Case Study 2: Document Retrieval

MAP EM, Latent Dirichlet Allocation, Gibbs Sampling

Machine Learning/Statistics for Big Data
CSE599C1/STAT592, University of Washington
Emily Fox
February 5th, 2013

©Emily Fox 2013

Gaussian Mixture Model

- Most commonly used mixture model
- Observations: $x^1, \ldots, x^N$
- Parameters: $\theta = \{\pi, \phi\}$
  
  \[
  \pi = [\pi_1, \ldots, \pi_K] \\
  \phi = \{\phi_k\} = \{\mu_k, \Sigma_k\}
  \]
- Likelihood:
  
  \[
  p(x^i | \theta) = \sum_k \pi_k p(x^i | \phi_k)
  \]
- Ex. $z^i = \text{country of origin}$, $x^i = \text{height of } i^{th} \text{ person}$
  
  $K^{th}$ mixture component = distribution of heights in country $k$
Motivates EM Algorithm

- Initial guess: $\hat{\theta}^{(0)}$
- Estimate at iteration $t$: $\hat{\theta}^{(t)}$

**E-Step**
Compute
$$U(\theta, \hat{\theta}^{(t)}) = E[\log p(y \mid \theta) \mid x, \hat{\theta}^{(t)}]$$

**M-Step**
Compute
$$\hat{\theta}^{(t+1)} = \arg \max_{\theta} U(\theta, \hat{\theta}^{(t)})$$

MAP Estimation

- Bayesian approach:
  - Place prior $p(\theta)$ on parameters
  - Infer posterior $p(\theta \mid x)$

- Many, many, many motivations and implications
  - For the sake of this class, simplest motivation is to think of this as akin to regularization

$$\hat{\theta}^{MAP} = \arg \max_{\theta} \log p(\theta \mid x)$$

- Saw importance of regularization in logistic regression (ML estimate can overfit data and lead to poor generalization)
EM Algorithm – MAP Case

- Re-derive EM algorithm for $p(\theta \mid x)$
- Add $\log p(\theta)$ to $U(\theta, \hat{\theta}^{(t)})$
  - What must be computed in E-Step remains unchanged because this term does not depend on $y$.
  - M-Step becomes:
    $$\hat{\theta}^{(t+1)} = \arg \max_{\theta} U(\theta, \hat{\theta}^{(t)})$$

MAP EM Example – MoG

- For mixture of Gaussians, conjugate priors are:
  $$\pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$$

<table>
<thead>
<tr>
<th>$\alpha$ Configuration</th>
<th>$\pi$ Distribution</th>
<th>$\tilde{x}$ Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,0,0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0,0,1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0,1,0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1/2,1/2,0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1/3,1/3,1/3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1/4,1/4,1/2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1/2,1/2,0)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$p(\pi | \alpha, \beta)$

$\alpha = 1.0, \beta = 1.0$
$\alpha = 1.0, \beta = 3.0$
$\alpha = 1.0, \beta = 0.3$
$\alpha = 0.3, \beta = 0.3$

$p(\pi | \alpha, \beta)$

$\alpha = 1.0, \beta = 1.0$
$\alpha = 1.0, \beta = 3.0$
$\alpha = 1.0, \beta = 0.3$
$\alpha = 0.3, \beta = 0.3$
For mixture of Gaussians, conjugate priors are:

\[ \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \quad p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1} \]

**Dirichlet posterior**

- Assume we condition on observations \( z^i \sim \pi \)
- Count occurrences of \( z^i = k \)
- Then,

\[ p(\pi \mid \alpha, z^1, \ldots, z^N) \propto \]

- Conjugacy: This **posterior** has same form as **prior**

---

**MAP EM Example – MoG**

- Results in following M-Step:

\[
\hat{\mu}_k = \frac{r_k \bar{x}_k + \kappa_0 m_0}{r_k + \kappa_0} \quad \hat{\pi}_k = \frac{r_k + \alpha_k - 1}{N + \sum_k \alpha_k - K} \\
\hat{\Sigma}_k = \frac{S_0 + r_k S_k + \frac{\kappa_0 r_k}{\kappa_0 + r_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)'}{\nu_0 + r_k + d + 2}
\]
Posterior Computations

- MAP EM focuses on point estimation:
  \[ \hat{\theta}^{MAP} = \arg \max_{\theta} p(\theta | x) \]

- What if we want a full characterization of the posterior?
  - Maintain a measure of uncertainty
  - Estimators other than posterior mode (different loss functions)
  - Predictive distributions for future observations

