Probabilistic Topic Models: Origins and Challenges

David M. Blei

Department of Computer Science
Princeton University

December 9, 2013
• ORGANIZE
• VISUALIZE
• SUMMARIZE
• SEARCH
• PREDICT
• UNDERSTAND
Probabilistic Topic Modeling

**Input:** An unorganized collection of documents
**Output:** An organized collection, and a description of how
<table>
<thead>
<tr>
<th>Year</th>
<th>Phase</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1880</td>
<td>electric machine</td>
<td>power, steam, two, iron, battery, wire</td>
</tr>
<tr>
<td>1890</td>
<td>electric apparatus</td>
<td>steam, power, electrical, two, motor, engine</td>
</tr>
<tr>
<td>1900</td>
<td>air apparatus</td>
<td>steam, power, engineering, water, construction, engineer, room, feet</td>
</tr>
<tr>
<td>1910</td>
<td>air apparatus</td>
<td>tube, water, engineering, apparatus, room, laboratory, engineer, made, gas, tube</td>
</tr>
<tr>
<td>1920</td>
<td>tube apparatus</td>
<td>glass, air, pressure, water, gas, made, laboratory, mercury</td>
</tr>
<tr>
<td>1930</td>
<td>tube apparatus</td>
<td>glass, air, mercury, laboratory, mercury, made, gas, small</td>
</tr>
<tr>
<td>1940</td>
<td>air apparatus</td>
<td>tube, apparatus, glass, laboratory, rubber, pressure, small, mercury, gas</td>
</tr>
<tr>
<td>1950</td>
<td>tube apparatus</td>
<td>glass, air, chamber, instrument, small, laboratory, pressure, rubber</td>
</tr>
<tr>
<td>1960</td>
<td>tube system</td>
<td>temperature, air, heat, chamber, power, high, instrument, control</td>
</tr>
<tr>
<td>1970</td>
<td>air system</td>
<td>heat, temperature, chamber, system, high, instrument, control, flow, tube, design</td>
</tr>
<tr>
<td>1980</td>
<td>high power system</td>
<td>heat, system, devices, instruments, control, large</td>
</tr>
<tr>
<td>1990</td>
<td>materials high power current applications</td>
<td>technology, devices, design, device, heat</td>
</tr>
<tr>
<td>2000</td>
<td>devices</td>
<td>device, materials, current, gate, high, light, silicon, material, technology</td>
</tr>
</tbody>
</table>
Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors. Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschew structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature. A recurring theme in his films is man's inhumanity to man. While often viewed as
This talk

1. The origins of probabilistic topic modeling
2. The basics of latent Dirichlet allocation
3. A couple ideas that we are excited about in my group
4. Open questions, challenges, and discussion
Latent Semantic Analysis (LSA)
(Deerwester et al., 1990)

This is the seminal work that launched topic modeling.

- Treat a collection as a document by term matrix of TFIDF scores.
- Choose a number of topics, and run SVD on the matrix.
- This results in
  - a matrix of per-document topic weights
  - a matrix of per-topic term weights
Probabilistic Latent Semantic Analysis (pLSA)  
(Hofmann, 1999)

- A probabilistic model based on the main ideas of LSA
- Define a **topic** as a distribution over terms.
- Describe each document as a distribution over topics.
- Learn these two sets of parameters with EM.
- Note: This model was also defined in Papadimitriou et al., 1998
Latent Dirichlet Allocation (LDA)
(Blei et al., 2001; Blei et al., 2003)
**Seeking Life’s Bare (Genetic) Necessities**

*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 250 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 Steve Anderson, a graduate student at the University in St. Petersburg, who arrived at the 200 number. But coming up with a consensus answer may be more than just a matter of numbers. Some, particularly more and more genome researchers, are beginning to question the sequencing. “It may be a way of organisms any newly sequenced genome,” explains Arca J. Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing it...

Seeking Life’s Bare (Genetic) Necessities

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 250 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 needed 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 Steve Anderson of Yale University in New Haven, who arrived at the 2,000 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 sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Musheghian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

Posterior inference
Seeking Life’s Bare (Genetic) Necessities

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 256 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 Siv Andersson 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 genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
Why does LDA “work”?

• LDA trades off two goals
  1. In each **document**, allocate its words to **few topics**.
  2. In each **topic**, assign high probability to **few terms**.

• We see this from the joint

\[
\log p(\cdot) = \ldots + \sum_d \sum_n \log p(z_{dn} \mid \theta_d) + \log p(w_{dn} \mid \beta_{z_{dn}}) + \ldots
\]

• Sparse proportions come from the 1st term.
  Sparse topics come from the 2nd term.
Why does LDA “work”?  

LDA trades off two goals  
1. In each **document**, allocate its words to **few topics**.  
2. In each **topic**, assign high probability to **few terms**.  

These goals are at odds.  
- Putting a document in a single topic makes #2 hard.  
- Putting very few words in each topic makes #1 hard.  

Trading off these goals finds groups of tightly co-occurring words.
Summary and other perspectives

- Disovers topics through posterior inference
- Can be seen as *multinomial PCA* (Buntine and Jakulin, 2004)
- Is a type of *mixed-membership model* (Erosheva, 2004)
- Independently invented in population genetics (Pritchard et al., 2000)
• LDA is a simple building block that enables many applications.
• Organizing and finding patterns in data has become important in the sciences, humanities, industry, and culture.
• Algorithmic improvements let us fit models to massive data.
- Case study in **text analysis with probability models**
- Topic modeling research
  - develops new models.
  - develops new inference algorithms.
  - develops new applications, visualizations, tools.
Some ideas we are excited about in my research group
Idea #1: User behavior data

People use documents.

