Probabilistic Topic Models

David M. Blei

Department of Computer Science
Princeton University

September 26, 2013
Probabilistic topic models

As more information becomes available, it becomes more difficult to find and discover what we need.

We need new tools to help us organize, search, and understand these vast amounts of information.
Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use annotations to organize, summarize, search, form predictions.
Probabilistic topic models

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

"Theoretical Physics"

FORC

RELATIVITY

LASER

"Neuroscience"

OXYGEN

NERVE

NEURON

1880 1900 1920 1940 1960 1980 2000

1880 1900 1920 1940 1960 1980 2000
Probabilistic topic models

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE

SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES
Probabilistic topic models
### Table 2: Top eight link predictions made by RTM ($\psi_e$) and LDA + Regression for two documents (italicized) from Cora. The models were fit with 10 topics. Boldfaced titles indicate actual documents cited by or citing each document. Over the whole corpus, RTM improves precision over LDA + Regression by 80% when evaluated on the first 20 documents retrieved.

<table>
<thead>
<tr>
<th>Title</th>
<th>RTM ($\psi_e$)</th>
<th>LDA + Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Markov chain Monte Carlo convergence diagnostics: A comparative review</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minorization conditions and convergence rates for Markov chain Monte Carlo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rates of convergence of the Hastings and Metropolis algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Possible biases induced by MCMC convergence diagnostics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bounding convergence time of the Gibbs sampler in Bayesian image restoration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self regenerative Markov chain Monte Carlo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auxiliary variable methods for Markov chain Monte Carlo with applications</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Rate of Convergence of the Gibbs Sampler by Gaussian Approximation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnosing convergence of Markov chain Monte Carlo algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact Bound for the Convergence of Metropolis Chains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self regenerative Markov chain Monte Carlo</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Minorization conditions and convergence rates for Markov chain Monte Carlo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gibbs-markov models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auxiliary variable methods for Markov chain Monte Carlo with applications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mediating instrumental variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A qualitative framework for probabilistic inference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptation for Self Regenerative MCMC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Probabilistic topic models

Wikipedia Topics

Relative Presence of Topics in all Documents

- (household, population, female)
- (film, series, show)
- (theory, work, human)
- (son, year, death)
- (war, force, army)
- (system, computer, user)
- (album, band, music)
- (government, party, election)
- (game, team, player)
- (god, call, give)
- (company, market, business)
- (math, number, function)
- (city, town, area)

{film, series, show}

<table>
<thead>
<tr>
<th>words</th>
<th>related documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>film</td>
<td>The X-Files</td>
</tr>
<tr>
<td>series</td>
<td>Orson Welles</td>
</tr>
<tr>
<td>show</td>
<td>Stanley Kubrick</td>
</tr>
<tr>
<td>character</td>
<td>B movie</td>
</tr>
<tr>
<td>play</td>
<td>Mystery Science Theater 3000</td>
</tr>
<tr>
<td>make</td>
<td>Monty Python</td>
</tr>
<tr>
<td>episode</td>
<td>Doctor Who</td>
</tr>
<tr>
<td>movie</td>
<td>Sam Peckinpah</td>
</tr>
<tr>
<td>good</td>
<td>Married... with Children</td>
</tr>
<tr>
<td>release</td>
<td>History of film</td>
</tr>
<tr>
<td>feature</td>
<td>The A-Team</td>
</tr>
<tr>
<td>television</td>
<td>Pulp Fiction (film)</td>
</tr>
<tr>
<td>star</td>
<td>Mad (magazine)</td>
</tr>
</tbody>
</table>

{theory, work, human}

<table>
<thead>
<tr>
<th>words</th>
<th>related documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>theory</td>
<td>Meme</td>
</tr>
<tr>
<td>work</td>
<td>Intelligent design</td>
</tr>
<tr>
<td>human</td>
<td>Immanuel Kant</td>
</tr>
<tr>
<td>idea</td>
<td>Philosophy of mathematics</td>
</tr>
<tr>
<td>term</td>
<td>History of science</td>
</tr>
<tr>
<td>study</td>
<td>Free will</td>
</tr>
<tr>
<td>view</td>
<td>Truth</td>
</tr>
<tr>
<td>science</td>
<td>Psychoanalysis</td>
</tr>
<tr>
<td>concept</td>
<td>Charles Peirce</td>
</tr>
<tr>
<td>form</td>
<td>Existentialism</td>
</tr>
<tr>
<td>world</td>
<td>Deconstruction</td>
</tr>
<tr>
<td>argue</td>
<td>Social sciences</td>
</tr>
<tr>
<td>social</td>
<td>Idealism</td>
</tr>
</tbody>
</table>

Stanley Kubrick

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 for 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...
Probabilistic topic models

• What are topic models?
• What kinds of things can they do?
• How do I compute with a topic model?
• How do I evaluate and check a topic model?
• What are some unanswered questions in this field?
• How can I learn more?
This is a case study in data analysis with probability models.

Our agenda is to teach about this kind of analysis through topic models.

Note: We are being “Bayesian” in this sense:

“[By Bayesian inference,] I simply mean the method of statistical inference that draws conclusions by calculating conditional distributions of unknown quantities given (a) known quantities and (b) model specifications.” (Rubin, 1984)

(The Bayesian versus Frequentist debate is not relevant to this talk.)
Probabilistic models

- **Specifying models**
  - Directed graphical models
  - Conjugate priors and nonconjugate priors
  - Time series modeling
  - Hierarchical methods
  - Mixed-membership models
  - Prediction from sparse and noisy inputs

- **Model selection and Bayesian nonparametric methods**

- **Approximate posterior inference**
  - MCMC
  - Variational inference

- **Using and evaluating models**
  - Exploring, describing, summarizing, visualizing data
  - Evaluating model fitness
Probabilistic models

Make assumptions

Collect data

Infer the posterior

Predict

Explore

Check
Organization of these lectures

1 Introduction to topic modeling: Latent Dirichlet allocation
2 Beyond latent Dirichlet allocation
   • Correlated and dynamic models
   • Supervised models
   • Modeling text and user data
3 Bayesian nonparametrics: A brief tutorial
4 Posterior computation
   • Scalable variational inference
   • Nonconjugate variational inference
5 Checking and evaluating models
   • Using the predictive distribution
   • Posterior predictive checks
6 Discussion, open questions, and resources
Introduction to Topic Modeling
Latent Dirichlet allocation (LDA)

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


Simple intuition: Documents exhibit multiple topics.
Latent Dirichlet allocation (LDA)

Topics
- gene 0.04
- dna 0.02
- genetic 0.01
- life 0.02
- evolve 0.01
- organism 0.01
- brain 0.04
- neuron 0.02
- nerve 0.01
- data 0.02
- number 0.02
- computer 0.01

Documents

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 eukaryotes 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 bacterium, notes John Suk et al., a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing the

• Each topic is a distribution over words
• Each document is a mixture of corpus-wide topics
• Each word is drawn from one of those topics
In reality, we only observe the documents.

