Nonparametric Bayesian Models
--Learning and Reasoning in Open Possible Worlds

Eric Xing
epxing@cs.cmu.edu
Machine Learning Dept./Language Technology Inst./Computer Science Dept.
Carnegie Mellon University

Outline

• Motivation and challenge

• Dirichlet Process and Infinite Mixture
  • Formulation
  • Approximate Inference algorithm
  • Example: population clustering

• Hierarchical Dirichlet Process and Multi-Task Clustering
  • Formulation
  • Transformed DP and HDP
  • Kernel stick-breaking process
  • Application: joint image segmentation

• Dynamic Dirichlet Process
  • Hidden Markov DP
  • Temporal DPM
  • Application: evolutionary clustering of documents

• Summary
Clustering

Image Segmentation

- How to segment images?
  - Manual segmentation (very expensive)
  - Algorithm segmentation
    - K-means
    - Statistical mixture models
    - Spectral clustering

- Problems with most existing algorithms
  - Ignore the spatial information
  - Perform the segmentation one image at a time
  - Need to specify the number of segments \textit{a priori}
Discover Object Categories

- Discover what objects are present in a collection of images in an unsupervised way
- Find those same objects in novel images
- Determine what local image features correspond to what objects; segmenting the image

Learn and Recognize Natural Scene Categories
Object Recognition and Tracking

\[(1.8, 7.4, 2.3) \rightarrow (1.9, 9.0, 2.1) \rightarrow (1.9, 6.1, 2.2) \rightarrow (0.7, 5.1, 3.2) \rightarrow (0.6, 5.9, 3.2)\]

\[t=1 \quad t=2 \quad t=3\]

The Evolution of Science
A Classical Approach

- Clustering as Mixture Modeling

Then "model selection"

Partially Observed, Open and Evolving Possible Worlds

- Unbounded # of objects/trajectories
- Changing attributes
- Birth/death, merge/split
- Relational ambiguity

The parametric paradigm:

- Finite
- Structurally unambiguous

How to open it up?
Model Selection vs. Posterior Inference

• Model selection
  • "intelligent" guess: ???
  • cross validation: data-hungry
  • information theoretic:
    - AIC
    - TIC
    - MDL: Parsimony, Ockam's Razor
  • Bayes factor: need to compute data likelihood

• Posterior inference:
  we want to handle uncertainty of model complexity explicitly
  \[
  p(M|D) \propto p(D|M) p(M)
  \]
  \[
  \hat{M} = \arg\min_{\theta, K} KL(f(\theta, K) | g(\theta_{ML}, K))
  \]
  • we favor a distribution that does not constrain \(M\) in a "closed" space!

Two "Recent" Developments

• First order probabilistic languages (FOPLs)
  • Examples: PRM, BLOG ...
  • Lift graphical models to "open" world (#rv, relation, index, lifespan …)
  • Focus on complete, consistent, and operating rules to instantiate possible worlds, and formal language of expressing such rules
  • Operational way of defining distributions over possible worlds, via sampling methods

• Bayesian Nonparametrics
  • Examples: Dirichlet processes, stick-breaking processes ...
  • From finite, to infinite mixture, to more complex constructions (hierarchies, spatial/temporal sequences, …)
  • Focus on the laws and behaviors of both the generative formalisms and resulting distributions
  • Often offer explicit expression of distributions, and expose the structure of the distributions --- motivate various approximate schemes
Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Transformed DP and HDP
  - Kernel stick-breaking process
  - Application: joint image segmentation
- Dynamic Dirichlet Process
  - Hidden Markov DP
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary
Random Partition of Probability Space

Stick-breaking Process

\[ G \sim \text{DP}(\alpha, G_0) \]

\[ G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k) \]

\[ \theta_k \sim G_0 \]

\[ \sum_{k=1}^{\infty} \pi_k = 1 \]

\[ \pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \]

\[ \beta_k \sim \text{Beta}(1, \alpha) \]

\[ \Pi_{j=1}^{k-1} (1 - \beta_j) \]

