Text Mining for Economics and Finance
Latent Dirichlet Allocation

Stephen Hansen
Recall we are interested in mixed-membership modeling, but that the pLSI model has a huge number of parameters to estimate.

One solution is to adopt a Bayesian approach; the pLSI model with a prior distribution on the document-specific mixing probabilities is called Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003).

LDA is widely used within computer science and, increasingly, social sciences.

LDA forms the basis of many, more complicated mixed-membership models.
1. Draw $\theta_d$ independently for $d = 1, \ldots, D$ from Dirichlet($\alpha$).

2. Each word $w_{d,n}$ in document $d$ is generated from a two-step process:
   
   2.1 Draw topic assignment $z_{d,n}$ from $\theta_d$.
   
   2.2 Draw $w_{d,n}$ from $\beta_{z_{d,n}}$.

Estimate hyperparameters $\alpha$ and term probabilities $\beta_1, \ldots, \beta_K$. 
1. Draw \( \beta_k \) independently for \( k = 1, \ldots, K \) from Dirichlet(\( \eta \)).

2. Draw \( \theta_d \) independently for \( d = 1, \ldots, D \) from Dirichlet(\( \alpha \)).

3. Each word \( w_{d,n} \) in document \( d \) is generated from a two-step process:
   3.1 Draw topic assignment \( z_{d,n} \) from \( \theta_d \).
   3.2 Draw \( w_{d,n} \) from \( \beta_{z_{d,n}} \).

Fix scalar values for \( \eta \) and \( \alpha \).
We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.
We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.
We have noticed a change in the relationship between the core CPI and the chained core CPI, which suggested to us that maybe something is going on relating to substitution bias at the upper level of the index. You focused on the nonmarket component of the PCE, and I wondered if something unusual might be happening with the core CPI relative to other measures.
Example statement: Yellen, March 2006, #51

We have noticed a change in the relationship between the two measures, which suggested to us that something is going on translating to substitution as at the per level of the index. You focused on the nmarket component of the E, and I wondered if something unusual might be happening with the measure relative to other measures.
Distribution of Attention
Advantage of Flexibility

‘measur’ has probability 0.026 in topic 25, and probability 0.021 in topic 11
Advantage of Flexibility

‘measur’ has probability 0.026 in topic 25, and probability 0.021 in topic 11.

It gets assigned to 25 in this statement consistently due to the presence of other topic 25 words.

In statements containing words on evidence and numbers, it consistently gets assigned to 11.

Sampling algorithm can help place words in their appropriate context.
External Validation—BBD

Text Mining
Lecture 5
Pro-Cyclical Topics

<table>
<thead>
<tr>
<th>FOMC Meeting Date</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987m7</td>
<td>0</td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>1991m1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994m7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998m1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001m7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005m1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008m7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Text Mining Lecture 5 10 / 45
Counter-Cyclical Topics

Text Mining

Lecture 5

11 / 45
Graphical Models

Consider a probabilistic model with joint distribution $f(x)$ over the random variables $x = (x_1, \ldots, x_N)$.

In high-dimensional models, it is useful to summarize relationships among random variables with directed graphs in which nodes represent random variables and links between nodes represent dependencies.

A node’s parents are the set of nodes that link to it; a node’s children are the set of nodes that it links to.

A Bayesian network is a probabilistic model whose joint distribution can be represented by a directed acyclic graph (DAG).

The nodes in a DAG can be ordered so that parents precede children.
LDA as a Bayesian Network

\[
\alpha \quad \theta_1 \quad z_{1,1} \quad z_{1,N_1} \quad \theta_D \quad z_{D,1} \quad z_{D,N_D}
\]

\[
w_{1,1} \quad w_{1,N_1} \quad \beta_1, \ldots, \beta_K \quad w_{D,1} \quad w_{D,N_D}
\]

\[
\eta
\]
**Graph Properties**

Let \( B = (\beta_1, \ldots, \beta_K) \) and \( T = (\theta_1, \ldots, \theta_D) \).

<table>
<thead>
<tr>
<th>Object</th>
<th>Parents</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>( \emptyset )</td>
<td>( T )</td>
</tr>
<tr>
<td>( \theta_d )</td>
<td>( \alpha )</td>
<td>( z_d )</td>
</tr>
<tr>
<td>( z_{d,n} )</td>
<td>( \theta_d )</td>
<td>( w_{d,n} )</td>
</tr>
<tr>
<td>( w_{d,n} )</td>
<td>( z_{d,n}, B )</td>
<td>( \emptyset )</td>
</tr>
<tr>
<td>( \beta_k )</td>
<td>( \eta )</td>
<td>( w_1, \ldots, w_D )</td>
</tr>
<tr>
<td>( \eta )</td>
<td>( \emptyset )</td>
<td>( B )</td>
</tr>
</tbody>
</table>
In a Bayesian network, nodes are independent of their ancestors conditional on their parents.