The other structure are **hidden variables**

Topic modeling algorithms **infer** these variables from data.
Our goal is to **infer** the hidden variables

I.e., compute their distribution conditioned on the documents

\[ p(\text{topics, proportions, assignments} | \text{documents}) \]
LDA as a graphical model

- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data
LDA as a graphical model

- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.
LDA as a graphical model

\[
\rho(\beta, \theta, z, w) = \left( \prod_{i=1}^{K} \rho(\beta_i | \eta) \right) \left( \prod_{d=1}^{D} \rho(\theta_d | \alpha) \prod_{n=1}^{N} \rho(z_{d,n} | \theta_d) \rho(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)
\]
LDA as a graphical model

- This joint defines a posterior, $p(\theta, z, \beta | w)$.

- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions $\theta_d$
  - Per-corpus topic distributions $\beta_k$

- Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.
Approximate posterior inference algorithms

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Distributed sampling (Newman et al., 2008; Ahmed et al., 2012)
- Collapsed variational inference (Teh et al., 2006)
- Online variational inference (Hoffman et al., 2010)
- Factorization based inference (Arora et al., 2012; Anandkumar et al., 2012)
Example inference

- **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 using variational inference.
**Example 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 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 Aracdy Mushegan, 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.
### Example inference

<table>
<thead>
<tr>
<th>human</th>
<th>genome</th>
<th>evolution</th>
<th>disease</th>
<th>computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>genome</td>
<td>dna</td>
<td>evolutionary</td>
<td>host</td>
<td>models</td>
</tr>
<tr>
<td>genetic</td>
<td>genes</td>
<td>species</td>
<td>bacteria</td>
<td>data</td>
</tr>
<tr>
<td>sequence</td>
<td>gene</td>
<td>organisms</td>
<td>diseases</td>
<td>computers</td>
</tr>
<tr>
<td>gene</td>
<td>molecular</td>
<td>life</td>
<td>resistance</td>
<td>system</td>
</tr>
<tr>
<td>sequencing</td>
<td>map</td>
<td>origin</td>
<td>bacterial</td>
<td>network</td>
</tr>
<tr>
<td>information</td>
<td>diversity</td>
<td>biology</td>
<td>new</td>
<td>systems</td>
</tr>
<tr>
<td>genetics</td>
<td>group</td>
<td>groups</td>
<td>strains</td>
<td>model</td>
</tr>
<tr>
<td>mapping</td>
<td>new</td>
<td>phylogenetic</td>
<td>control</td>
<td>parallel</td>
</tr>
<tr>
<td>project</td>
<td>two</td>
<td>living</td>
<td>infectious</td>
<td>methods</td>
</tr>
<tr>
<td>sequences</td>
<td>common</td>
<td>infectious</td>
<td>malaria</td>
<td>networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>parasites</td>
<td>parasite</td>
<td>software</td>
</tr>
<tr>
<td></td>
<td></td>
<td>united</td>
<td>united</td>
<td>new</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tuberculosis</td>
<td>tuberculosis</td>
<td>simulations</td>
</tr>
<tr>
<td>Column 1</td>
<td>Column 2</td>
<td>Column 3</td>
<td>Column 4</td>
<td>Column 5</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>DNA</td>
<td>Protein</td>
<td>Water</td>
<td>Says</td>
<td>Mantle</td>
</tr>
<tr>
<td>Gene</td>
<td>Cell</td>
<td>Climate</td>
<td>Researchers</td>
<td>High</td>
</tr>
<tr>
<td>Sequence</td>
<td>Cells</td>
<td>Atmospheric</td>
<td></td>
<td>Earth</td>
</tr>
<tr>
<td>Genes</td>
<td>Proteins</td>
<td>Temperature</td>
<td></td>
<td>Pressure</td>
</tr>
<tr>
<td>Genome</td>
<td>Receptor</td>
<td>Global</td>
<td>University</td>
<td>Seismic</td>
</tr>
<tr>
<td></td>
<td>Activity</td>
<td>Surface</td>
<td>Just</td>
<td>Crust</td>
</tr>
<tr>
<td></td>
<td>Activation</td>
<td>Ocean</td>
<td>Science</td>
<td>Temperature</td>
</tr>
<tr>
<td></td>
<td>Kinase</td>
<td>Carbon</td>
<td>Like</td>
<td>Earths</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Atmosphere</td>
<td>Work</td>
<td>Lower</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Changes</td>
<td>First</td>
<td>Earthquakes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Years</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>End</td>
<td>Time</td>
<td>Materials</td>
<td>DNA</td>
<td>Disease</td>
</tr>
<tr>
<td>Article</td>
<td>Data</td>
<td>Surface</td>
<td>RNA</td>
<td>Cancer</td>
</tr>
<tr>
<td>Start</td>
<td>Two</td>
<td>High</td>
<td>Protein</td>
<td>Patients</td>
</tr>
<tr>
<td>Science</td>
<td>Model</td>
<td>Structure</td>
<td>Transcription</td>
<td>Human</td>
</tr>
<tr>
<td>Readers</td>
<td>Fig</td>
<td>Temperature</td>
<td></td>
<td>Gene</td>
</tr>
<tr>
<td>Service</td>
<td>System</td>
<td>Molecules</td>
<td></td>
<td>Medical</td>
</tr>
<tr>
<td>News</td>
<td>Number</td>
<td>Chemical</td>
<td></td>
<td>Studies</td>
</tr>
<tr>
<td>Card</td>
<td>Other</td>
<td>Molecular</td>
<td></td>
<td>Drug</td>
</tr>
<tr>
<td>Circle</td>
<td>Fig</td>
<td>Molecular</td>
<td></td>
<td>Normal</td>
</tr>
<tr>
<td>Letters</td>
<td>University</td>
<td>Structural</td>
<td></td>
<td>Drugs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years</td>
<td>Species</td>
<td>Proteins</td>
<td>Cells</td>
<td>Space</td>
</tr>
<tr>
<td>Million</td>
<td>Evolution</td>
<td>Two</td>
<td>Cell</td>
<td>Solar</td>
</tr>
<tr>
<td>Ago</td>
<td>Population</td>
<td>Amino</td>
<td>Virus</td>
<td>Stars</td>
</tr>
<tr>
<td>Age</td>
<td>Evolutionary</td>
<td>Binding</td>
<td>HIV</td>
<td>University</td>
</tr>
<tr>
<td>University</td>
<td>Populations</td>
<td>Acid</td>
<td>Immune</td>
<td>Mass</td>
</tr>
<tr>
<td>North</td>
<td>Natural</td>
<td>Residues</td>
<td>Human</td>
<td>Sun</td>
</tr>
<tr>
<td>Early</td>
<td>Studies</td>
<td>Molecular</td>
<td>Antigen</td>
<td>Astronomers</td>
</tr>
<tr>
<td>Fig</td>
<td>Genetic</td>
<td>Structural</td>
<td>Infected</td>
<td>Telescope</td>
</tr>
<tr>
<td>Evidence</td>
<td>Biology</td>
<td></td>
<td>Viral</td>
<td></td>
</tr>
<tr>
<td>Record</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tax</td>
<td>Cells</td>
<td>Energy</td>
<td>Research</td>
<td>Neurons</td>
</tr>
<tr>
<td>Manager</td>
<td>Cell</td>
<td>Electron</td>
<td></td>
<td>Brain</td>
</tr>
<tr>
<td>Science</td>
<td>Gene</td>
<td>State</td>
<td></td>
<td>Activity</td>
</tr>
<tr>
<td>AAAS</td>
<td>Genes</td>
<td>Light</td>
<td></td>
<td>Cells</td>
</tr>
<tr>
<td>Advertising</td>
<td>Expression</td>
<td>Quantum</td>
<td></td>
<td>Channels</td>
</tr>
<tr>
<td>Sales</td>
<td>Development</td>
<td>Physics</td>
<td></td>
<td>University</td>
</tr>
<tr>
<td>Member</td>
<td>Mutant</td>
<td>Electrons</td>
<td></td>
<td>Cortex</td>
</tr>
<tr>
<td>Recruitment</td>
<td>Mice</td>
<td>Electrons</td>
<td></td>
<td>Neuronal</td>
</tr>
<tr>
<td>Associate</td>
<td>Fig</td>
<td>Laser</td>
<td></td>
<td>Visual</td>
</tr>
<tr>
<td>Washington</td>
<td>Biology</td>
<td>Magnetic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Aside: The Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