<table>
<thead>
<tr>
<th>Location</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8</td>
</tr>
</tbody>
</table>
**DP – a Pólya urn Process**

$$p = \frac{2}{5+\alpha}$$

$$p = \frac{3}{5+\alpha}$$

$$p = \frac{\alpha}{5+\alpha}$$

$$G_0 := p(\bullet \bullet \bullet \ldots)$$

**Joint:**

$$G(\mathcal{X}) \sim DP(\alpha G_0)$$

**Marginal:**

$$\phi_i | \phi, \alpha, G_0 \sim \sum_{i=1}^{\mathcal{K}} \frac{n_i}{\mathcal{K}-1+\alpha} \delta_{\phi_i} + \frac{\alpha}{\mathcal{K}-1+\alpha} G_0.$$  

- Self-reinforcing property
- Exchangeable partition of samples

---

**Clustering and DP Mixture**

$$p = \frac{2}{5+\alpha}$$

$$p = \frac{3}{5+\alpha}$$

$$p = \frac{\alpha}{5+\alpha}$$

$$G_0 := p(\bullet \bullet \bullet \ldots)$$

- We can associate mixture components with colors in the Pólya urn model and thereby define a *clustering* of the data
Chinese Restaurant Process

\[ P(c_i = k \mid c_{-i}) = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{1 + \alpha} & \frac{\alpha}{1 + \alpha} & 0 \\ \frac{1}{2 + \alpha} & \frac{1}{2 + \alpha} & \frac{\alpha}{2 + \alpha} \\ \frac{1}{3 + \alpha} & \frac{2}{3 + \alpha} & \frac{\alpha}{3 + \alpha} \\ \frac{m_i}{i + \alpha - 1} & \frac{m_i}{i + \alpha - 1} & \frac{\alpha}{i + \alpha - 1} \end{bmatrix} \]

Dirichlet Process

- A CDF, \( G \), on possible worlds of random partitions follows a Dirichlet Process if for any measurable finite partition \( (\phi_1, \phi_2, \ldots, \phi_m) \):

\[ (G(\phi_1), G(\phi_2), \ldots, G(\phi_m)) \sim \text{Dirichlet}(\alpha G_0(\phi_1), \ldots, \alpha G_0(\phi_m)) \]

where \( G_0 \) is the base measure and \( \alpha \) is the scale parameter.

Thus a Dirichlet Process \( G \) defines a distribution of distribution...
Graphical Model
Representations of DP

The Pólya urn construction

The Stick-breaking construction

Example: DP-haplotyper [Xing et al, 2004]

- Clustering human populations

- Inference: Markov Chain Monte Carlo (MCMC)
  - Gibbs sampling
  - Metropolis Hasting
Inheritance and Observation Models

- **Single-locus mutation model**
  \[ A_{C_i} \rightarrow H_i \]
  
  \[ P_a(h_i \mid a_i, \theta) = \begin{cases} 
  \theta & \text{for } h_i = a_i \\
  1 - \theta & \text{for } h_i \neq a_i 
  \end{cases} \]

- **Noisy observation model**
  \[ H_i, h_i \rightarrow G_i \]
  
  \[ P_o(g \mid h_1, h_2) = g_i = h_i \oplus h_j \text{ with prob. } \lambda \]

MCMC for Haplotype Inference

- **Gibbs sampling for exploring the posterior distribution under the proposed model**
  - Integrate out the parameters such as \( \theta \) or \( \lambda \), and sample \( c_i \), \( a_k \) and \( h_i \)

\[ p(c_i = k \mid c_{[-i]}, h, a) \propto p(c_i = k \mid c_{[-i]}, c) p(h_i \mid a_k, h_{[-i]}, c) \]

- **Gibbs sampling algorithm**: draw samples of each random variable to be sampled given values of all the remaining variables
MCMC for Haplotype Inference