This means we can write \( f(x) = \prod_{i=1}^{N} f(x_i | \text{parents}(x_i)) \), which can greatly simplify joint distributions.

Applying this formula to LDA yields

\[
\left( \prod_d \prod_n \Pr [ w_{d,n} | z_{d,n}, \mathbf{B} ] \right) \cdot \left( \prod_d \prod_n \Pr [ z_{d,n} | \theta_d ] \right) \times \\
\left( \prod_d \Pr [ \theta_d | \alpha ] \right) \cdot \left( \prod_k \Pr [ \beta_k | \eta ] \right)
\]
The Markov blanket $\text{MB}(x_i)$ of a node $x_i$ in a Bayesian network is the set of nodes consisting of $x_i$’s parents, children, and children’s parents.

Conditional on its Markov blanket, the node $x_i$ is independent of all nodes outside its Markov blanket.

So $f(x_i \mid x_{-i})$ has the same distribution as $f(x_i \mid \text{MB}(x_i))$. 


**Posterior Distribution**

The inference problem in LDA is to compute the posterior distribution over $z$, $T$, and $B$ given the data $w$ and Dirichlet hyperparameters.

Let’s consider the simpler problem of inferring the latent variables taking the parameters as given. Posterior distribution is

$$
Pr[z = z' \mid w, T, B] = \frac{Pr[w \mid z = z', T, B] Pr[z = z' \mid T, B]}{\sum_{z'} Pr[w \mid z = z', T, B] Pr[z = z' \mid T, B]}.
$$

We can compute the numerator easily, and each element of denominator. But $z' \in \{1, \ldots, K\}^N \Rightarrow$ there are $K^N$ terms in the sum $\Rightarrow$ intractable problem.

For example, a 100 word corpus with 50 topics has $\approx 7.88e169$ terms.
The inference problem in LDA is to compute the posterior distribution over \( z, T, \) and \( B \) given the data \( w \) and Dirichlet hyperparameters.

Let’s consider the simpler problem of inferring the latent variables taking the parameters as given. Posterior distribution is

\[
\Pr[z = z' \mid w, T, B] = \frac{\Pr[w \mid z = z', T, B] \Pr[z = z' \mid T, B]}{\sum_{z'} \Pr[w \mid z = z', T, B] \Pr[z = z' \mid T, B]}.
\]

We can compute the numerator easily, and each element of denominator.

But \( z' \in \{1, \ldots, K\}^N \Rightarrow \) there are \( K^N \) terms in the sum \( \Rightarrow \) intractable problem.

For example, a 100 word corpus with 50 topics has \( \approx 7.88e169 \) terms.
Approximate Inference

Instead of obtaining a closed-form solution for the posterior distribution, we must approximate it.

Markov chain Monte Carlo methods provide a stochastic approximation to the true posterior.

The general idea is to define a Markov chain whose stationary distribution is equivalent to the posterior distribution, which we then draw samples from.

There are several MCMC methods, but we will consider Gibbs sampling.
**Gibbs Sampling**

We want to draw samples from some joint distribution over $\mathbf{x} = (x_1, \ldots, x_N)$ given by $f(\mathbf{x})$ (e.g. a posterior distribution).

Suppose we can compute the conditional distribution $f_i \equiv f(x_i \mid \mathbf{x}_{-i})$.

Then we can use the following algorithm:

1. Randomly allocate an initial value for $\mathbf{x}$, say $\mathbf{x}^0$

2. Let $S$ be the number of iterations to run chain. For each $s \in \{1, \ldots, S\}$, draw $x^s_i$ according to

   $$x^s_i \sim f(x_i \mid x^s_1, \ldots, x^s_{i-1}, x^{s-1}_{i+1}, \ldots, x^{s-1}_N).$$

3. Discard initial iterations (burn in), and collect samples from every $m$th (thinning interval) iteration thereafter.

4. Use collected samples to approximate joint distribution, or related distributions and moments.
Sampling Equations for $\theta_d$

The Markov blanket of $\theta_d$ is:

- The parent $\alpha$.
- The children $z_d$.

So we need to draw samples from $\Pr[\theta_d | \alpha, z_d]$. This is the posterior distribution for $\theta_d$ given a fixed value for the vector of allocation variables $z_d$.