\[ p(\theta | \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i-1}. \]

- It is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of \( \theta \) is a Dirichlet.

- The parameter \( \alpha \) controls the mean shape and sparsity of \( \theta \).

- The topic proportions are a \( K \) dimensional Dirichlet. The topics are a \( V \) dimensional Dirichlet.
\( \alpha = 1 \)
\( \alpha = 10 \)
\[ \alpha = 1 \]
\( \alpha = 0.01 \)
$\alpha = 0.001$
Why does LDA “work”?

- LDA trades off two goals.
  1. For each document, allocate its words to as few topics as possible.
  2. For each topic, assign high probability to as few terms as possible.

- These goals are at odds.
  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability under that topic.
  - Putting very few words in each topic makes #1 hard:
    To cover a document’s words, it must assign many topics to it.

- Trading off these goals finds groups of tightly co-occurring words.
LDA summary

- LDA is a probabilistic model of text. It casts the problem of discovering themes in large document collections as a posterior inference problem.

- It lets us visualize the hidden thematic structure in large collections, and generalize new data to fit into that structure.

- Builds on latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999) It is a mixed-membership model (Erosheva, 2004). It relates to PCA and matrix factorization (Jakulin and Buntine, 2002). Was independently invented for genetics (Pritchard et al., 2000)
LDA summary

- LDA is a simple building block that enables many applications.
- It is popular because organizing and finding patterns in data has become important in the sciences, humanities, industry, and culture.
- Further, algorithmic improvements let us fit models to massive data.
Example: LDA in R (Jonathan Chang)

docs <- read.documents("mult.dat")
K <- 20
alpha <- 1/20
eta <- 0.001
model <- lda.collapsed.gibbs.sampler(documents, K, vocab, 1000, alpha, eta)

...
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>dna</td>
<td>protein</td>
<td>water</td>
<td>says</td>
<td>mantle</td>
</tr>
<tr>
<td>gene</td>
<td>cell</td>
<td>climate</td>
<td>researchers</td>
<td>high</td>
</tr>
<tr>
<td>sequence</td>
<td>cells</td>
<td>atmospheric</td>
<td>new</td>
<td>earth</td>
</tr>
<tr>
<td>genes</td>
<td>proteins</td>
<td>temperature</td>
<td>university</td>
<td>pressure</td>
</tr>
<tr>
<td>sequences</td>
<td>receptor</td>
<td>global</td>
<td>just</td>
<td>seismic</td>
</tr>
<tr>
<td>human</td>
<td>fig</td>
<td>surface</td>
<td>science</td>
<td>crust</td>
</tr>
<tr>
<td>genome</td>
<td>binding</td>
<td>ocean</td>
<td>like</td>
<td>temperature</td>
</tr>
<tr>
<td>genetic</td>
<td>activity</td>
<td>carbon</td>
<td>work</td>
<td>earths</td>
</tr>
<tr>
<td>analysis</td>
<td>activation</td>
<td>atmosphere</td>
<td>first</td>
<td>lower</td>
</tr>
<tr>
<td>two</td>
<td>kinase</td>
<td>changes</td>
<td>years</td>
<td>earthquakes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>end</td>
<td>time</td>
<td>materials</td>
<td>dna</td>
<td>disease</td>
</tr>
<tr>
<td>article</td>
<td>data</td>
<td>surface</td>
<td>rna</td>
<td>cancer</td>
</tr>
<tr>
<td>start</td>
<td>two</td>
<td>high</td>
<td>transcription</td>
<td>patients</td>
</tr>
<tr>
<td>science</td>
<td>model</td>
<td>structure</td>
<td>protein</td>
<td>human</td>
</tr>
<tr>
<td>readers</td>
<td>fig</td>
<td>temperature</td>
<td>binding</td>
<td>gene</td>
</tr>
<tr>
<td>service</td>
<td>system</td>
<td>molecules</td>
<td>sequence</td>
<td>medical</td>
</tr>
<tr>
<td>news</td>
<td>number</td>
<td>chemical</td>
<td>proteins</td>
<td>studies</td>
</tr>
<tr>
<td>card</td>
<td>alphabet</td>
<td>molecular</td>
<td>specific</td>
<td>drug</td>
</tr>
<tr>
<td>circle</td>
<td>math</td>
<td>fig</td>
<td>sequences</td>
<td>normal</td>
</tr>
<tr>
<td>letters</td>
<td>math</td>
<td>university</td>
<td></td>
<td>drugs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>years</td>
<td>species</td>
<td>protein</td>
<td>cells</td>
<td>space</td>
</tr>
<tr>
<td>million</td>
<td>evolution</td>
<td>structure</td>
<td>cell</td>
<td>solar</td>
</tr>
<tr>
<td>ago</td>
<td>population</td>
<td>proteins</td>
<td>virus</td>
<td>observations</td>
</tr>
<tr>
<td>age</td>
<td>evolutionary</td>
<td>two</td>
<td>hiv</td>
<td>earth</td>
</tr>
<tr>
<td>university</td>
<td>university</td>
<td>amino</td>
<td>infection</td>
<td>stars</td>
</tr>
<tr>
<td>north</td>
<td>populations</td>
<td>binding</td>
<td>immune</td>
<td>university</td>
</tr>
<tr>
<td>early</td>
<td>natural</td>
<td>acid</td>
<td>human</td>
<td>mass</td>
</tr>
<tr>
<td>fig</td>
<td>studies</td>
<td>residues</td>
<td>antigen</td>
<td>sun</td>
</tr>
<tr>
<td>evidence</td>
<td>genetic</td>
<td>molecular</td>
<td>infected</td>
<td>astronomers</td>
</tr>
<tr>
<td>record</td>
<td>biology</td>
<td>structural</td>
<td>viral</td>
<td>telescope</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>tax</td>
<td>cells</td>
<td>energy</td>
<td>research</td>
<td>neurons</td>
</tr>
<tr>
<td>manager</td>
<td>cell</td>
<td>electron</td>
<td>science</td>
<td>brain</td>
</tr>
<tr>
<td>science</td>
<td>gene</td>
<td>state</td>
<td>national</td>
<td>cells</td>
</tr>
<tr>
<td>aaas</td>
<td>expression</td>
<td>light</td>
<td>scientific</td>
<td>activity</td>
</tr>
<tr>
<td>advertising</td>
<td>genes</td>
<td>quantum</td>
<td>scientists</td>
<td>channels</td>
</tr>
<tr>
<td>sales</td>
<td>development</td>
<td>physics</td>
<td>new</td>
<td>university</td>
</tr>
<tr>
<td>member</td>
<td>mutant</td>
<td>electrons</td>
<td>states</td>
<td>cortex</td>
</tr>
<tr>
<td>recruitment</td>
<td>mice</td>
<td>high</td>
<td>university</td>
<td>neuronal</td>
</tr>
<tr>
<td>associate</td>
<td>fig</td>
<td>laser</td>
<td>state</td>
<td>visual</td>
</tr>
<tr>
<td>washington</td>
<td>biology</td>
<td>magnetic</td>
<td>energy</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Open source document browser (with Allison Chaney)