1. Sample \( c_{ik}^{(j)} \) from
\[
p(c_{ik}^{(j)} = k | d_{i,j}^{(j)}, h, a) = \frac{p(c_{ik}^{(j)} = k | d_{i,j}^{(j)}, h, a)}{\sum_k p(c_{ik}^{(j)} = k | d_{i,j}^{(j)}, h, a)}
\]
\[
\propto p(c_{ik}^{(j)} = k | d_{i,j}^{(j)}, m, n) p(h_{ik}^{(j)} | a_k, c, h^{(j-1)})
\]
\[
\propto (m_k^{(j-1)} + \tau \beta_k) p(h_{ik}^{(j)} | a_k, l_k^{(j-1)}), \text{ for } k = 1, \ldots, K + 1
\]

2. Sample \( a_k \) from
\[
p(a_k, l_k | c, h) \propto \prod_{j, i} p(h_{ik}^{(j)} | a_k, l_k, t_{ik})
\]
\[
= \frac{\Gamma(\alpha_h + l_{ik}) \Gamma(\beta_h + t_{ik})}{\Gamma(\alpha_h + \beta_h + m_k)(|B| - 1)_{ik}} R(\alpha_h, \beta_h)
\]

3. Sample \( h_{ik}^{(j)} \) from
\[
p(h_{ik}^{(j)} | h_{ik}^{(j-1)}, c, a, g)
\]

- For DP scale parameter \( \alpha \): a vague inverse Gamma prior

Convergence of Ancestral Inference

- Gibbs sampling solution is not efficient enough to scale up to the large scale problems.
- Truncated stick-breaking approximation can be formulated in the space of explicit, non-exchangeable cluster labels.
- Variational inference can now be applied to such a finite-dimensional distribution

- Variational Inference:
  - For a complicated \( P(X_1, X_2, \ldots, X_n) \), approximate it with \( Q(\lambda) \):

\[
Q(\mathbf{X}) = \prod_i Q(\mathbf{X}_{C_i})
\]

\[
\{ Q^*(\mathbf{X}_{C_i}) \} = \arg \min K L(Q(\mathbf{X})|P(\mathbf{X}))
\]
Approximations to DP

- Truncated stick-breaking representation

\[ \begin{align*}
\alpha_i &\sim \beta(\alpha_0; 1, \alpha) \\
\pi^* &\equiv 1 \\
\pi_i &\equiv \alpha_0 \prod_{j=1}^{i-1} (1 - \alpha_j) \\
\pi_i &\equiv 0 \quad \text{for } i > T
\end{align*} \]

- Finite symmetric Dirichlet approximation

\[ \pi \sim \text{Dir}(\pi; \frac{\alpha}{N}, \ldots, \frac{\alpha}{N}) \]

The joint distribution can be expressed as:

\[ P(X, z, \pi, q) = \prod_{i=1}^{K} \prod_{l=1}^{n_i} p(c_l | \pi_i) p(z_l | c_l, \pi_i) \prod_{l=1}^{T} p(q_l | \pi_i, 1, \alpha) \]

\[ P(X, z, \pi, q) = \prod_{i=1}^{K} \prod_{l=1}^{n_i} p(c_l | \pi_i) p(z_l | c_l, \pi_i) \prod_{l=1}^{T} p(q_l | \pi_i, \frac{\alpha}{N}, \ldots, \frac{\alpha}{N}) \]

TDP vs. TSB

- TDP is size biased
- Cluster labels is NOT interchangeable under TDP but is interchangeable under TSB
**Marginalization**

- In variational Bayesian approximation, we assume a factorized form for the posterior distribution.
- However, it is not a good assumption since changes in $\pi$ will have a considerable impact on $z$.

If we can integrate out $\pi$, the joint distribution is given by

$$P(X, z, \eta) = \left[ \prod_{i=1}^{N} p(x_i | y_{i,z_i}) \right] p(z) \prod_{i=1}^{N} p(\eta_i)$$

For the TSB representation:

$$p_{\text{TSB}}(x) = \prod_{i \in T} \frac{\Gamma(1 + N_i) \Gamma(\alpha + N_{z_i})}{\Gamma(1 + \alpha + N_{z_i})}$$

For the FSD representation:

$$p_{\text{FSD}}(x) = \frac{\prod_{i=1}^{N} \Gamma(\alpha) \Gamma(N + \frac{\alpha}{2})}{\Gamma(N + \alpha) \Gamma(\frac{\alpha}{2})}$$

