Lecture 6:
Topic Models
(Latent Dirichlet Allocation)

Motivating questions:
- What are the topics that a document is about?
- Given one document, can we find other documents about the same topics?
- How do topics in a field change over time?

A hierarchical Bayesian approach:
- Assume each document defines a distribution over (hidden) topics
- Assume each topic defines a distribution over words
- The posterior probability of these latent variables given a document collection determines a hidden decomposition of the collection into topics.

References


Setting up a generative model
- We have $D$ documents using a vocabulary of $V$ word types
- Each documents contains (up to) $N$ word tokens.
- We assume $K$ topics.
- Each document has a $K$-dimensional multinomial $\theta_d$ over topics with a common Dirichlet prior $Dir(\alpha)$
- Each topic has a $V$-dimensional multinomial $\beta_k$ over words with a common symmetric Dirichlet prior $D(\eta)$.
**The generative process**

For each topic \(1 \ldots k\):
- Draw a multinomial over words \(\beta_k \sim \text{Dir}(\eta)\)

For each document \(1 \ldots d\):
- Draw a multinomial over topics \(\theta_d \sim \text{Dir}(\alpha)\)
- For each word \(w_{d,n}\):
  - Draw a topic \(Z_{d,n} \sim \text{Mult}(\theta_d)\) with \(Z_{d,n} \in \{1..K\}\)
  - Draw a word \(w_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}})\)

**As a graphical model**

![Graphical model of the generative process]

**A geometric interpretation**

![Geometric interpretation of topic words]

**The generative process**

![Diagram of the generative process]

---


The inference problem

**STATISTICAL INFERENCE**

![Diagram of topic modeling]


**Gibbs sampling**

Represent corpus as an array of words \( w[i,j] \), document indices \( d[i] \) and topics \( z[i] \).

Words \( w[i,j] \) and document indices \( d[i] \) are fixed.
Only the topics \( z[i] \) change.

States of Markov chain = topic assignments to words.

"Macrosteps": assign a new topic to all of the words
"Microsteps": assign a new topic to each word \( w[i] \).

The inference problem

What is the posterior of the hidden variables given the observed variables (and hyperparameters)?

\[
p(\theta_1,: z_{1:D}, \beta_{1,k} \mid w_{1:D,1:N}, \alpha, \eta) = \frac{p(\theta_1,: z_{1:D}, \beta_{1,k} \mid w_{1:D}, \alpha, \eta)}{\int_\beta \int_\theta p(\theta_1,: z_{1:D}, \beta_{1,k} \mid w_{1:D}, \alpha, \eta)}
\]

Problem: the integral in the denominator is intractable!

Solutions: approximate inference
- Gibbs sampling (Griffiths & Steyvers)
- Variational Inference (Blei et al.)

**Assigning a new topic to \( w_i \)**

The probability \( P(z_i = j \mid z_{<i}, w, d) \) is proportional to the probability of \( w_i \) under topic \( j \) times the probability of topic \( j \) given document \( d_i \)

Define \( n_{i,j}(w_i) \) as freq. of \( w_i \) labeled as topic \( j \).
Define \( n_{i,j}(d_i) \) as number of words in \( d_i \) labeled as topic \( j \).

\[
P(z_i = j \mid z_{<i}, w, d) \propto \frac{n_{i,j}(w_i) + \eta}{V_n + \sum_{v=1}^{V_n} n_{i,j}(w_v)} \cdot \frac{n_{i,j}(d_i) + \alpha}{K\alpha + \sum_{k=1}^{K} n_{i,j}(d_k)}
\]

Prob of \( w_i \) under topic \( z_i \)
Prob of topic \( z_i \) in document \( d_i \)
What other quantities do we need?

As Bayesians, we want to compute the expected value of the parameters given the observed data. Our data is the set of words $w_{1:D,1:N}$. Hence, we need to compute $E[\ldots | w_{1:D,1:N}]$.

The **topic probability** of the word $v$ according to topic $k$:

$$\langle \beta_{k,v} \rangle = E[\beta_{k,v} | w_{1:D,1:N}]$$

The **topic proportions of each document $d$**:

$$\langle \theta_{d,k} \rangle = E[\theta_{d,k} | w_{1:D,1:N}]$$

The **topic assignment of each word $w_{d,n}$**:

$$\langle z_{d,n,k} \rangle = E[Z_{d,n} = k | w_{1:D,1:N}]$$

---

**Running the Gibbs sampler**

A toy example from Griffiths, T., & Steyvers, M. (2004):
- 25 words.
- 10 predefined topics
- 2000 documents generated according to known distributions.
- Each document = 5x5 image. Pixel intensity = Frequency of word.

From any single sample, we can estimate:

The **topic probability** of the word $v$ according to topic $k$:

$$\langle \beta_{k,v} \rangle = E[\beta_{k,v} | w_{1:D,1:N}]$$

The **topic proportions of each document $d$**:

$$\langle \theta_{d,k} \rangle = E[\theta_{d,k} | w_{1:D,1:N}]$$

The graphical representation of 10 topics, combined to produce “documents” like those shown in b, where each image is the result of 100 samples from a unique mixture of these topics.

Griffiths T. L, Steyvers M PNAS 2004;101:5228-5235

©2004 by National Academy of Sciences
Visualizing a topic

What is the distribution of words defined by a topic? Instead of using term probabilities directly, downweight them by how likely they are to be generated by any topic:

$$\text{term-score}_{w,v} = \hat{p}_{k,v} \log \left( \frac{\hat{p}_{k,v}}{\prod_{j=1}^{K} \hat{p}_{j,v}} \right)$$

contractual, expectation, gain, promises, breach, enforcing, note, perform

employment, industrial, local, employees, relations, unfair, economic, case

female, men, women, sexual, structure, employer, discrimination, risk

markets, earnings, investors, justice, federal, managers, firm, parole

criminal, discretion, civil, process, structure, officer, risk, inmates


Things you can do with LDA

Top words from the top topics (by term score)

sequence, region, pcr, identified, fragments, genes, three, cdna, analysis

measured, average, range, values, different, size, three, calculated, two

residues, binding, domains, helix, cys, regions, structure, terminus, terminal, site

computer methods, number, two, principle, design, access, processing, advantage, important

Visualizing a document

Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence tests and for evaluating sequence comparisons, (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations, and (iii) the statistical methods that can be applied to the analysis of spacings of sequence markers.

Use the posterior topic probabilities of each document and the posterior topic assignments to each word


Document similarity

Two documents are similar if they assign similar probabilities to topics:

$$\text{document-similarity}_{d,f} = \sum_{k=1}^{K} \left( \sqrt{\hat{\theta}_{d,k}} - \sqrt{\hat{\theta}_{f,k}} \right)^2.$$