Wikipedia Topics
Relative Presence of Topics in all Documents

Stanley Kubrick

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 eschews 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

{film, series, show}

words

related topics

Stanley Kubrick

series

Orson Welles

work

Mystery Science Theater 3000

theory

Doctor Who

human

Marty Pyro

idea

Sam Peckinpah

term

good

release

feature

story

company, market, business

god, call, give

wars, peace, war

(good, call, give)

history of film

posal

history of science

(on, year, death)

Man vs. Man... with Children

war, force, army

{theory, work, human}

words

related documents

related topics

theory

theory

Meme

(work, book, publish)

work

work

Intelligent design

(law, state, case)

human

human

Immanuel Kant

(on, year, death)

idea

idea

Philosophy of mathematics

(work, book, publish)

term

term

History of science

(on, year, death)

study

study

Free will

(on, year, death)

science

science

Psychology

(on, year, death)

cancept

cancept

Charles Peirce

(work, book, publish)

form

form

Existentialism

(work, book, publish)

world

world

Deconstruction

(work, book, publish)

argue

argue

Social sciences

(work, book, publish)

social

social

Idealism

(work, book, publish)
Beyond Latent Dirichlet Allocation
Extending LDA

- LDA is a simple topic model.
- It can be used to find topics that describe a corpus.
- Each document exhibits multiple topics.
- How can we build on this simple model of text?
Extending LDA

- Make assumptions
- Collect data
- Infer the posterior
- Explore
- Predict
- Check
LDA can be embedded in more complicated models, embodying further intuitions about the structure of the texts.

E.g., it can be used in models that account for syntax, authorship, word sense, dynamics, correlation, hierarchies, and other structure.
The data generating distribution can be changed. We can apply mixed-membership assumptions to many kinds of data.

E.g., we can build models of images, social networks, music, purchase histories, computer code, genetic data, and other types.
The **posterior** can be used in creative ways.

E.g., we can use inferences in information retrieval, recommendation, similarity, visualization, summarization, and other applications.
Extending LDA

- These different kinds of extensions can be combined.

- (Really, these ways of extending LDA are a big advantage of using probabilistic modeling to analyze data.)

- To give a sense of how LDA can be extended, I’ll describe several examples of extensions that my group has worked on.

- We will discuss
  - Correlated topic models
  - Dynamic topic models & measuring scholarly impact
  - Supervised topic models
  - Relational topic models
  - Ideal point topic models
  - Collaborative topic models
Correlated and Dynamic Topic Models
Correlated topic models

- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*. 
Correlated topic models

- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

\[
X \sim \mathcal{N}_K(\mu, \Sigma) \\
\theta_i \propto \exp\{x_i\}.
\]
Correlated topic models

- Draw topic proportions from a logistic normal
- This allows topic occurrences to exhibit correlation.
- Provides a “map” of topics and how they are related
- Provides a better fit to text data, but computation is more complex
LDA assumes that the order of documents does not matter.

Not appropriate for sequential corpora (e.g., that span hundreds of years)

Further, we may want to track how language changes over time.

Dynamic topic models let the topics \textit{drift} in a sequence.
Topics drift through time
Dynamic topic models

- Use a logistic normal distribution to model topics evolving over time.
- Embed it in a state-space model on the log of the topic distribution

\[
\beta_{t,k} \mid \beta_{t-1,k} \sim \mathcal{N} \left( \beta_{t-1,k}, \sigma^2 \right)
\]

\[
p(w \mid \beta_{t,k}) \propto \exp \{ \beta_{t,k} \}
\]

- As for CTMs, this makes computation more complex. But it lets us make inferences about sequences of documents.
Dynamic topic models

Original article

Sequencing the Genome, Fast
Jame C. McElhinny and Amanda A. MacKenzie

The Genome Sequencing Projects reveal the genetic makeup of organisms—bacteria, viruses, and eukaryotes—that are linked together into long double-stranded chains. Once the DNA is sequenced, these base sequences can be used to infer the entire world of genes, including those that encode for proteins. The first step in the sequencing process is to create a library of DNA fragments that are synthesized using the polymerase chain reaction (PCR). These fragments are then inserted into individual wells of a 96-well plate, which is then loaded into a sequencing instrument.

The sequencers operate in a similar manner, but they use different technologies. One common method is to use capillary electrophoresis, where the DNA fragments are separated based on size and base composition. Another method is to use shotgun sequencing, where the DNA is randomly sheared and then sequenced in parallel. These methods can be used to generate short reads or long reads, depending on the application.

Topic proportions

The proportion of genes, including 31 Mb of sequence from the human genome, that are assigned to the top 1% of genes varies from 75% to 95% across libraries. The number of reads assigned to each gene is typically large, ranging from 1 to 100,000 reads per gene.

With this information, researchers can infer the function of genes and their role in the organism. For example, in the human genome, genes that are involved in immune function and cancer have a higher proportion of reads assigned to them.

In conclusion, the sequencing of genomes has become an essential tool in biology and medicine. The ability to rapidly generate large amounts of sequence data has revolutionized our understanding of the genetic makeup of organisms and has led to the development of new technologies and applications.
Dynamic topic models

Original article

Sequencing the Genome, Fast
Javene C. Mathikiz and Amanda A. McPherron

Quantum sequencing projects reveal the genetic makeup of an organism from its DNA base, which includes all of the information necessary for the life of the organism. The base sequence contains four nucleotides: adenine, thymine, guanine, and cytosine—which are linked together into double-stranded helices. Once the DNA sequence is obtained, scientists can use this information to understand the organism's biology, disease susceptibility, and evolution.

The current technology for sequencing DNA involves the use of chemical and physical processes to degrade the DNA into smaller fragments, which are then amplified and read. This process is time-consuming and costly, with limited throughput.

In contrast, the Quantum Sequencer, a revolutionary new technology, can sequence DNA at a speed of 100 million bases per second. This is made possible by the use of quantum dots, which are molecules that can excite and emit light when they absorb light. These dots are used to detect the presence of specific DNA sequences, allowing for rapid and accurate sequencing.

The Quantum Sequencer has been shown to be reliable and cost-effective, making it a promising technology for future genetic research.

