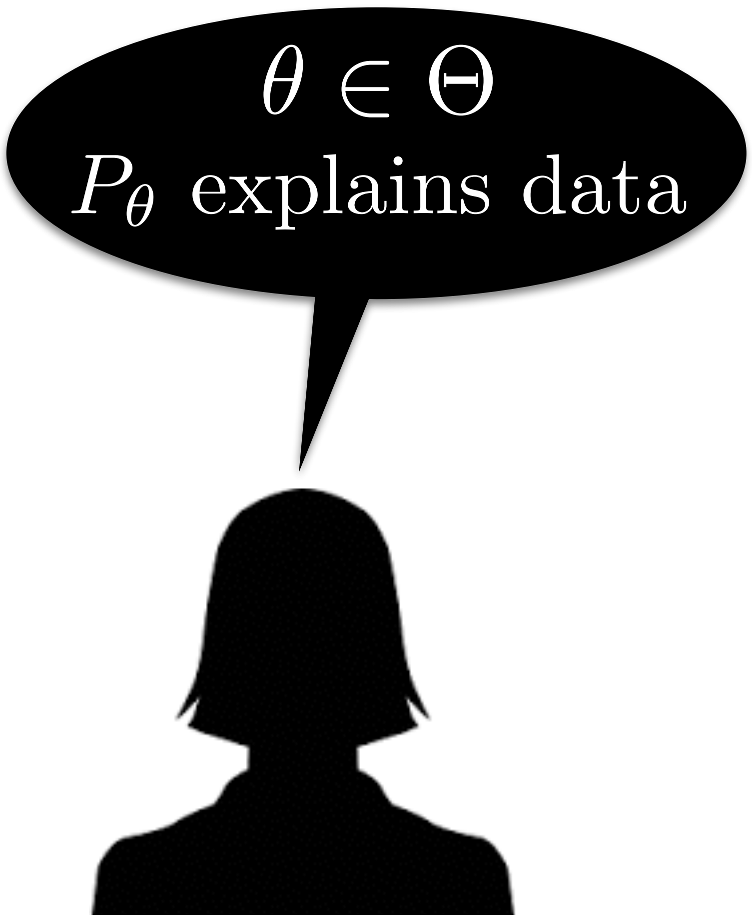
Machine Learning for Data Science (CS4786)
Lecture 16

Latent Dirchlet Allocation

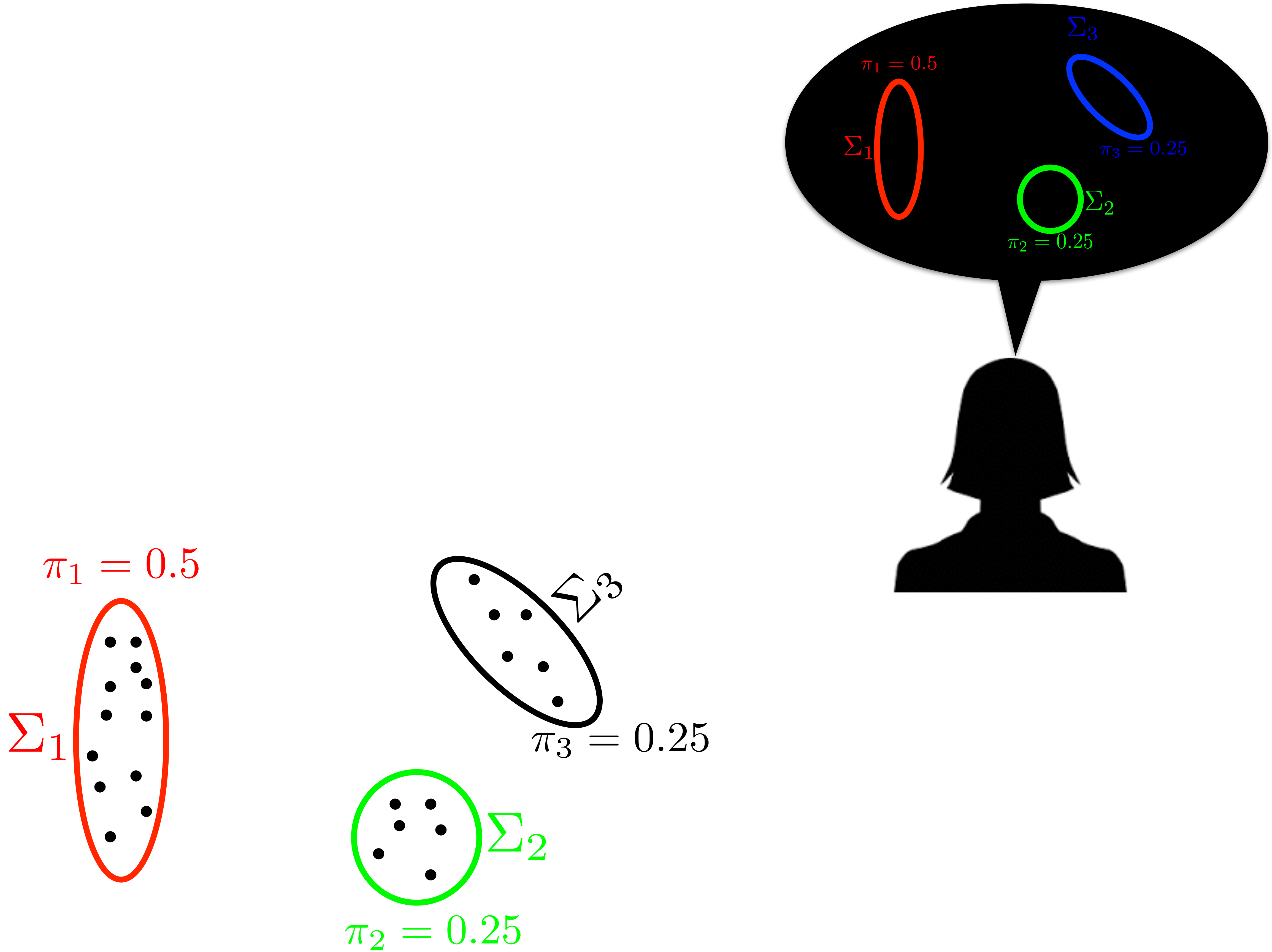
Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016fa/