- Often no closed-form characterization (e.g., mixture models)
- Alternatives:
  - Monte Carlo based estimates using samples from posterior
  - Variational approximations to posterior (more next time)

Gibbs Sampling

- Want draws:

  - Construct Markov chain whose steady state distribution is

  - Simplest case:
Example – Mixture of Gaussians

- Recall model
  - Observations: $x^1, \ldots, x^N$
  - Cluster indicators: $z^1, \ldots, z^N$
  - Parameters: $\theta = \{\pi, \phi\}$
    \[ \pi = [\pi_1, \ldots, \pi_K], \quad \phi = \{\phi_k\} = \{\mu_k, \Sigma_k\} \]
  - Generative model:
    \[ \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \]
    \[ \{\mu_k, \Sigma_k\} \sim F(\phi) \]
    \[ z^i \sim \pi \]
    \[ x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i}) \]

- Want to draw posterior samples of model parameters
  \[ \pi \sim p(\pi \mid \phi, x^1, \ldots, x^N) \]
  \[ \phi \sim p(\phi \mid \pi, x^1, \ldots, x^N) \]

Auxiliary Variable Samplers

- Augment variables of interest $\theta$ with variables $z$ to allow closed-form for sampling, just like in EM

- In both cases, simply looking at subchain $\{\theta^{(t)}\}$ converges to draws from marginal distribution $\pi(\theta)$
Example – Mixture of Gaussians

\[ \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \]
\[ \{\mu_k, \Sigma_k\} \sim F(\phi) \]
\[ z^i \sim \pi \]
\[ x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i}) \]

- Try auxiliary variable sampler
  - Introduce cluster indicators into sampler

Example – Clustering Results I

\[
\begin{align*}
\log p(x \mid \pi, \theta) &= -539.17 \\
\log p(x \mid \pi, \theta) &= -497.77 \\
\log p(x \mid \pi, \theta) &= -404.18 \\
\log p(x \mid \pi, \theta) &= -454.15 \\
\log p(x \mid \pi, \theta) &= -397.40 \\
\log p(x \mid \pi, \theta) &= -442.89 
\end{align*}
\]

Figure 2.18. Learning a mixture of \( K = 4 \) Gaussians using the Gibbs sampler of Alg. 2.1. Columns show the current parameters after \( T=2 \) (top), \( T=10 \) (middle), and \( T=50 \) (bottom) iterations from two random initializations. Each plot is labeled by the current data log–likelihood.

Figure courtesy of Erik Sudderth
Collapsed Gibbs Samplers

- Marginalize a set of latent variables or parameters
  - Sometimes marginalized variables are nuisance parameters
  - Other times what gets marginalized are the variables
    - Make post-facto inferences on variables of interest based on sampled variables
  - Can improve efficiency if marginalized variables are high-dim
    - Reduced dimension of search space
    - But, often introduces dependences!

Example – Collapsed MoG Sampling

\[
\pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \quad z_i \sim \pi \\
\{\mu_k, \Sigma_k\} \sim F(\phi) \quad x^i | z^i \sim N(x^i; \mu_{z_i}, \Sigma_{z_i})
\]

-Collapsed sampler
Example – Collapsed MoG Sampling

\[ \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \quad z^i \sim \pi \]
\[ \{\mu_k, \Sigma_k\} \sim F(\phi) \quad x^i \mid z^i \sim N(x^i; \mu_{z^i}, \Sigma_{z^i}) \]

- Derivation

- Important facts:

\[
p(z_{1:N} \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k) \prod_k \Gamma(n_k + \alpha_k)}{\prod_k \Gamma(\alpha_k) \Gamma(\sum_k n_k + \alpha_k)} \quad \frac{\Gamma(m+1)}{\Gamma(m)} = m
\]

Example – Clustering Results II

\[
\log p(x \mid \pi, \theta) = -399.06 \quad \log p(x \mid \pi, \theta) = -461.94
\]
\[
\log p(x \mid \pi, \theta) = -397.38 \quad \log p(x \mid \pi, \theta) = -449.23
\]
\[
\log p(x \mid \pi, \theta) = -396.53 \quad \log p(x \mid \pi, \theta) = -448.68
\]

Figure 2.19. Learning a mixture of \( K = 4 \) Gaussians using the Rao–Blackwellized Gibbs sampler of Alg. 2.2. Columns show the current parameters after \( T=2 \) (top), \( T=10 \) (middle), and \( T=50 \) (bottom) iterations from two random initializations. Each plot is labeled by the current data log–likelihood.