Most likely words from top topics

sequence
genome
genomes
sequences
human
gene
dna
sequencing
chromosome
regional
analysis
data
genomic
number
devices
device
devices
materials
material
current
current
high
high
gate
gate
light
light
silicon
silicon
technology
technology
electrical
electrical
fiber
fiber
power
power
based
based

data
information
network
web
computer
language
networks
time
software
system
words
algorithm
number
internet
Dynamic topic models

- 1880: electric machine, power, engine, steam, two machines, iron, battery, wire
- 1890: electric power company, steam, power, engine, electrical machine, two systems, motor, engine
- 1900: apparatus, steam power, engineering, water, construction, engineer, room, feet
- 1910: air, water, engineering apparatus room, laboratory, engineer, made, gas, tube
- 1920: apparatus, tube, air pressure, water, glass, gas, made, laboratory, mercury
- 1930: tube, apparatus, glass, air, mercury, laboratory, pressure, made, gas, small
- 1940: air, tube, apparatus, glass, laboratory, rubber, pressure, small, mercury, gas
- 1950: tube, apparatus, glass, air, chamber, instrument, small, laboratory, pressure, rubber
- 1960: tube, system, temperature, air, heat, chamber, power, high, instrument, control
- 1970: air, heat, power, system, temperature, chamber, high, flow, tube, design
- 1980: high, power, design, heat, system, systems, devices, instruments, control, large
- 1990: materials, high power, current, applications, technology, devices, design, device, heat
- 2000: devices, device, materials, current, gate, high, light, silicon, material, technology
Dynamic topic models

"Theoretical Physics"

FORCE
RELATIVITY
LASER

"Neuroscience"

OXYGEN
NERVE
NEURON

1880 1900 1920 1940 1960 1980 2000
Dynamic topic models

- **Time-corrected similarity** shows a new way of using the posterior.
- Consider the expected Hellinger distance between the topic proportions of two documents,

\[
d_{ij} = \mathbb{E} \left[ \sum_{k=1}^{K} (\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}})^2 \mid w_i, w_j \right]
\]
- Uses the latent structure to define similarity
- Time has been factored out because the topics associated to the components are different from year to year.
- Similarity based only on topic proportions
The Brain of the Orang (1880)

In the brain, the parietal convolution passes over the cerebellum, being separated from it by the transverse fissure of the Sylvian fissure. The parietal lobes are separated by the longitudinal fissure of the Sylvian fissure, and the occipital lobes are separated by the transverse fissure of the Sylvian fissure. The parietal lobes are separated by the transverse fissure of the Sylvian fissure, and the occipital lobes are separated by the transverse fissure of the Sylvian fissure.
Dynamic topic models

Representation of the Visual Field on the Medial Wall of Occipital-Parietal Cortex in the Owl Monkey (1976)

...
We built on the DTM to measure **scholarly impact** with sequences of text.

Influential articles reflect future changes in language use.

The “influence” of an article is a latent variable.

Influential articles affect the drift of the topics that they discuss.

The posterior gives a retrospective estimate of influential articles.
Per-document influence
Measuring scholarly impact

- Each document has an influence score $I_d$.
- Each topic drifts in a way that is biased towards the documents with high influence.
- We can examine the posterior of the influence scores to retrospectively find articles that best explain the changes in language.
Measuring scholarly impact

<table>
<thead>
<tr>
<th>Number of topics</th>
<th>Correlation to citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.15</td>
</tr>
<tr>
<td>40</td>
<td>0.20</td>
</tr>
<tr>
<td>60</td>
<td>0.25</td>
</tr>
<tr>
<td>80</td>
<td>0.30</td>
</tr>
<tr>
<td>100</td>
<td>0.35</td>
</tr>
</tbody>
</table>

- This measure of impact only uses the words of the documents. It correlates strongly with citation counts.
- “Low” impact, high citation: “Building a large annotated corpus of English: the Penn Treebank” (Marcus et al., 1993)
Measuring scholarly impact

- PNAS, Science, and Nature from 1880–2005
- 350,000 Articles
- 163M observations
- Year-corrected correlation is 0.166
The Dirichlet assumption on topics and topic proportions makes strong conditional independence assumptions about the data.

The **correlated topic model** uses a logistic normal on the topic proportions to find patterns in how topics tend to co-occur.

The **dynamic topic model** uses a logistic normal in a linear dynamic model to capture how topics change over time.

What’s the catch? These models are harder to compute with. (Stay tuned.)
Supervised Topic Models
Supervised LDA

• LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?

• Many data are paired with **response variables**.
  - User reviews paired with a number of stars
  - Web pages paired with a number of “likes”
  - Documents paired with links to other documents
  - Images paired with a category

• **Supervised LDA** are topic models of documents and responses. They are fit to find topics predictive of the response.
Supervised LDA

1. Draw topic proportions $\theta | \alpha \sim \text{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n | \theta \sim \text{Mult}(\theta)$.
   - Draw word $w_n | z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$.
3. Draw response variable $y | z_{1:N}, \eta, \sigma^2 \sim \mathcal{N}\left(\eta^T \bar{z}, \sigma^2\right)$, where
   \[
   \bar{z} = \left(1/N\right) \sum_{n=1}^{N} z_n.
   \]
Supervised LDA

- Fit sLDA parameters to documents and responses. This gives: topics $\beta_{1:K}$ and coefficients $\eta_{1:K}$.
- Given a new document, predict its response using the expected value:

$$E[Y \mid w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2] = \eta^\top E[\tilde{Z} \mid w_{1:N}]$$

- This blends generative and discriminative modeling.
• 10-topic sLDA model on movie reviews (Pang and Lee, 2005).
• Response: number of stars associated with each review
• Each component of coefficient vector $\eta$ is associated with a topic.
Supervised LDA

Model
lda
slda

Number of topics
Correlation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
5 10 15 20 25 30

Model
lda
slda
SLDA enables model-based regression where the predictor is a document.

It can easily be used wherever LDA is used in an unsupervised fashion (e.g., images, genes, music).

SLDA is a supervised dimension-reduction technique, whereas LDA performs unsupervised dimension reduction.
• SLDA has been extended to generalized linear models, e.g., for image classification and other non-continuous responses.

• We will discuss two extensions of sLDA
  - **Relational topic models**: Models of networks and text
  - **Ideal point topic models**: Models of legislative voting behavior
Relational topic models

Many data sets contain **connected observations**.

For example:

- Citation networks of documents
- Friend-connected social network profiles
• Research has focused on finding communities and patterns in the link-structure of these networks. But this ignores content.

• We adapted sLDA to pairwise response variables. This leads to a model of **content and connection**.