We computed this posterior above. Let $n_{d,k}$ be the number of words in document $d$ that have topic allocation $k$.

Then $\Pr[\theta_d | \alpha, z_d] = \text{Dir}(\alpha + n_{d,1}, \ldots, \alpha + n_{d,K})$. 
The Markov blanket of $\beta_k$ is:

- The parent $\eta$.
- The children $w = (w_1, \ldots, w_D)$.
- The children’s parents:
  1. $z = (z_1, \ldots, z_D)$.
  2. $B_{-k}$.

Consider $\Pr[\beta_k | \eta, w, z]$. Only the allocation variables assigned to $k$—and their associated words—are informative about $\beta_k$.

Let $m_{k,v}$ be the number of times topic $k$ allocation variables generate term $v$.

Then $\Pr[\beta_k | \eta, w, z] = \text{Dir}(\eta + m_{k,1}, \ldots, \eta + m_{k,v})$. 
The Markov blanket of $z_{d,n}$ is:

- The parent $\theta_d$.
- The child $w_{d,n}$.
- The child’s parents $\beta_1, \ldots, \beta_K$.

\[
\Pr[z_{d,n} = k \mid w_{d,n} = \nu, B, \theta_d] = \frac{\Pr[w_{d,n} = \nu \mid z_{d,n} = k, B, \theta_d] \Pr[z_{d,n} = k \mid B, \theta_d]}{\sum_k \Pr[w_{d,n} = \nu \mid z_{d,n} = k, B, \theta_d] \Pr[z_{d,n} = k \mid B, \theta_d]} = \frac{\theta^k_d \beta^\nu_k}{\sum_k \theta^k_d \beta^\nu_k}.
\]
To complete one iteration of Gibbs sampling, we need to:

1. Sample from a multinomial distribution $N$ times for the topic allocation variables.

2. Sample from a Dirichlet $D$ times for the document-specific mixing probabilities.

3. Sample from a Dirichlet $K$ times for the topic-specific term probabilities.

Sampling from these distributions is standard, and implemented in many programming languages.
Collapsed sampling refers to analytically integrating out some variables in the joint likelihood and sampling the remainder.

This tends to be more efficient because we reduce the dimensionality of the space we sample from.

Griffiths and Steyvers (2004) proposed a collapsed sampler for LDA that integrates out the $T$ and $B$ terms and samples only $z$.

For details see Heinrich (2009) and technical appendix of Hansen, McMahon, and Prat (2015).
The sampling equation for the $n$th allocation variable in document $d$ is:

$$
\Pr \left[ z_{d,n} = k \mid z_{-(d,n)}, w, \alpha, \eta \right] \propto \frac{m_{k,v_{d,n}}^- + \eta}{\sum_v m_{k,v}^- + \eta V} \left( n_{d,k}^- + \alpha \right)
$$

where the $-$ superscript denotes counts excluding $(d, n)$ term.
The sampling equation for the $n$th allocation variable in document $d$ is:

$$\Pr \left[ z_{d,n} = k \mid z_{-(d,n)}, w, \alpha, \eta \right] \propto \frac{m_{k,v_{d,n}}^- + \eta}{\sum_v m_{k,v}^- + \eta V} \left( n_{d,k}^- + \alpha \right)$$

where the $-$ superscript denotes counts excluding $(d, n)$ term.

Probability term $n$ in document $d$ is assigned to topic $k$ is increasing in:

1. The number of other terms in document $d$ that are currently assigned to $k$.
2. The number of other occurrences of the term $v_{d,n}$ in the entire corpus that are currently assigned to $k$.

Both mean that terms that regularly co-occur in documents will be grouped together to form topics.

Property 1 means that terms within a document will tend to be grouped together into few topics rather than spread across many separate topics.
Predictive Distributions

Collapsed sampling gives the distribution of the allocation variables, but we also care about variables we integrated out.

Their predictive distributions are easy to form given topic assignments:

\[
\hat{\beta}_{k,v} = \frac{m_{k,v} + \eta}{\sum_{v=1}^{V} (m_{k,v} + \eta)} \quad \text{and} \quad \hat{\theta}_{d,k} = \frac{n_{d,k} + \alpha}{\sum_{k=1}^{K} (n_{d,k} + \alpha)}.
\]
Recall from the first lecture that text data is one instance of count data.

Although typically applied to natural language, LDA is in principle applicable to any count data.

We recently applied it to CEO survey data to estimate management “behaviors” with $K = 2$.

