For each customer draw mixture of types \( \pi_t \sim \text{Dirchlet}(\alpha) \)
  For \( i = 1 \) to \( m \)
    For each item to purchase, first draw type \( c_{t[i]} \sim \pi_t \)
    Next, given the type draw \( x_{t[i]} \sim p_{c_{t[i]}} \)
  End For
End For

Parameters, \( \alpha \) for the Dirichlet distribution and \( p_1, \ldots, p_K \)
Generative story:

For $t = 1$ to $n$

For each customer draw mixture of types $\pi_t \sim \text{Dirichlet}(\alpha)$

For $i = 1$ to $m$

For each item to purchase, first draw type $c_t[i] \sim \pi_t$

Next, given the type draw $x_t[i] \sim p_{c_t[i]}$

End For

End For

Parameters, $\alpha$ for the Dirichlet distribution and $p_1, \ldots, p_K$
Its a distribution over distributions!

Parameters $\alpha_1, \ldots, \alpha_K$ s.t. $\alpha_k > 0$

The density function is given as

$$p(\pi; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k}$$

where $B(\alpha) = \prod_{k=1}^{K} \Gamma(\alpha_k)/\Gamma(\sum_{k=1}^{K} \alpha_k)$
Dirichlet Distribution

Dirichlet(.5,.5,.5)

Dirichlet(1,1,1)

Dirichlet(5,10,8)
What is the Dirichlet distribution doing?

Say we didn’t have the $\text{Dir}(\alpha)$, and we had one $\pi$ for all customers. Two choices:

1. For each customer $t$ draw customer type $c_t$ from $\pi$ and then draw all products $i$ from 1 to $m$, based on $p_{c_t}$. What is this model?

2. For each customer $t$ and each product $i$ the customer buys, draw $c_t[i] \sim \pi$ and then draw $x_t[i] \sim p_{c_t[i]}$. 

What is the Dirichlet distribution doing?

Next, say we didn’t have $\text{Dir}(\alpha)$ but each customer separate $\pi_t$?
Next, say we didn’t have $\text{Dir}(\alpha)$ but each customer separate $\pi_t$? This model is often called probabilistic latent semantic analysis.
What is the Dirichlet distribution doing?

Next, say we didn’t have Dir(α) but each customer separate π_t? 
- This model is often called probabilistic latent semantic analysis 
- Number of parameters is \( n \), grows with number of customers
What is the Dirichlet distribution doing?

Next, say we didn’t have $\text{Dir}(\alpha)$ but each customer separate $\pi_t$?
- This model is often called probabilistic latent semantic analysis
- Number of parameters is $n$, grows with number of customers
- Since each customer gets her/his own mixture distribution without restriction, model can overfit easily.
Next, say we didn’t have \( \text{Dir}(\alpha) \) but each customer separate \( \pi_t \)?

- This model is often called probabilistic latent semantic analysis
- Number of parameters is \( n \), grows with number of customers
- Since each customer gets her/his own mixture distribution without restriction, model can overfit easily.
- Further, since there are as many \( \pi \)'s as customers, when a new customer walks in there is no way of extending \( \pi_{n+1} \) is any meaningful way to use our model.
Next, say we didn’t have \( \text{Dir}(\alpha) \) but each customer separate \( \pi_t \)?

- This model is often called probabilistic latent semantic analysis.
- Number of parameters is \( n \), grows with number of customers.
- Since each customer gets her/his own mixture distribution without restriction, model can overfit easily.
- Further, since there are as many \( \pi \)'s as customers, when a new customer walks in there is no way of extending \( \pi_{n+1} \) is any meaningful way to use our model.

Dirichlet prior helps us get a model for new, unseen customers.
If we haven’t seen a customer type yet, that's ok.
Generative Story:

For each customer type $k$ from 1 to $K$,
   Draw $p_k \sim \text{Dir}(\beta)$ (smooth $p_k$'s)
End

For each customer $t$ from 1 to $n$
   Draw $\pi_t \sim \text{Dir}(\alpha)$
   For each purchase $i$ from 1 to $m$ for this customer,
      Draw the customer type $c_{t[i]} \sim \pi_t$ for the purchase
      Given customer type, draw the item $x_{t[i]} \sim p_{c_{t[i]}}$ purchased
End
End

Parameters: $\alpha$ a K-dimensional vector and $\beta$ a d-dimensional vector.
Say $z_1, \ldots, z_n$ are Latent variables. Eg. cluster assignments

- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

  (E step) For every $t$, define distribution $Q_t$ over the latent variable $c_t$ as:
  
  $$Q_t^{(i)}(z_t) = P(z_t | x_t, \theta^{(i-1)})$$

  (M step)

  $$\theta^{(i)} = \arg\max_{\theta \in \Theta} \sum_{t=1}^n \sum_{z_t} Q_t^{(i)}(z_t) \log P(x_t, z_t | \theta) \quad \text{if MLE}$$
Say $z_1, \ldots, z_n$ are Latent variables. Eg. cluster assignments

- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

  (E step) For every $t$, define distribution $Q_t$ over the latent variable $c_t$ as:

  $$Q_t^{(i)}(z_t) = P(z_t|x_t, \theta^{(i-1)})$$

  (M step)

  $$\theta^{(i)} = \arg\max_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{z_t} Q_t^{(i)}(z_t) \log P(x_t, z_t|\theta) \quad \text{if MLE}$$

Latent variables $c_t[i]$’s, $p_k$’s and $\pi_t$’s.
EM Algorithm for LDA
EM Algorithm for LDA

- There are infinite possibilities for $\pi'_t s$ and $p'_k s$
EM Algorithm for LDA

• There are infinite possibilities for $\pi'_t$ and $p'_k$.

• Only think of $c_t[i]'s$ as latent variables.
EM Algorithm for LDA

• There are infinite possibilities for $\pi_t$'s and $p_k$'s
• Only think of $c_t[i]$'s as latent variables
• E-step becomes intractable!
EM Algorithm for LDA

- There are infinite possibilities for $\pi'_t$ and $p'_k$.
- Only think of $c_t[i]'s$ as latent variables.
- E-step becomes intractable!
- Use approximate E-step (Variational approximation).
EM Algorithm for LDA

- There are infinite possibilities for $\pi'_t$'s and $p'_k$'s
- Only think of $c_t[i]'s$ as latent variables
- E-step becomes intractable!
- Use approximate E-step (Variational approximation)
- M-step involves convex optimization
What was common between the various mixture models?