• Relational topic models find related hidden structure in both types of data.
Adapt fitting algorithm for sLDA with binary GLM response
- RTMs allow predictions about new and unlinked data.
- These predictions are out of reach for traditional network models.
### Relational topic models

<table>
<thead>
<tr>
<th>Markov chain Monte Carlo convergence diagnostics: A comparative review</th>
<th>RTM (ψ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minorization conditions and convergence rates for Markov chain Monte Carlo</td>
<td>Rates of convergence of the Hastings and Metropolis algorithms</td>
</tr>
<tr>
<td>Possible biases induced by MCMC convergence diagnostics</td>
<td>Bounding convergence time of the Gibbs sampler in Bayesian image restoration</td>
</tr>
<tr>
<td>Self regenerative Markov chain Monte Carlo</td>
<td></td>
</tr>
<tr>
<td>Auxiliary variable methods for Markov chain Monte Carlo with applications</td>
<td></td>
</tr>
<tr>
<td>Rate of Convergence of the Gibbs Sampler by Gaussian Approximation</td>
<td>Diagnosing convergence of Markov chain Monte Carlo algorithms</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exact Bound for the Convergence of Metropolis Chains</th>
<th>LDA + Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self regenerative Markov chain Monte Carlo</td>
<td></td>
</tr>
<tr>
<td>Gibbs-markov models</td>
<td></td>
</tr>
<tr>
<td>Minorization conditions and convergence rates for Markov chain Monte Carlo</td>
<td>Auxiliary variable methods for Markov chain Monte Carlo with applications</td>
</tr>
<tr>
<td>Gibbs-markov models</td>
<td></td>
</tr>
<tr>
<td>Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models</td>
<td>Mediating instrumental variables</td>
</tr>
<tr>
<td>Mediating instrumental variables</td>
<td></td>
</tr>
<tr>
<td>A qualitative framework for probabilistic inference</td>
<td></td>
</tr>
<tr>
<td>Adaptation for Self Regenerative MCMC</td>
<td></td>
</tr>
</tbody>
</table>

Given a new document, which documents is it likely to link to?
### Table 2

| Top eight link predictions made by RTM ($\psi_e$) and LDA + Regression for two documents (italicized) from Cora. The models were fit with 10 topics. Boldfaced titles indicate actual documents cited by or citing each document. Over the whole corpus, RTM improves precision over LDA + Regression by 80% when evaluated on the first 20 documents retrieved. |

<table>
<thead>
<tr>
<th>Competitive environments evolve better solutions for complex tasks</th>
<th>Coevolving High Level Representations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic Algorithms in Search, Optimization and Machine Learning</td>
<td>A Survey of Evolutionary Strategies</td>
</tr>
<tr>
<td>Strongly typed genetic programming in evolving cooperation strategies</td>
<td>Solving combinatorial problems using evolutionary algorithms</td>
</tr>
<tr>
<td></td>
<td>A promising genetic algorithm approach to job-shop scheduling…</td>
</tr>
<tr>
<td></td>
<td>Evolutionary Module Acquisition</td>
</tr>
<tr>
<td></td>
<td>An Empirical Investigation of Multi-Parent Recombination Operators…</td>
</tr>
</tbody>
</table>

| | RTM ($\psi_e$) |

<table>
<thead>
<tr>
<th>A New Algorithm for DNA Sequence Assembly</th>
<th>Identification of protein coding regions in genomic DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solving combinatorial problems using evolutionary algorithms</td>
<td>A promising genetic algorithm approach to job-shop scheduling…</td>
</tr>
<tr>
<td></td>
<td>A genetic algorithm for passive management</td>
</tr>
<tr>
<td></td>
<td>The Performance of a Genetic Algorithm on a Chaotic Objective Function</td>
</tr>
<tr>
<td></td>
<td>Adaptive global optimization with local search</td>
</tr>
<tr>
<td></td>
<td>Mutation rates as adaptations</td>
</tr>
</tbody>
</table>

Given a new document, which documents is it likely to link to?
The **ideal point model** uncovers voting patterns in legislative data. We observe roll call data \( v_{ij} \).

- Bills attached to discrimination parameters \( a_j \).
- Senators attached to ideal points \( x_i \).
Posterior inference reveals the political spectrum of senators

Widely used in quantitative political science.
Ideal point topic models

• We can predict a missing vote.
• But we cannot predict all the missing votes from a bill.
• Cf. the limitations of collaborative filtering
• Use supervised LDA to predict bill discrimination from bill text.
• But this is a latent response.
Ideal point topic models

\[ \theta_d \]
\[ Z_{dn} \]
\[ W_{dn} \]
\[ \beta_k \]
\[ \alpha \]
\[ \eta \]
\[ \sigma^2_d \]
\[ \sigma^2_u \]
\[ A_d, B_d \]
\[ V_{ud} \]
\[ X_u \]
\[ N \]
\[ D \]
\[ U \]

Bill content  Bill sentiment  Votes  Ideal points
In addition to senators and bills, IPTM places **topics** on the spectrum.
Many documents are associated with response variables.

**Supervised LDA** embeds LDA in a generalized linear model that is conditioned on the latent topic assignments.

**Relational topic models** use sLDA assumptions with pair-wise responses to model networks of documents.

**Ideal point topic models** demonstrates how the response variables can themselves be latent variables. In this case, they are used downstream in a model of legislative behavior.

(SLDA, the RTM, and others are implemented in the R package “lda.”)
Modeling User Data and Text
In many settings, we have information about **how people use documents**.

With new models, this can be used to

- Help people find documents that they are interested in
- Learn about what the documents mean to the people reading them
- Learn about the people reading (or voting on) the documents.

(We also saw this in ideal point topic models.)
Online communities of scientists’ allow for new ways of connecting researchers to the research literature.

With **collaborative topic models**, we recommend scientific articles based both on other scientists’ preferences and their content.

We can form both “in-matrix” and “out-of-matrix” predictions. We can learn about which articles are important, and which are interdisciplinary.
Consider EM (Dempster et al., 1977). The text lets us estimate its topics:

With user data, we adjust the topics to account for who liked it:

We can then recommend to users:
Topic models for recommendation

- \( \theta_d \)
- \( \alpha \)
- \( \sigma_d^2 \)
- \( \sigma_u^2 \)
- \( Z_{dn} \)
- \( W_{dn} \)
- \( \beta_k \)
- \( N \)
- \( D \)
- \( K \)
- \( \zeta_d \)
- \( V_{ud} \)
- \( X_u \)
- \( U \)

User Preferences

Ratings

Article content

Topic correction

User Preferences
Topic models for recommendation

- Big data set from Mendeley.com
- Fit the model with **stochastic optimization**
- The data—
  - 261K documents
  - 80K users
  - 10K vocabulary terms
  - 25M observed words
  - 5.1M entries (sparsity is 0.02%)
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.
Can make predictions about current articles and new articles
More than recommendation

- The users also **tell us about the data**.