Figure courtesy of Erik Sudderth
Comparing Collapsed vs. Regular

Given previous cluster assignments $z^{(t-1)}$, sequentially sample new assignments as follows:

1. Sample a random permutation $\tau(\cdot)$ of the integers $\{1, \ldots, N\}$.
2. Set $z = z^{(t-1)}$. For each $i \in \{\tau(1), \ldots, \tau(N)\}$, sequentially resample $z_i$ as follows:
   
   (a) For each of the $K$ clusters, determine the predictive likelihood $f_k(x_i) = p(x_i | \{x_j | z_j = k, j \neq i\}, \lambda)$.
   
   This likelihood can be computed from cached sufficient statistics via Prop. 2.1.4.
   
   (b) Sample a new cluster assignment $z_i$ from the following multinomial distribution:
      
      $z_i \sim \frac{1}{Z_i} \sum_{k=1}^{K} \left(N - i_k + \frac{\alpha}{K}\right) f_k(x_i) \delta(z_i, k)$
      
      $Z_i = \sum_{k=1}^{K} \left(N - i_k + \frac{\alpha}{K}\right) f_k(x_i)$
      
      $N - i_k$ is the number of other observations assigned to cluster $k$ (see eq. (2.162)).
   
   (c) Update cached sufficient statistics to reflect the assignment of $x_i$ to cluster $z_i$.

3. Set $z^{(t)} = z$. Optionally, mixture parameters may be sampled via steps 2–3 of Alg. 2.1.

Algorithm 2.2. Rao–Blackwellized Gibbs sampler for a $K$-component exponential family mixture model, as defined in Fig. 2.9. Each iteration sequentially resamples the cluster assignments for all $N$ observations $x = \{x_i\}_{i=1}^{N}$ in a different random order. Mixture parameters are integrated out of the sampling recursion using cached sufficient statistics of the parameters assigned to each cluster.

Figure 2.20. Comparison of standard (Alg. 2.1, dark blue) and Rao–Blackwellized (Alg. 2.2, light red) Gibbs samplers for a mixture of $K = 4$ two–dimensional Gaussians. We compare data log–likelihoods at each of 1000 iterations for the single $N = 300$ point dataset of Figs. 2.18 and 2.19.

Left: Log–likelihood sequences for 20 different random initializations of each algorithm.

Right: From 100 different random initializations, we show the median (solid), 0.25 and 0.75 quantiles (thick dashed), and 0.05 and 0.95 quantiles (thin dashed) of the resulting log–likelihood sequence $s$. The Rao–Blackwellized sampler has superior typical performance, but occasionally remains trapped in local optima for many iterations. These results suggest that while Rao–Blackwellization can usefully accelerate mixing, convergence diagnostics are still important.

Figure courtesy of Erik Sudderth

Task 2: Cluster Documents

- Previously:
  - Cluster documents based on topic
A Generative Model

- Documents: $x^1, \ldots, x^D$
- Associated topics: $z^1, \ldots, z^D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:

Task 2: Cluster Documents

- **Now:** Document may belong to multiple clusters
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 30 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough. Although the numbers don’t match precisely, those predictions:

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Fred Anderson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genomes," explains Araceli Masignan, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Computing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, Mar 8 to 12.

©Emily Fox 2013
Latent Dirichlet Allocation (LDA)

Topics

Documents

Topic proportions and assignments

-seeking Life's Bare (Genetic) Necessities

Data: The OCR’ed collection of Science from 1990-2000

- 17K documents
- 11M words
- 20K unique terms (stop words and rare words removed)

Model: 100-topic LDA model

Example Inference – Topic Weights

- "we test all that for grains," especially as compared to the 74,000 grains in the key experiment.

- "the test all that for grains," especially as compared to the 74,000 grains in the key experiment.

- "the text all that for grains," especially as compared to the 74,000 grains in the key experiment.

- "we test all that for grains," especially as compared to the 74,000 grains in the key experiment.

- "we test all that for grains," especially as compared to the 74,000 grains in the key experiment.