Can visualize results in terms of likelihood ratios of marginals over specific data features.
Activity Type

- Meeting
- Site visit
- Communications
Planning; Size; Number of Functions

- Planned
- >1hr
- 2 people or more
- 2 functions or more

Bar chart showing the number of functions for different categories.
Insiders vs Outsiders

- Insiders
- Outsiders
- Insiders & Outsiders

Bar chart showing the comparison between Insiders, Outsiders, and Insiders & Outsiders.
There are three parameters to set to run the Gibbs sampling algorithm: number of topics $K$ and hyperparameters $\alpha, \eta$.

Priors don’t receive too much attention in literature. Griffiths and Steyvers recommend $\eta = 200/V$ and $\alpha = 50/K$. Smaller values will tend to generate more concentrated distributions. (See also Wallach et. al. 2009).

$K$ is less clear. Two potential goals:

1. Predict text well. Statistical criteria to select $K$.
Chang et. al. (2009) propose an objective way of determining whether topics are indeed interpretable.

Two tests:

1. *Word intrusion*. Form set of words consisting of top five words from topic $k$ + word with low probability in topic $k$. Ask subjects to identify inserted word.

2. *Topic intrusion*. Show subjects a snippet of a document + top three topics associated to it + randomly drawn other topic. Ask to identify inserted topic.

Estimate LDA and other topic models on NYT and Wikipedia articles for $K = 50, 100, 150$. 
Figure 5: A scatter plot of model precision (top row) and topic log odds (bottom row) vs. predictive log likelihood. Each point is colored by model and sized according to the number of topics used to fit the model. Each model is accompanied by a regression line. Increasing likelihood does not increase the agreement between human subjects and the model for either task (as shown by the downward-sloping regression lines).
Topics seem objectively interpretable in many contexts.

Tradeoff between goodness-of-fit and interpretability, which is generally more important in social science.

Potential development of statistical models in future to explicitly maximize interpretability.
Determining when a chain has converged can be tricky.

One approach is to measure how well different states of the chain predict the data, and determine convergence in terms of its stability.

Standard practice is run chains from different starting values, after which you can select the best-fit chain for analysis.

For LDA, a typical goodness-of-fit measure is perplexity, given by

\[
\exp \left[ - \frac{\sum_{d=1}^{D} \sum_{v=1}^{V} x_{d,v} \log \left( \sum_{k=1}^{K} \hat{\theta}_{d,k} \hat{\beta}_{k,v} \right)}{\sum_{d=1}^{D} N_d} \right].
\]
Perplexity 1

![Perplexity Graph]

- Chain 1
- Chain 2
- Chain 3
- Chain 4
- Chain 5

Perplexity vs. Iteration
Perplexity 2

![Graph showing perplexity over iterations for multiple chains](image-url)
Out-of-Sample Documents

We are sometimes interested in obtaining the document-topic distribution for out-of-sample documents.

We can perform Gibbs sampling treating estimated topics as fixed

\[
\Pr \left[ z_{d,n} = k \mid z_{-(d,n)}, w_d, \alpha, \eta \right] \propto \hat{\beta}_{k, v_{d,n}} \left[ n_{d,k} - 1 + \alpha \right]
\]

for each out-of-sample document \( d \).

Only 10-20 iterations necessary since topics already estimated.
Dictionary Methods + LDA

The terms in dictionaries come labeled, so can be seen as a type of supervised approach to information retrieval.

One can combine dictionary methods with the output of LDA to weight words counts by topic.

Recent application to minutes of the Federal Reserve to extract index of economic situation and forward guidance.

First step is to run 15-topic model and identify two separate kinds of topic.
## Monetary Measures of Tone

<table>
<thead>
<tr>
<th>Contraction</th>
<th>Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td>decreases*</td>
<td>increases*</td>
</tr>
<tr>
<td>decelerate*</td>
<td>accelerate*</td>
</tr>
<tr>
<td>slow*</td>
<td>fast*</td>
</tr>
<tr>
<td>weak*</td>
<td>strong*</td>
</tr>
<tr>
<td>low*</td>
<td>high*</td>
</tr>
<tr>
<td>loss*</td>
<td>gain*</td>
</tr>
<tr>
<td>contract*</td>
<td>expand*</td>
</tr>
</tbody>
</table>
Key ideas from this lecture:

1. Latent Dirichlet Allocation. Influential in computer science, many potential applications in economics.

2. Graphical models to simplify and visualize complex joint likelihoods.

3. Gibbs sampling to stochastically approximate posterior.

4. Interpretable output in several domains.

Topic for future: incorporate economic structure into LDA.