- We can look at posterior estimates to find
  - Widely read articles in a field
  - Articles in a field that are widely read in other fields
  - Articles from other fields that are widely read in a field

- These kinds of explorations require **interpretable dimensions**. They are not possible with classical matrix factorization.
# Maximum Likelihood Estimation

<table>
<thead>
<tr>
<th>Topic</th>
<th>estimates, likelihood, maximum, parameters, method</th>
</tr>
</thead>
</table>
| In-topic, **read in topic** | Maximum Likelihood Estimation of Population Parameters  
Bootstrap Methods: Another Look at the Jackknife  
R. A. Fisher and the Making of Maximum Likelihood |
| In-topic, **read in other topics** | Maximum Likelihood from Incomplete Data with the EM Algorithm  
Bootstrap Methods: Another Look at the Jackknife  
Tutorial on Maximum Likelihood Estimation |
| Out-of-topic, **read in topic** | Random Forests  
Identification of Causal Effects Using Instrumental Variables  
Matrix Computations |
<table>
<thead>
<tr>
<th>Topic</th>
<th>networks, topology, connected, nodes, links, degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-topic, read in topic</td>
<td>Assortative Mixing in Networks</td>
</tr>
<tr>
<td></td>
<td>Characterizing the Dynamical Importance of Network Nodes and Links</td>
</tr>
<tr>
<td></td>
<td>Subgraph Centrality in Complex Networks</td>
</tr>
<tr>
<td>In-topic, read in other topics</td>
<td>Assortative Mixing in Networks</td>
</tr>
<tr>
<td></td>
<td>The Structure and Function of Complex Networks</td>
</tr>
<tr>
<td></td>
<td>Statistical Mechanics of Complex Networks</td>
</tr>
<tr>
<td>Out-of-topic, read in topic</td>
<td>Power Law Distributions in Empirical Data</td>
</tr>
<tr>
<td></td>
<td>Graph Structure in the Web</td>
</tr>
<tr>
<td></td>
<td>The Origins of Bursts and Heavy Tails in Human Dynamics</td>
</tr>
</tbody>
</table>
Our earlier ideal point model uses topics to predict votes from new bills. Alternatively, we can use the text to characterize how legislators diverge from their usual ideal points. For example: A senator might be left wing, but vote conservatively when it comes to economic matters.
Issue-adjusted ideal points

\[ \alpha \quad \eta \quad \sigma^2 \]

\[ \theta_d \quad A_d, B_d \quad \sigma_d^2 \]

\[ Z_{dn} \quad W_{dn} \quad \sigma_u^2 \]

\[ W_{dn} \quad N \]

\[ D \]

\[ \beta_k \]

\[ K \]

\[ \text{Bill content} \]

\[ \text{Bill sentiment} \]

\[ \text{Global ideal point} \]

\[ \text{Observed votes} \]

\[ \text{Issue adjustments} \]
Extending LDA

New applications—
- Syntactic topic models
- Topic models on images
- Topic models on social network data
- Topic models on music data
- Topic models for recommendation systems

Testing and relaxing assumptions—
- Spike and slab priors
- Models of word contagion
- N-gram topic models
• Each of these models is tailored to solve a problem.
  • Some problems arise from new kinds of data.
  • Others arise from an issue with existing models.
• Probabilistic modeling is a *flexible and modular language for designing solutions to specific problems.*
Extending LDA

- Make assumptions
- Infer the posterior
- Explore
- Collect data
- Predict
- Check
Bayesian Nonparametric Models
Bayesian nonparametric models

- Why Bayesian nonparametric models?
- The Chinese restaurant process
- Chinese restaurant process mixture models
- The Chinese restaurant franchise
- Bayesian nonparametric topic models
- Random measures and stick-breaking constructions
Why Bayesian nonparametric models?

• Topic models assume that the number of topics is fixed.

• It is a type of **regularization parameter**. It can be determined by cross validation and other model selection techniques.

• Bayesian nonparametric methods skirt model selection—
  • The data determine the number of topics during inference.
  • Future data can exhibit new topics.

• (This is a field unto itself, but has found wide application in topic modeling.)
The Chinese restaurant process (CRP)

- $N$ customers arrive to an infinite-table restaurant. Each sits down according to how many people are sitting at each table,

$$p(z_i = k | z_{1:(i-1)}, \alpha) \propto \begin{cases} n_k & \text{for } k \leq K \\ \alpha & \text{for } k = K + 1. \end{cases}$$

- The resulting seating plan provides a partition

- This distribution is **exchangeable**: Seating plan probabilities are the same regardless of the order of customers (Pitman, 2002).
CRP mixture models

• Associate each table with a topic ($\beta^*$).
  Associate each customer with a data point (grey node).

• The number of clusters is infinite a priori;
  the data determines the number of clusters in the posterior.

• Further: the next data point might sit at new table.

• Exchangeability makes inference easy (Escobar and West, 1995; Neal, 2000).
The CRP is not a mixed-membership model

- Mixture models draw each data point from one component.
- The advantage of LDA is that it's a mixed-membership model.
- This is addressed by the Chinese restaurant franchise.
At the corpus level, topics are drawn from a prior.

Each document-level table is associated with a customer at the corpus level restaurant.

Each word is associated with a customer at the document's restaurant. It is drawn from the topic that its table is associated with.
The CRF selects the “right” number of topics (Teh et al., 2006)
Extended to find hierarchies (Blei et al., 2010)
Random measures

- The CRP metaphors are the best first way to understand BNP methods.
- BNP models were originally developed as random measure models.
- E.g., data drawn independently from a random distribution:

\[
G \sim \text{DP}(\alpha G_0) \\
X_n \sim G
\]

- The random measure perspective helps with certain applications (such as the BNP correlated topic model) and for some approaches to inference.
The Dirichlet process \( (\text{Ferguson, 1973}) \)

- The Dirichlet process is a distribution of distributions, \( G \sim \text{DP}(\alpha, G_0) \)
  - \textit{concentration parameter} \( \alpha \) (a positive scalar)
  - \textit{base distribution} \( G_0 \).

- It produces distributions defined on the same space as its base distribution.
The Dirichlet process (Ferguson, 1973)

- Consider a partition of the probability space \((A_1, \ldots, A_K)\).

- Ferguson: If for all partitions,

\[
\langle G(A_1), \ldots, G(A_k) \rangle \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))
\]

then \(G\) is distributed with a Dirichlet process.

- Note: In this process, the random variables \(G(A_k)\) are indexed by the Borel sets of the probability space.
The Dirichlet process (Ferguson, 1973)

- $G$ is discrete; it places its mass on a countably infinite set of atoms.
- The distribution of the locations is the base distribution $G_0$.
- As $\alpha$ gets large, $G$ looks more like $G_0$.
- The conditional $P(G|X_{1:N})$ is a Dirichlet process.
The Dirichlet process (Ferguson, 1973)

- Marginalizing out $G$ reveals the **clustering property**.
- The joint distribution of $X_{1:N}$ will exhibit fewer than $N$ unique values.
- These unique values are drawn from $G_0$.
- The distribution of the partition structure is a $\text{CRP}(\alpha)$.
The draw from $G$ can be a latent parameter to an observed variable:

$$G \sim \text{DP}(\alpha, G_0)$$
$$\theta_n \sim G$$
$$x_n \sim p(\cdot | \theta_n).$$

This smooths the random discrete distribution to a \textit{DP mixture}.

Because of the clustering property, marginalizing out $G$ reveals that this model is the same as a CRP mixture.
The hierarchical Dirichlet process (HDP) models grouped data.

\[ G_0 \sim \text{DP}(\gamma, H) \]
\[ G_m \sim \text{DP}(\alpha, G_0) \]
\[ \theta_{mn} \sim G_m \]
\[ x_{mn} \sim p(\cdot | \theta_{mn}) \]

Marginalizing out \( G_0 \) and \( G_m \) reveals the Chinese restaurant franchise.
Hierarchical Dirichlet processes (Teh et al., 2006)

- In topic modeling—
  - The atoms of $G_0$ are all the topics.
  - Each $G_m$ is a document-specific distribution over those topics.
  - The variable $\theta_{mn}$ is a topic drawn from $G_m$.
  - The observation $x_{mn}$ is a word drawn from the topic $\theta_{mn}$.

