Statistical Topic Models for Science and Innovation Policy

Hanna M. Wallach
University of Massachusetts Amherst
wallach@cs.umass.edu
“Whether it's improving our health or harnessing clean energy, protecting our security or succeeding in the global economy, our future depends on reaffirming America's role as the world's engine of scientific discovery and technological innovation.”

— President Barack Obama
... Behind the Scenes

“The public has generally treated this progress as something that just happened, without recognizing that it is, in fact, largely the result of a sustained federal commitment to support science through science policies.”

— http://science-policy.net
Science and Innovation Policy

- Goal: identify administrative, financial, political actions
- Actions chosen to have impact on, e.g.,
  - Stimulating breakthrough research
  - Increasing economic prosperity
  - Broadening participation
- Government, private sector, education
- This talk: statistical models for facilitating efficient, data-driven science policy decisions
Examples of Policy Actions

- **Funding actions:**
  - Using federal funds for research on human stem cells
  - “People not projects” vs. pre-defined deliverables

- **Patenting actions:**
  - Granting software patents

- **Educational actions:**
  - Running high school outreach activities
  - Providing mentoring programs
Data-Driven Policy Decisions

- Discovery: identifying possible policy actions
- Prediction: estimating expected impact
- Evaluation: assessing observed outcomes

⇒ Automated data analysis
“Scientific information is both the basic raw material for, and one of the principal products of, scientific research […] Scientists find out what other scientists are accomplishing through [...] journals, books, abstracts and indexes, bibliographies, reviews.”

— NSF Brochure, 1962
Approach: Statistical Models

- Modeling challenges:
  - Aggregating and representing large data sets
  - Handling data from sources with disparate emphases
  - Reasoning under uncertain information
  - Performing efficient inference

- Bayesian latent (hidden) variable models:
  - Powerful and flexible [Wallach et al. & Adams et al., AISTATS '10]
  - This talk: statistical topic models
My Research Goal

To develop new statistical models and computational tools for representing and analyzing large quantities of complex data in order to better enable scientific policy-makers to identify and evaluate high-impact policy actions and advance the study of science and innovation policy.
Collaborate to Study Collaboration

“There needs to be a greater focus on what these [science interaction] data mean [...] This requires the input of social scientists, rather than just those more traditionally involved in data capture, such as computer scientists.”

— Julia Lane, NSF, 24 March 2010
This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Evaluating statistical topic models

Collaborators: Sarah Kaplan, Rotman, University of Toronto; Andrew McCallum, UMass Amherst; David Mimno, UMass Amherst; Ned Talley, NIH
This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Evaluating statistical topic models

Collaborators: Sarah Kaplan, Rotman, University of Toronto; Andrew McCallum, UMass Amherst; David Mimno, UMass Amherst; Ned Talley, NIH
Why Topic Models?

From (9) it can then be shown that (Exercise 3.2)
\[ \lambda = (K^{-1} - K^{-1} M (M^T K^{-1} M)^{-1} M^T K^{-1}) \]
so that the resulting predictor
\[ Z = k^T \lambda \]
which is identical to what we called generalized least squares estimates
\[ k_n - k^T K \lambda \]
where \( \gamma = m(x_0) - M^T K \lambda \).

Best linear unbiased prediction (BLUP), named after the Solow (1951) and Huijbregts (1965) process is assumed to be an ordinary least squares estimator for the more general \( m \) is known a priori with the mean assumed 0 is generally called objective analysis (Pedder 1987 and Daley 1991). The linear unbiased prediction for regression modeling did not explicitly consider the spatial setting. For further discussion on the history of various formulations, see Haining et al. (1993).