Probabilistic Model



Data: $\mathbf{x}_1, \ldots, \mathbf{x}_n$

Probabilistic Model



Probabilistic Models

- Set of models Θ consists of parameters s.t. P_θ 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

$$\theta_{MLE} = \mathrm{argmax}_{\theta \in \Theta} \underbrace{\log P_\theta(x_1, \ldots, x_n)}_{\text{Likelihood}}$$

- 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

$$\theta_{MAP} = \mathrm{argmax}_{\theta \in \Theta} P(\theta | x_1, \ldots, x_n)$$

$$= \mathrm{argmax}_{\theta \in \Theta} \; \log P(x_1, \ldots, x_n | \theta) + \log P(\theta)$$

Expectation Maximization Algorithm

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:

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

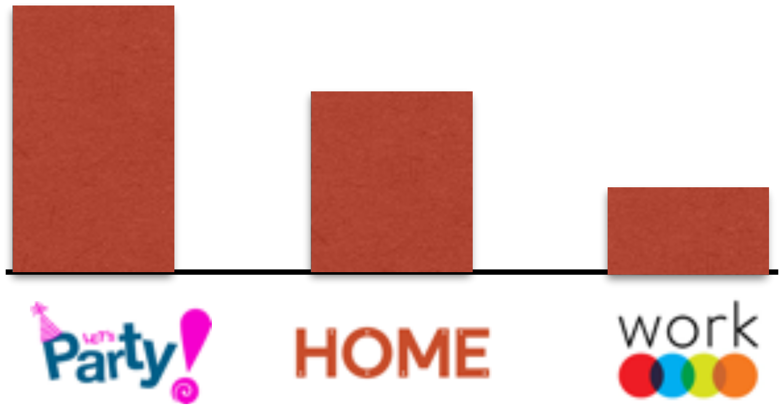
(M step)