**VB inference**

- We can then apply the VB inference on the four approximations

$$\{Q^*(X_{C_i})\} = \arg \min KL(Q(X)||P(X))$$

The approximated posterior distribution for TSB and FSD are

$$Q_{\text{TSB}}(x, \eta, \nu) = \left[ \prod_{i=1}^{N} q(z_i) \right] \left[ \prod_{i=1}^{T} q(\eta_i) q(\nu_i) \right]$$

$$Q_{\text{FSD}}(x, \eta, \pi) = \left[ \prod_{i=1}^{N} q(z_i) \right] \left[ \prod_{i=1}^{K} q(\eta_i) \right] q(\pi)$$

Depending on marginalization or not, $\nu$ and $\pi$ may be integrated out.
Experimental results

Figure 2: Average log probability per data point for test data as a function of N.

Figure 4: Average log probability per data point for test data as a function of T or TSN methods or T or FSV methods.

Outline

- Motivation and challenge

- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering

- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Transformed DP and HDP
  - Kernel stick-breaking process
  - Application: joint image segmentation

- Dynamic Dirichlet Process
  - Hidden Markov DP
  - Temporal DPM
  - Application: evolutionary clustering of documents

- Summary
Solving Multiple Clustering Problems

- Solve separately
  - Fail to capture correlation
  - Fail to cross-reinforce shared information (i.e., topic specific lexicon)
  - Data fragmentation

- Solve together
  - Then what is the difference between all these journals?
Hierarchical Dirichlet Process

- Two level Pólya urn scheme
  - At the $i$-th step in $j$-th "group",

\[ \theta_i | \theta_i \sim \frac{\sum_{k} n_k}{\sum_{k} n_k + \gamma} \delta_{\theta_i} + \frac{\gamma}{\sum_{k} n_k + \gamma} H(\theta) \]

- Conditioning on $DP(\gamma, H)$, the $m$th draw from the $m$th bottom-level urn also form a Dirichlet measure

\[ \theta_m | \theta_m \sim \frac{\sum_{k} m_k + a}{m_k + a} \delta_{\theta_m} + \frac{a}{m_k + a} \sum_{k} \frac{m_k + a}{i + \gamma} H(\theta_k) \]

\[ + \sum_{k} \frac{m_k + a}{m_k + a + i + \gamma} p_{i,j} \delta_{\mu_{i,j}}(\theta_k) + \frac{a}{m_k + a + i + \gamma} H(\theta_k) \]

\[ + \sum_{k} \frac{m_k + a}{m_k + a + i + \gamma} p_{i,j} \delta_{\mu_{i,j}}(\theta_k) + \frac{a}{m_k + a + i + \gamma} H(\theta_k) \]
Recall: Graphical Model Representations of DP

The Pólya urn construction

The Stick-breaking construction

Hierarchical DP Mixture

$\text{Stick}(\alpha, \beta):$

$\pi'_j = \text{Beta}(\alpha_j, \beta_j) + \sum_i \pi_i \left(1 - \pi'_j\right)$
Topic Models for Images

Latent Dirichlet Allocation (LDA)

"beach"

Image Representation

representation vector \( [r_{11}, \ldots, r_{1d}] \):
annotation vector \( [w_1, \ldots, w_{|V|}] \):

(real, 1 per image segment)
(binary, same for each segment)

\[ [r_{n1}, \ldots, r_{nd}] , [w_1, \ldots, w_{|V|}] \]
Infinite Topic Model for Image

A single image with \( k \) topic

An LDA

A single image with inf-topic

A DP

\( J \) images with inf-topic

An HDP

Problem with HDP

- Every group (i.e., image) has exactly the same set of visual-vocabulary topics, albeit with different frequency
Transformed Dirichlet Process
[Sudderth et al, 2005]

- An extension of HDP in which global mixture components undergo a set of random transformations before being reused in each group.