This information can be used to

- Help people find documents that they are interested in
- Learn about how the documents are implicitly organized
- Learn about the people reading the documents
Idea #1: User behavior data

- **Collaborative topic models** analyze text and user data.
- They can be used to
  - recommend articles to readers: old and new
  - describe users in terms of their preferences
  - identify impactful, interdisciplinary articles
• Consider EM (Dempster et al., 1977). We infer topics from its text:

Maximum Likelihood from Incomplete Data via the EM Algorithm

By A. P. Dempster, N. M. Laird and D. B. Rubin
Harvard University and Educational Testing Service

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, December 8th, 1976, Professor S. D. Silvey in the Chair]

SUMMARY
A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

• Suppose there are two types of scientists

STATISTICIAN

VISION RESEARCHER

• We first recommend the EM paper to statisticians.
• With user data, we can adjust the topics to account for who liked it:

![Diagram showing user data and paper recommendations.]

• Consider again the scientists

![Diagram showing statistics and vision research.]  

• We now recommend the EM paper to **vision researchers**.
1. Without text, we cannot initially recommend to anyone.
2. Without user data, we cannot recommend to vision researchers.
3. We learned about the special interdisciplinary status of the EM paper.
The collaborative topic model

(Wang and Blei, 2011)
• Trades off matrix factorization and content recommendation
• The dimensions of user preferences also explain the text.
• Thus, they are interpretable.
Maximum Likelihood from Incomplete Data via the $EM$ Algorithm

By A. P. Dempster, N. M. Laird and D. B. Rubin

Harvard University and Educational Testing Service

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 8th, 1976, Professor S. D. Silvey in the Chair]

Summary

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.
general, examples, presented, discussed, algorithm, algorithms, optimization, problem, efficient estimates, likelihood, maximum, parameters
Maximum Likelihood Estimation
{Estimates, Likelihood, Maximum, Parameters, Method}

Widely read
- Maximum Likelihood Estimation of Population Parameters
- Bootstrap Methods: Another Look at the Jackknife
- R. A. Fisher and the Making of Maximum Likelihood

Interdisciplinary MLE articles
- Maximum Likelihood from Incomplete Data with the EM Algorithm
- Bootstrap Methods: Another Look at the Jackknife
- Tutorial on Maximum Likelihood Estimation

Outside influences
- Random Forests
- Identification of Causal Effects Using Instrumental Variables
- Matrix Computations
Idea #1: User behavior data

- Collaborative topic models give good recommendations.
- User behavior data give us a new window into the collection.
- Q: What if the users are in a network?
- Q: What if the users write reviews?
Idea #2: Poisson factorization

1. For each term $v$ and topic $k$: draw $\beta_{kv} \sim \text{Gamma}(a, b)$
2. For each document $d$:
   a. For each topic $k$: draw $\theta_{dk} \sim \text{Gamma}(c, d)$.
   b. For each term $v$: draw $n_{dv} \sim \text{Poisson}(\theta_d^T \beta_v)$. 
Idea #2: Poisson factorization

- Shows better perplexity than LDA. (Canny, 2004)
- Easy to fit with auxiliary variables
- Easy to extend the Poisson additive model on word counts
- Equivalent to LDA when we condition on document length (It is multinomial PCA.)
- Is a Bayesian form of NMF with “KL loss” (Lee and Seung, 2000)
Idea #2: Poisson factorization

- Works well in other settings
  - networks (Ball et al., 2012); recommendation (Gopalan et al., 2013)
- We can build Bayesian nonparametric versions (Gopalan et al., yesterday)
- Why is it better than LDA?
  - Explicitly models document length?
  - Avoids pesky normalizations?
Idea #3: Stochastic Variational Inference

Massive document collection

Topics

SUBSAMPLE DOCUMENTS

INFER LOCAL HIDDEN VARIABLES

UPDATE TOPICS

(Hoffman et al., 2010, 2013)
Challenges to topic modeling
• Topic modeling research
  • develops new models.
  • develops new inference algorithms.
  • develops new applications, visualizations, tools.

• Workshops are also for half-baked ideas and difficult-to-articulate problems.
How do we explore?

- Topic models are used to explore collections.
- How can we build and evaluate models with this goal?
- Brings to focus thorny issues
  - Visualization, Interpretability
  - Interactivity, Never-ending collections
- Theory of exploration (Tukey, 1962; Good, 1983; Diaconis, 1985)
How do we select and revise?

- Which model should I choose for my problem?
- Where does my model go right? Where does it go wrong?
- More thorny issues
  - Model evaluation
  - Posterior predictive checks (Box, 1980; Rubin, 1984; Gelman et al., 1996)
How do we apply?

- Topic modeling moves in useful directions when we solve real problems.
- Collaborate with scientists/scholars that want to analyze texts
  - E.g., History, Comparative Literature, Political Science The Law, Cognitive Science, Sociology, Media Theory, Linguistics, Biology
- Create usable open-source tools for topic modeling.
- Success story: MALLET and the digital humanities.
Box’s loop

Build model

Infer hidden variables

Data

Predict & Explore

Criticize model

Revise