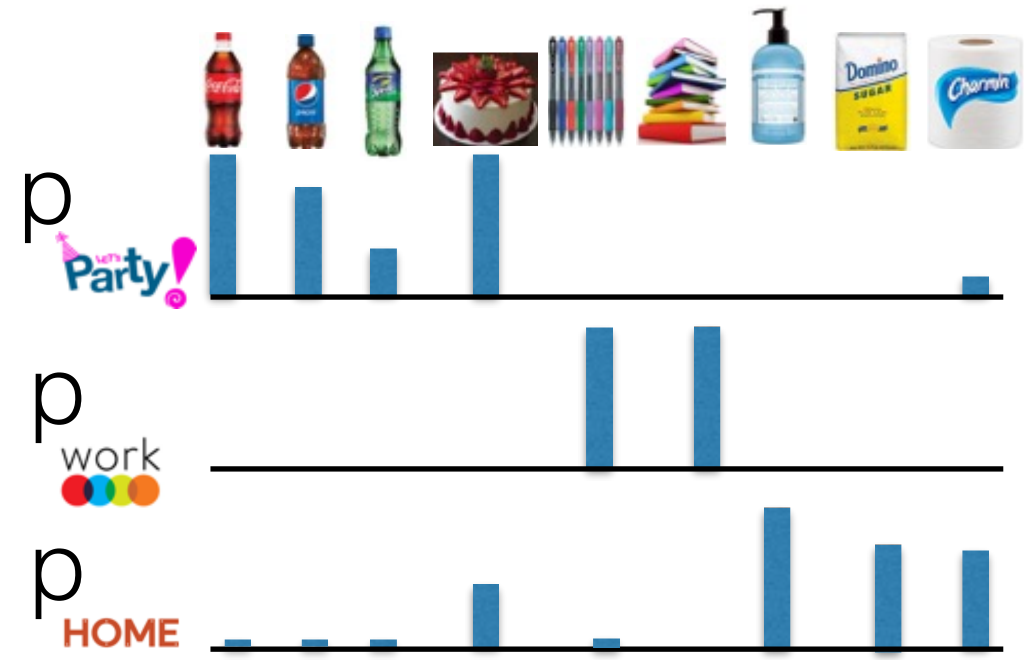
$$\theta^{(i)} = \mathrm{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t | \theta) \qquad \text{if MLE}$$

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

Mixture of Multinomials

$\pi =$





π ~									
Party! ~	10	10	5	2	0	0	0	0	5
HOME ~	1	0	0	1	0	0	0	1	10
work ~	0	0	0	0	1	1	0	0	0
Party! ~	20	15	10	5	0	0	0	0	0
work ~	10	5	5	2	1	1	1	1	5

Multinomial Distribution

$$P(x|p) = \frac{m!}{x[1]! \cdot \ldots \cdot x[d]!} p[1]^{x_t[1]} \cdot \ldots \cdot p[d]^{x_t[d]}$$

Probability of purchase vector x while drawing products independently m times from p

Mixture of Multinomials

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

Everyone is a bit of party and a bit of work!

LATENT DIRICHLET ALLOCATION

- 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, α for the Dirichlet distribution and $p_1, \ldots, p_K$