![Diagram of transformed Dirichlet Process](image)

Synthetic Data Results

![Synthetic Data Results](image)

Figure 3: Comparison of hierarchical models learned via Gibbs sampling from synthetic 2D data. *Left:* Four of 50 “images” used for training. *Center:* Global distribution \(G_0(\theta)\) for the HDP, where ellipses are covariance estimates and intensity is proportional to prior probability. *Right:* Global TDP distribution \(G_0(\theta, \rho)\) over both clusters \(\theta\) (solid) and translations \(\rho\) of those clusters (dashed).

- HDP uses a large set of global clusters to discretize the transformations underlying the data, and may have poor generalization for modeling visual scenes.
For image analysis, we want to impose the belief that spatially proximate patches are more probable to be associated with the same cluster.

We augmented the stick-breaking representation of DP to employ a kernel function to quantify some additional prior.
KSBP for image analysis

- Consider an image composed of $N$ patches, the features vectors $\{x_n\}_{n=1}^N$, and the associated locations $\{r_n\}_{n=1}^N$ can be modeled as follows:

$$x_n \sim f(\phi_n)$$

$$\phi_n \sim G_r$$

$$G_r = \sum_{b=1}^{\infty} \pi_b(r; V_b, \Gamma_b, \psi) \delta_{\phi_b}$$

$$\pi_b(r; V_b, \Gamma_b, \psi) = V_b K(r, \Gamma_b, \psi) \prod_{l=1}^{b-1} [1 - V_l K(r, \Gamma_l, \psi)]$$

$$V_b \sim \text{Beta}(a, b)$$

$$\Gamma_b \sim H$$

$$\theta_b \sim G_o$$

---

Multi-task Image Segmentation

[Diagram of multi-task image segmentation]
Segmentation results [An et al, 2008]

Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Transformed DP and HDP
  - Kernel stick-breaking process
  - Application: joint image segmentation
- Dynamic Dirichlet Process
  - Hidden Markov DP
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary
Object Recognition and Tracking

- Each chain corresponds to the trajectory of a specific object

Hidden Markov Dirichlet Process

- Hidden Markov Dirichlet process mixtures
  - Extension of HMM model to infinite ancestral space
    - Infinite dimensional transition matrix
    - Each row of the transition matrix is modeled with a DP: \( G_i | \alpha, G_0 \sim \text{DP}(\alpha, G_0) \)

\[ G_0 | \gamma, H \sim \text{DP}(\gamma, H) \]
HMDP as a Graphical Model

Evolutionary Clustering

- Adapts the number of mixture components over time
- Mixture components can die out
- New mixture components are born at any time
- Retained mixture components parameters evolve according to a Markovian dynamics
The Chinese Restaurant Process

- Customers correspond to data points
- Tables correspond to clusters/mixture components
- Dishes correspond to parameter of the mixtures

Temporal DPM [Ahmed and Xing 2008]

- The Recurrent Chinese Restaurant Process
  - The restaurant operates in epochs
  - The restaurant is closed at the end of each epoch
  - The state of the restaurant at time epoch \( t \) depends on that at time epoch \( t-1 \)
    - Can be extended to higher-order dependencies.
Customers at time $T=1$ are seated as before:
- Choose table $j \propto N_{j,1}$ and Sample $x_i \sim f(\phi_{j,1})$
- Choose a new table $K+1 \propto \alpha$
- Sample $\phi_{K+1,1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1,1})$

Dish eaten at table 3 at time epoch 1
OR the parameters of cluster 3 at time epoch 1

$N_{1,1}=2 \quad N_{2,1}=3 \quad N_{3,1}=1$
\[ \phi_{1,1} \quad \phi_{2,1} \quad \phi_{3,1} \]

\[ \phi_{1,2} \quad \phi_{2,1} \quad \phi_{3,1} \]

\[ N_{1,1}=2 \quad N_{2,1}=3 \quad N_{3,1}=1 \]

\[ \frac{1}{6+\alpha} \quad \frac{\alpha}{6+\alpha} \quad \frac{3}{6+\alpha} \]

And so on ...