As noted in 1.3, a useful characterization of

**Definition 2.1** A Gaussian process is a complete finite number of which have a joint Gaussian distribution. We define mean function of the process \( f(x) \) as
\[ m(x) = \mathbb{E}[f(x)], \]
\[ (x, x') = \mathbb{E}[(f(x) - m(x))] \]
Gaussian process as
\[ f(x) \sim \mathcal{G}(m(x)) \]
weional simplicity we will not be concerned here. The random variables representing the data at \( \mathcal{X} \), Gaussian processes are non-stationary when the random variables is time.

where the index set \( \mathcal{X} \) is the be more general, e.g., \( \mathbb{R}^D \). For notation and enumeration of the cases in the training set, such that \( f_i \triangleq f(x_i) \) is the random variable as would be expected.
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 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.
Topics and Words

- human
dna
human
evolution
disease
computer
genome
evolutionary
evolutionary
disease
data

- species
organisms
host
host
biology
bacteria

- life
origin
diseases
resistance

- host
bacterial

- new
strains
control

- organisms
infectious

- living

- computer
models
information

- information

- models

- data
computers

- system
network

- system
model

- model
parallel

- parallel
Generative Statistical Modeling

- Assume data was generated by a probabilistic model:
  - Model may have hidden structure (latent variables)
  - Model defines a joint distribution over all variables
  - Model parameters are unknown
- Infer hidden structure and model parameters from data
- Situate new data into estimated model
Generative Process
Choose a Distribution Over Topics

...
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 150 genes, and that the earliest life forms required a mere 128.
Choose a Word

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.
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, these predictions are not all that far apart, especially in comparison to the 75,000 genes in the human genome, notes Sti 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 genome," explains Aracly Mazzina, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Sti 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 genome," explains Aracly Mazzina, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Sti 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 genome," explains Aracly Mazzina, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an


SCIENCE • VOL. 272 • 24 MAY 1996

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

Hanna M. Wallach :: UMass Amherst :: 21
Real Data: Statistical 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 230 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, these predictions


“are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Stév 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 genome,” explains Araceli Muñóz-Montes, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

[Diagram]

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

SCIENCE • VOL. 272 • 24 MAY 1996

probability
Directed Graphical Models

\[ P(y, x_1, \ldots, x_N) = P(y) \prod_{n=1}^{N} P(x_n | y) \]

- Nodes: random variables (latent or observed)
- Edges: probabilistic dependencies between variables
- Plates: “macros” that allow subgraphs to be replicated
Statistical Topic Modeling

[Perfmann, '99]
Latent Dirichlet Allocation (LDA)

[Blei, Ng & Jordan, '03]
The State of The Art

- Topic models are extremely popular
- ... but they're not always usable by non-experts
- Need to bridge this gap between producers and consumers of topic modeling technology:
  - Address problems/challenges faced by practitioners
  - Question unquestioned assumptions
  - Explore the interplay between theory and practice
This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Evaluating statistical topic models

Collaborators: Sarah Kaplan, Rotman, University of Toronto; Andrew McCallum, UMass Amherst; David Mimno, UMass Amherst; Ned Talley, NIH
“Off-the-Shelf” Topic Modeling

I want to model technology emergence by analyzing patent abstracts...

I have a statistical model that you can use...
I want to model technology emergence by analyzing patent abstracts...

I have a statistical model that you can use...

| a field emission an electron ... | a the carbon and gas ... | the of a to and ... | the invention of to present ...

"Off-the-Shelf" Topic Modeling
“Off-the-Shelf” Topic Modeling?

Help! All my topics consist of “the, and of, to, a ...”

Now they all consist of “invention, present, thereof ...”

Wait, but how do I choose the right number of topics?

Preprocess your data to remove stop words...

Make a domain-specific list of stop words...