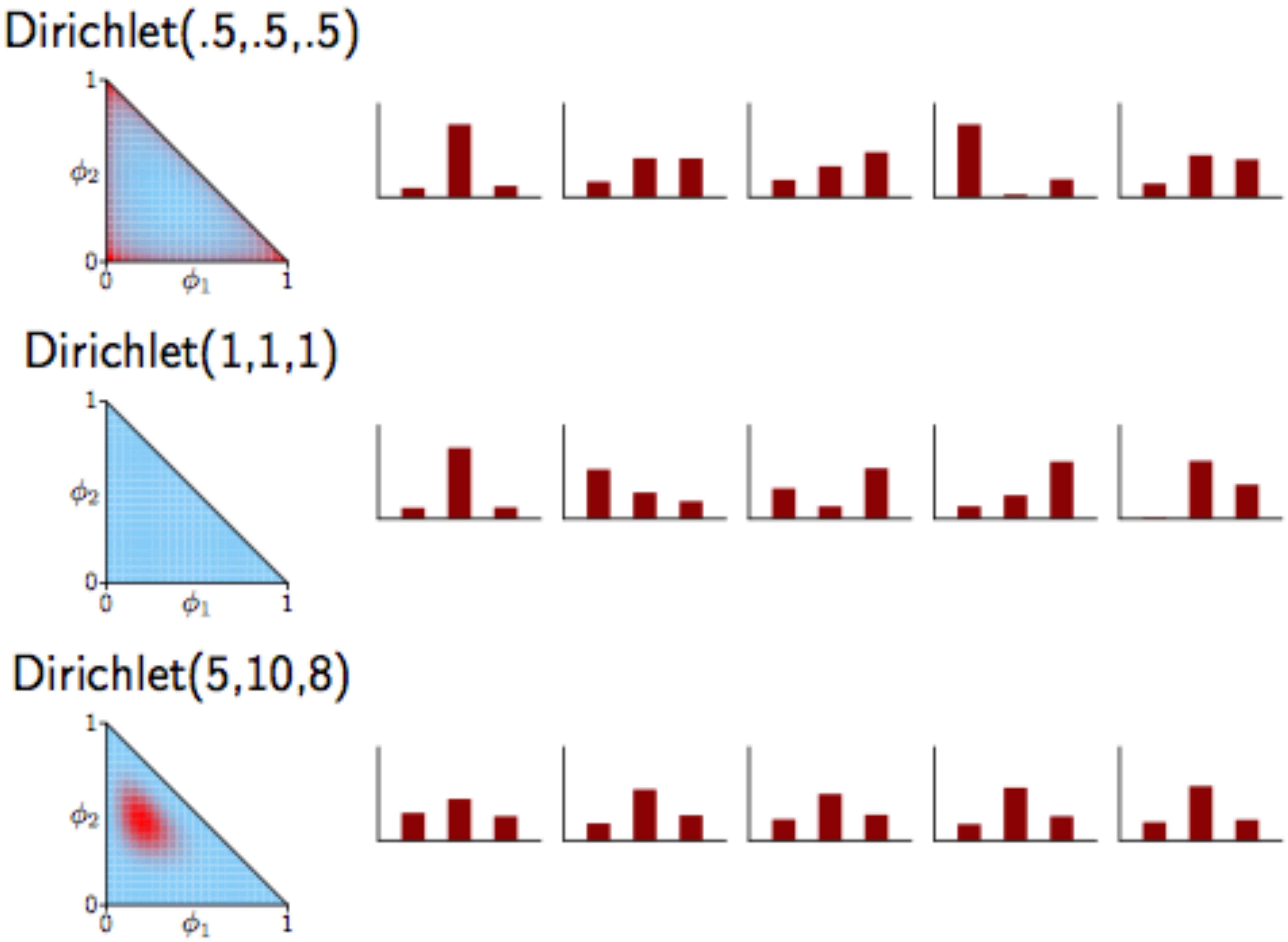
Dirichlet Distribution

- 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



WHAT IS THE DIRICHLET DISTRIBUTION DOING?

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

  1. For each customer t draw customer type c_t from π 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 π_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 π's as customers, when a new customer walks in there is no way of extending π_{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, thats ok.

A Refined Generative Story

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: α a K-dimensional vector and β a d-dimensional vector.

Expectation Maximization Algorithm

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)

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

Latent variables $c_t[i]$'s, p_k's and π_t's.

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?

Graphical Models

- Abstract away the parameterization specifics
- Focus on relationship between random variables

Graphical Models

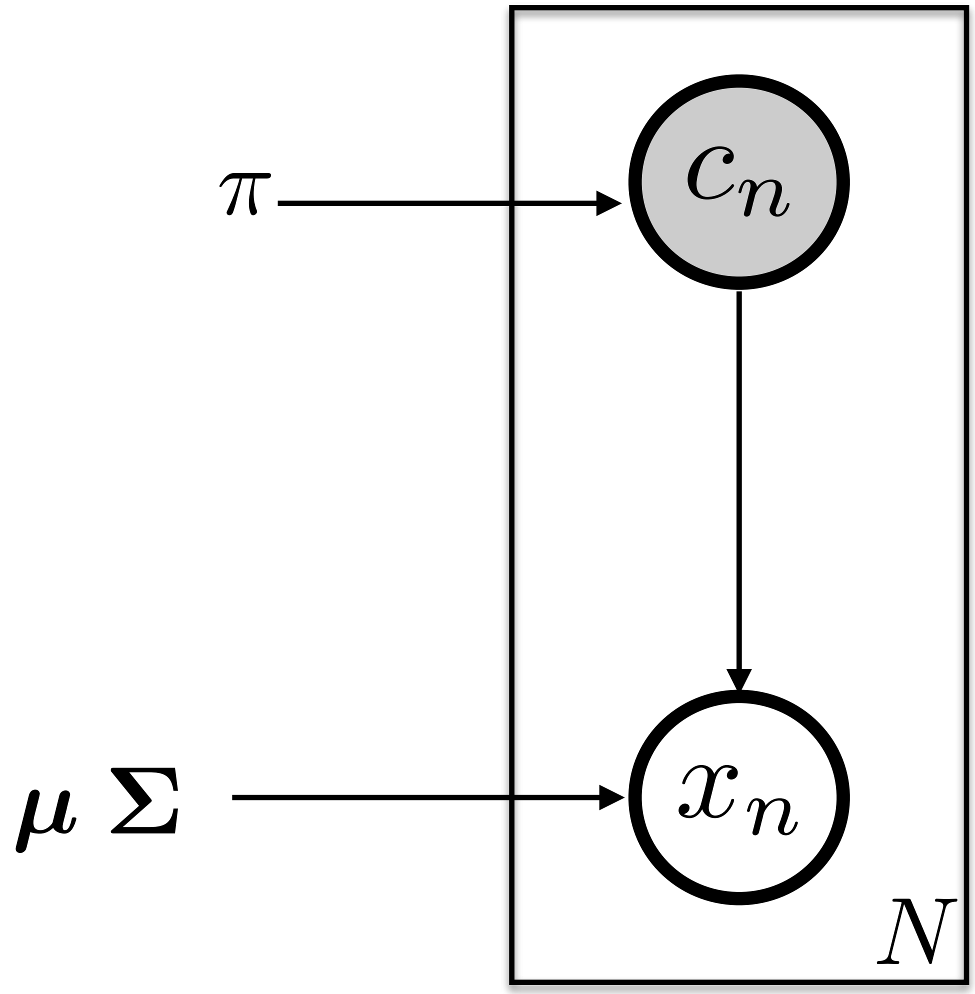
- A graph whose nodes are variables $X_1, \ldots, X_N$
- Graphs are an intuitive way of representing relationships between large number of variables
- Allows us to abstract out the parametric form that depends on θ and the basic relationship between the random variables.