At the end of epoch 2

Newly born cluster

Died out cluster
Temporal DPM

- Can be extended to model higher-order dependencies
- Can decay dependencies over time
  - Pseudo-counts for table $k$ at time $t$ is

$$\sum_{w=1}^{W} \left( e^{-w} \frac{-w}{N_{k,t-w}} \right)$$

- History size
- Decay factory
- Number of customers sitting at table $K$ at time epoch $t-w$
\[ \phi_{1,1}, \phi_{2,1}, \phi_{3,1} \]
\[ \phi_{1,2}, \phi_{2,2}, \phi_{3,1}, \phi_{4,2} \]
\[ N_{1,1} = 2, N_{2,1} = 3, N_{3,1} = 1 \]
\[ N_{2,3} = \sum_{w=1}^{W} e^{-x} N_{k,t-w} \]

TDPM Generative Power

DPM
- \( W = T \)
- \( \lambda = \infty \)

TDPM
- \( W = 4 \)
- \( \lambda = .4 \)

Independent DPMs
- \( W = 0 \)
- \( \lambda = ? (\text{any}) \)
Experiments

- **Simulated data**
- Chain dynamics is modeled as random walk
  \[ \phi_{k,t} | \phi_{k,t-1} \sim N(\phi_{k,t-1}, \rho I) \]
- Gaussian emission: \[ x_{t,i} | \phi_{t,i} = k \sim N(\phi_{k,t}, \Sigma) \]
- Simulated 30 epochs with 100 data points in each epoch
- Can TDPM recover the ground truth clustering?
  - Posterior inference ran using Gibbs sampling [Ahmed and Xing 2008]
- Compare with fixed-dimension dynamic models
Results: NIPS 12

- Building a simple dynamic topic model
- Chain dynamics is as before
- Emission model for document $x_{k,t}$ is:
  - Project $\phi_{t}$ over the simplex
  - Sample $x_{k,t} \sim \text{Multinomial}(. | \text{Logistic}(\phi_{t}))$
- Unlike LDA here a document belongs to one topic
- Use this model to analyze NIPS12 corpus
  - Proceeding of NIPS conference 1987-1999
Summary

- A non-parametric Bayesian model for Pattern Uncovery
  - Finite mixture model of latent patterns (e.g., image segments, objects)
  - infinite mixture of prototypes: alternative to model selection
  - hierarchical infinite mixture
  - infinite hidden Markov model
  - temporal infinite mixture model

- Applications in general data-mining …
How to Model Semantic?

- **Q:** What is it about?
- **A:** Mainly MT, with syntax, some learning

### A Hierarchical Phrase-Based Model for Statistical Machine Translation

We present a statistical phrase-based Translation model that uses hierarchical phrases—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the formal machinery of syntax based translation systems without any linguistic commitment. In our experiments using BLEU as a metric, the hierarchical Phrase-based model achieves a relative improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.

<table>
<thead>
<tr>
<th>Mixing Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT: 0.6</td>
</tr>
<tr>
<td>Syntax: 0.3</td>
</tr>
<tr>
<td>Learning: 0.1</td>
</tr>
</tbody>
</table>

### Unigram over vocabulary

<table>
<thead>
<tr>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Target SMT Alignment Score BLEU</td>
</tr>
<tr>
<td>Parse Tree Noun Phrase Grammar CFG</td>
</tr>
<tr>
<td>likelihood EM Hidden Parameters Estimation argMax</td>
</tr>
</tbody>
</table>

### Admixture Models

- **Objects** are bags of elements
- **Mixtures** are distributions over elements

**Objects have mixing vector** $\theta$
- Represents each mixtures’ contributions

**Object is generated as follows:**
- Pick a mixture component from $\theta$
- Pick an element from that component
Topic Models = Admixture Models

Generating a document

- Draw $\theta$ from the prior
  For each word $n$
    - Draw $z_n$ from multinomial $l(\theta)$
    - Draw $w_n | z_n, \{\beta_{zk}\}$ from multinomial $l(\beta_{zk})$

Which prior to use?

Variational Inference

Approximate the Integral

Approximate the Posterior

$P(\gamma, z_{1:n} | D)$

arg min $KL(q||p)$

$\mu^*, \Sigma^*, \phi_{1:n}^*$

Solve

Optimization Problem