- "the test all that for grains," especially as compared to the 74,000 grains in the key experiment.
Example Inference – Topic Words

- human, evolution, disease, computer
- genome, evolutionary, host, models
- dna, species, bacteria, information
- genetic, organisms, diseases, data
- genes, life, resistance, computers
- sequence, origin, bacterial, system
- gene, biology, new, network
- molecular, groups, strains, systems
- sequencing, phylogenetic, control, model
- map, living, infectious, parallel
- information, diversity, malaria, methods
- genetics, group, parasite, networks
- mapping, new, parasites, software
- project, two, united, new
- sequences, common, tuberculosis, simulations

LDA Generative Model

- Observations: \( w_1^d, \ldots, w_{N_d}^d \)
- Associated topics: \( z_1^d, \ldots, z_{N_d}^d \)
- Parameters: \( \theta = \{ \{ \pi_k^d \} \}, \{ \beta_k \} \)
- Generative model:
LDA Generative Model

$$p(\cdot) = \prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} p(\pi^d \mid \alpha) \left( \prod_{i=1}^{N_d} p(z_i^d \mid \pi^d) p(w_i^d \mid z_i^d, \beta) \right)$$

Collapsed LDA Sampling

- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions
- Sample topic indicators for each word
  - Derivation:

$$p(z_i^d \mid N_d, \alpha) = \frac{\Gamma(\sum_k \alpha_k) \prod_k \Gamma(n_k^d + \alpha_k)}{\prod_k \Gamma(\alpha_k) \prod_k \Gamma(n_k^d + \alpha_k)}$$

$$p(w_i^d \mid z_i^d = k, \lambda) = \frac{\Gamma(\sum_{v} \lambda_v) \prod_v \Gamma(v_k^d + \lambda_v)}{\prod_v \Gamma(\lambda_v) \prod_v \Gamma(v_k^d + \lambda_v)}$$

$$p(z \mid \alpha) = \prod_{d=1}^{D} p(z_i^d \mid N_d, \alpha)$$

$$p(w \mid z, \lambda) = \prod_{k=1}^{K} p(w_i^d \mid z_i^d = k, \lambda)$$
Collapsed LDA Sampling

- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions
- Sample topic indicators for each word
  - Algorithm:

Sample Document

<table>
<thead>
<tr>
<th>Etruscan</th>
<th>trade</th>
<th>price</th>
<th>temple</th>
<th>market</th>
</tr>
</thead>
</table>

©Emily Fox 2013
### Randomly Assign Topics

<table>
<thead>
<tr>
<th>3</th>
<th>2</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
</tbody>
</table>

### Randomly Assign Topics

<table>
<thead>
<tr>
<th>3</th>
<th>2</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
</tbody>
</table>
### Maintain Global Statistics

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>temple</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>market</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>1</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>market</td>
<td>50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>price</td>
<td>42</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>temple</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>trade</td>
<td>10</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

Total counts from **all** docs

### Resample Assignments

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>price</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>temple</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>market</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>1</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>market</td>
<td>50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>price</td>
<td>42</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>temple</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>trade</td>
<td>10</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

©Emily Fox 2013
What is the conditional distribution for this topic?

Part I: How much does this document like each topic?

<table>
<thead>
<tr>
<th></th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>temple</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>market</td>
</tr>
</tbody>
</table>

©Emily Fox 2013
Part I: How much does this document like each topic?

Part II: How much does each topic like this word?

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>?</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
<td></td>
</tr>
</tbody>
</table>

What is the conditional distribution for this topic?

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>trade</td>
<td>10</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

©Emily Fox 2013
What is the conditional distribution for this topic?

- Part I: How much does this document like each topic?
- Part II: How much does each topic like this word?

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>?</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan trade price temple market</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Topic 1**

**Topic 2**

**Topic 3**

\[
\begin{align*}
\gamma & = \frac{n^d_k + \alpha_k}{\sum_{j=1}^K n^d_j + \alpha_j} \sum_{j=1}^Y v^k_j + \lambda_j \\
\end{align*}
\]

Sample a New Topic Indicator

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>?</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan trade price temple market</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Topic 1**

**Topic 2**

**Topic 3**

©Emily Fox 2013
Update Counts

<table>
<thead>
<tr>
<th>3</th>
<th>?</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>market</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>price</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>temple</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>trade</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

Geometrically…

<table>
<thead>
<tr>
<th>3</th>
<th>1</th>
<th>1</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etruscan</td>
<td>trade</td>
<td>price</td>
<td>temple</td>
<td>market</td>
</tr>
</tbody>
</table>

Topic 1  Topic 2  Topic 3
Issues with Generic LDA Sampling

- Slow mixing rates → Need many iterations
- Each iteration cycles through sampling topic assignments for all words in all documents
- Modern approaches:
- Next time: Variational methods instead of sampling

Acknowledgements

- Thanks to Dave Blei, David Mimno, and Jordan Boyd-Graber for some material in this lecture relating to LDA