Evaluate the probability of unseen data for different numbers...
3-dimensional discrete probability distributions can be visually represented in 2-dimensional space:
Dirichlet Distribution

- Distribution over discrete probability distributions:

\[ p \sim \text{Dir}(\alpha m) \]

- Base measure (mean)
- Concentration parameter

\[ \equiv \]

\[ \sim \]

\[ \ldots \]
Dirichlet Parameters

$\alpha = 14$

$m = \left(\frac{5}{7}, \frac{1}{7}, \frac{1}{7}\right)$

$m = \left(\frac{1}{7}, \frac{5}{7}, \frac{1}{7}\right)$

$m = \left(\frac{1}{7}, \frac{1}{7}, \frac{5}{7}\right)$

$m = u = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$

$\alpha = 3$

$\alpha = 6$

$\alpha = 30$
Dirichlet Priors for LDA

symmetric priors: uniform base measures
Dirichlet Priors for LDA

- Two scalar concentration parameters: $\alpha$ and $\beta$
- Concentration parameters are usually set heuristically
  - e.g., $\alpha = 50$ and $\beta = 0.01W$
- Some recent work on learning optimal values for the concentration parameters from data
- No rigorous study of the Dirichlet priors:
  - e.g., asymmetric vs. symmetric base measures
  - Effects of the base measures on the inferred topics
Symmetric $\rightarrow$ Asymmetric

- Use prior over $\Theta = \{\theta_1, \ldots, \theta_D\}$ as a running example
- Uniform base measure $\rightarrow$ nonuniform base measure

$$\Theta \sim \text{Dir}(\alpha m) \quad \Theta \sim \text{Dir}(\alpha m)$$

- Asymmetric prior: some topics more likely a priori
Hierarchical Asymmetric Dirichlet

- Which topics should be more probable a priori?
  - Draw $m$ from a Dirichlet distribution:
Symmetric Dirichlet is a special case of the hierarchical asymmetric Dirichlet (large concentration parameter)
Putting Everything Together

- Asymmetric hierarchical Dirichlet priors
- Integrate out $\Theta$, $\Phi$ and base measures
- Learn $z$ and concentration parameters from data
Data Sets

- Carbon nanotechnology patents:
  - Ultimate goal: track innovation and emergence
  - Fullerene and carbon nanotube patents
  - 1,016 abstracts (~100 words each)
  - 103,499 total words; 6,068 unique words
- 20 Newsgroups data (80,012 total words)
- New York Times articles (477,465 total words)
## Inferred Topics

<table>
<thead>
<tr>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>a field emission an electron ...</td>
<td>the invention of to present ...</td>
</tr>
<tr>
<td>a the carbon and gas ...</td>
<td>the of a to and ...</td>
</tr>
<tr>
<td>the of to and ...</td>
<td>metal catalytic transition catalyst from ...</td>
</tr>
<tr>
<td>the of to and ...</td>
<td>the composite polymer matrix weight fiber ...</td>
</tr>
<tr>
<td>a of to and ...</td>
<td>carbon nanotubes catalytic polymer substrate ...</td>
</tr>
<tr>
<td>the of to and ...</td>
<td>metal catalytic transition catalyst from ...</td>
</tr>
<tr>
<td>the of to and ...</td>
<td>the invention of to present ...</td>
</tr>
<tr>
<td>a field emission an electron ...</td>
<td>the of to and ...</td>
</tr>
</tbody>
</table>
Sampled Concentration Parameters
Sampled Concentration Parameters
Intuition

- Topics should be distinct from each other:
  - Asymmetric prior over topics makes topics more similar to each other (and to corpus-wide word frequencies)
  - Want a symmetric prior to preserve topic “distinctness”

- Still have to account for power-law word usage:
  - Asymmetric prior over document-specific topic distributions means some topics (e.g., “the, a, of, to ...”) can be used more often than others in all documents
“Off-the-Shelf” Topic Modeling

I can model technology emergence by analyzing patent abstracts!

Great! Let me know if you need any more help!