Graphical Models

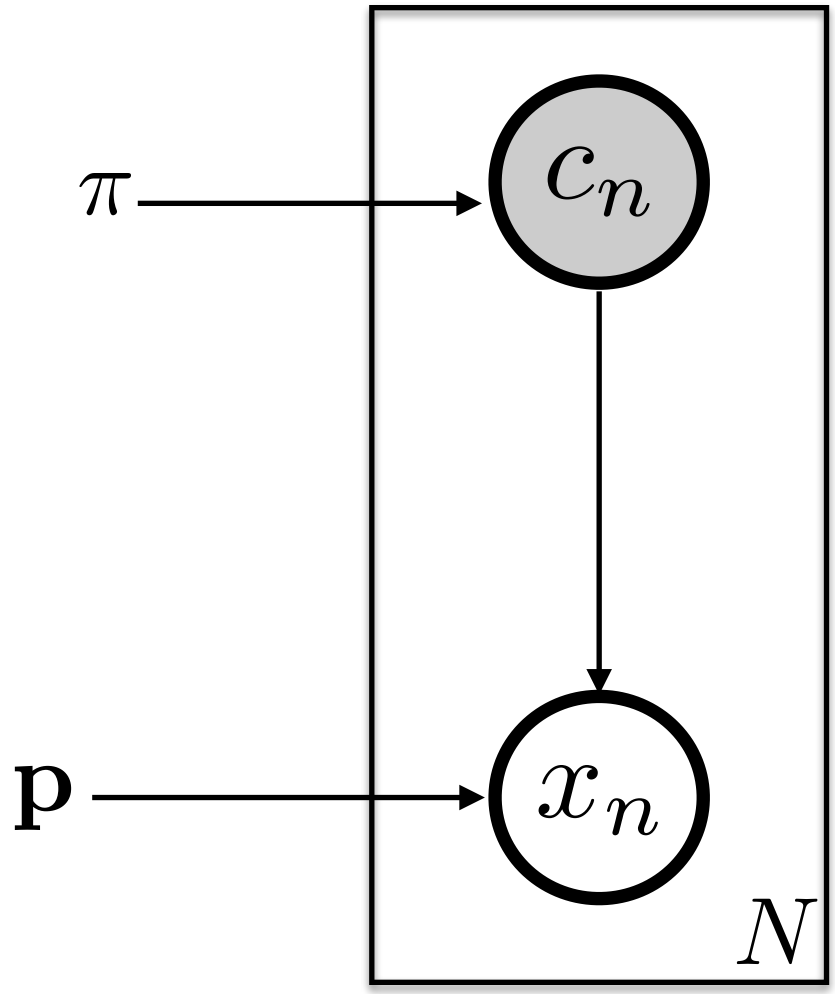
- A graph whose nodes are variables $X_1, \ldots, X_N$
- Graphs are an intuitive way of representing relationships between large number of variables
- Allows us to abstract out the parametric form that depends on θ and the basic relationship between the random variables.

Draw a picture for the generative story
that explains what generates what.

Gaussian Mixture Model



Mixture of Multinomials



Example: Latent Dirichlet Allocation