<table>
<thead>
<tr>
<th>the</th>
<th>carbon nanotubes</th>
<th>metal catalytic transition catalyst</th>
<th>composite polymer matrix weight fiber</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>nanotube catalyst</td>
<td>from</td>
<td>weight</td>
</tr>
<tr>
<td>of</td>
<td>substrate</td>
<td></td>
<td>fiber</td>
</tr>
<tr>
<td>to</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Great! Let me know if you need any more help!
Declining Topics

- T19 = fullerene synthesis
- T61 = (fullerenes and polymers)
- T77 = fullerene chemistry

Year: 1991 to 2004
Rising Topics

The graph shows the rising topics over time, with labels for specific topics such as carbon nanotube transistors and carbon nanotube electronics.
Analyzing Debian Mailing Lists

- women, men, debian, debian-women, men women, women men
- packages, package, debian, maintainers, maintainer, upstream
- dd, upload rights, people, dds, voting rights, rights
- debian, packages, package, sponsor, nm process, process
Building Other Tools

● Topic-based language modeling [Wallach, ICML '06]
  - Predict the next word given previous words
  - Have to model stop words

● Polylingual topic modeling [Mimno et al., EMNLP '09]
  - Track scientific progress in other countries
  - Simultaneously model text in many languages
  - Need robustness to word usage in many languages
This Talk

- Background: statistical topic models
- Building “off-the-shelf” statistical topic models
- Evaluating statistical topic models

Collaborators: Sarah Kaplan, Rotman, University of Toronto; Andrew McCallum, UMass Amherst; David Mimno, UMass Amherst; Ned Talley, NIH
Evaluating Topic Models

• Topic models are unsupervised so evaluation is hard
• A lot of topic modeling research has skirted this issue
• Easy to get a sense of topics from “eyeballing” output
  – ... but this isn't rigorous evaluation
• One common evaluation metric is the probability of held-out documents [Wallach et al., ICML '09]
• Also need expert-driven evaluation
Expert-Driven Evaluation

- Scientific policy-makers know their own domains
- Invaluable resource for model evaluation:
  - Identification of good/poor quality topics
  - Characterization of different types of topics
- Collaborative research:
  - Automated evaluation metrics
  - Prior distributions that influence model output
Evaluation of NIH Topics

- 2 experts from NIH, 150 topics (NINDS coverage)
- Collaboratively developed 3-stage evaluation protocol
- 4 classes of poor quality topics:
  - Intruded: 2 or more unrelated concepts
  - Chained: e.g., “fatty acids” → “acids” → “nucleic acids”
  - Unbalanced: mix of general and specific terms
  - Random: no clear concept represented
Evaluation Metrics

- Number of words assigned to each topic (topic size)
- Within-document co-occurrence of the top words

<table>
<thead>
<tr>
<th>Intruded</th>
<th>Chained</th>
</tr>
</thead>
<tbody>
<tr>
<td>sleep</td>
<td>cerebellar</td>
</tr>
<tr>
<td>sars</td>
<td>cerebellum</td>
</tr>
<tr>
<td>insomnia</td>
<td>pb</td>
</tr>
<tr>
<td>cov</td>
<td>purkinje</td>
</tr>
<tr>
<td>disturbances</td>
<td>ag</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Intruded</th>
<th>Chained</th>
</tr>
</thead>
<tbody>
<tr>
<td>cerebellar</td>
<td>1149</td>
<td>499</td>
</tr>
<tr>
<td>cerebellum</td>
<td>499</td>
<td>1283</td>
</tr>
<tr>
<td>pb</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>purkinje</td>
<td>318</td>
<td>228</td>
</tr>
<tr>
<td>ag</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Automated Evaluation

- Word co-occurrence-based metric:
  - 17 of 20 worst-scoring topics are “bad”
  - 18 of 20 best-scoring topics are “good”
- Goal: incorporate co-occurrence information directly into the model to prevent poor quality topics:
  - Words that do not co-occur in documents should not have high probability within the a single topic
Generalized Polya Urns

- The topic-word component of LDA is a Polya urn
- Can be replaced with a generalized Polya urn
  - Can then incorporate co-occurrence statistics directly into the model via the generalized Polya urn schema
- Relatively little computational cost beyond LDA
- Resultant topics are more coherent:
  - Much better evaluation scores (automated, humans)
Thanks!

Acknowledgements: Sarah Kaplan, University of Toronto; Andrew McCallum, University of Massachusetts Amherst; David Mimno, University of Massachusetts Amherst; Ned Talley, NIH