- Note that in the original topic modeling story, we worked with pointers to topics. Here the $\theta_{mn}$ variables are distributions over words.
• Bayesian nonparametric modeling is a growing field (Hjort et al., 2011).

• BNP methods can define priors over latent combinatorial structures.

• In the posterior, the documents determine the particular form of the structure that is best for the corpus at hand.

• **Recent innovations:**
  • Improved inference (Blei and Jordan, 2006, Wang et al. 2011)
  • BNP models for language (Teh, 2006; Goldwater et al., 2011)
  • Dependent models, such as time series models (MacEachern 1999, Dunson 2010, Blei and Frazier 2011)
  • Predictive models (Hannah et al. 2011)
  • Factorization models (Griffiths and Ghahramani, 2011)
Posterior Inference
• We can express many kinds of assumptions.
• How can we analyze the collection under those assumptions?
Posterior inference is the main computational problem.

Inference links observed data to statistical assumptions.

Inference on large data is crucial for topic modeling applications.
• Our goal is to compute the distribution of the hidden variables conditioned on the documents

\[ p(\text{topics, proportions, assignments} | \text{documents}) \]
Posterior inference for LDA

The joint distribution of the latent variables and documents is

$$\prod_{i=1}^{K} p(\beta_i | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \left( \prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right).$$

The posterior of the latent variables given the documents is

$$p(\beta, \theta, z | w).$$
This is equal to

\[
\frac{p(\beta, \theta, z, w)}{\int_{\beta} \int_{\theta} \sum_{z} p(\beta, \theta, z, w)}.
\]

We can’t compute the denominator, the marginal \( p(w) \).

This is the crux of the inference problem.
There is a large literature on approximating the posterior, both within topic modeling and Bayesian statistics in general.

We will focus on mean-field variational methods.

We will derive stochastic variational inference, a generic approximate inference method for very large data sets.
Variational inference

- Variational inference turns posterior inference into **optimization**.

- The main idea—
  - Place a distribution over the hidden variables with free parameters, called **variational parameters**.
  - Optimize the variational parameters to make the distribution close (in KL divergence) to the true posterior

- Variational inference can be faster than sampling-based approaches.

- It is easier to handle **nonconjugate** models with variational inference. (This is important in the CTM, DTM, and legislative models.)

- It can be scaled up to very large data sets with **stochastic optimization**.
We want to condition on large data sets and approximate the posterior.

In **variational inference**, we optimize over a family of distributions to find the member closest in KL divergence to the posterior.

Variational inference usually results in an algorithm like this:

- Infer local variables for each data point.
- Based on these local inferences, re-infer global variables.
- Repeat.
• This is inefficient. We should know something about the global structure after seeing part of the data.

• And, it assumes a finite amount of data. We want algorithms that can handle data sources, information arriving in a constant stream.

• With stochastic variational inference, we can condition on large data and approximate the posterior of complex models.
Stochastic variational inference

- The structure of the algorithm is:
  - Subsample the data—one data point or a small batch.
  - Infer local variables for the subsample.
  - Update the current estimate of the posterior of the global variables.
  - Repeat.

- This is **efficient**—we need only process one data point at a time.

- We will show: Just as easy as “classical” variational inference
Stochastic variational inference for LDA

1. Sample a document $w_d$ from the collection
2. Infer how $w_d$ exhibits the current topics
3. Create intermediate topics, formed as though the $w_d$ is the only document.
4. Adjust the current topics according to the intermediate topics.
5. Repeat.
Stochastic variational inference for LDA

Documents seen (log scale)

Perplexity

Documents analyzed

Top eight words

2048 systems made service announced national west
4096 systems health communication service billion language
care road
8192 service service systems companies business
companies billion health industry
12288 service service systems companies business
companies billion health industry
16384 service service systems companies business
companies billion health industry
32768 business service companies business
companies industry management
49152 business service companies business
companies industry management
65536 business service companies business
companies industry management public

Batch 98K
Online 98K
Online 3.3M
We have developed stochastic variational inference algorithms for

- Latent Dirichlet allocation
- The hierarchical Dirichlet process
- The discrete infinite logistic normal
- Mixed-membership stochastic blockmodels
- Bayesian nonparametric factor analysis
- Recommendation models and legislative models
• Describe a generic class of models
• Derive mean-field variational inference in this class
• Derive natural gradients for the variational objective
• Review stochastic optimization
• Derive stochastic variational inference
• We consider a **generic model**.
  • Hidden variables are local or global.

• We use **variational inference**.
  • Optimize a simple proxy distribution to be close to the posterior
  • Closeness is measured with Kullback-Leibler divergence

• Solve the optimization problem with **stochastic optimization**.
  • Stochastic gradients are formed by subsampling from the data.
The observations are $x = x_1:n$.

The **local** variables are $z = z_1:n$.

The **global** variables are $\beta$.

The $i$th data point $x_i$ only depends on $z_i$ and $\beta$.

Our goal is to compute $p(\beta, z | x)$. 

$$p(\beta, z_1:n, x_1:n) = p(\beta) \prod_{i=1}^{n} p(z_i | \beta) p(x_i | z_i, \beta)$$
A complete conditional is the conditional of a latent variable given the observations and other latent variable.

Assume each complete conditional is in the exponential family,

\[
p(z_i | \beta, x_i) = h(z_i) \exp \{ \eta_\ell(\beta, x_i)^T z_i - a(\eta_\ell(\beta, x_i)) \}
\]

\[
p(\beta | z, x) = h(\beta) \exp \{ \eta_g(z, x)^T \beta - a(\eta_g(z, x)) \}.
\]
Generic model

\[
p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i | \beta) p(x_i | z_i, \beta)
\]

- Bayesian mixture models
- Time series models (variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization (e.g., factor analysis, PCA, CCA)

- Dirichlet process mixtures, HDPs
- Multilevel regression (linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models (LDA and some variants)
Mean-field variational inference

- Introduce a **variational distribution** over the latent variables $q(\beta, z)$.
- We optimize the **evidence lower bound** (ELBO) with respect to $q$,

\[
\log p(x) \geq E_q[\log p(\beta, Z, x)] - E_q[\log q(\beta, Z)].
\]

- Up to a constant, this is the negative KL between $q$ and the posterior.
We can derive the ELBO with Jensen’s inequality:

\[
\log p(x) = \log \int p(\beta, Z, x) dZd\beta \\
= \log \int p(\beta, Z, x) \frac{q(\beta, Z)}{q(\beta, Z)} dZd\beta \\
\geq \int q(\beta, Z) \log \frac{p(\beta, Z, x)}{q(\beta, Z)} dZd\beta \\
= E_q[\log p(\beta, Z, x)] - E_q[\log q(\beta, Z)].
\]
Mean-field variational inference

- We specify $q(\beta, z)$ to be a fully factored variational distribution,
  $$q(\beta, z) = q(\beta | \lambda) \prod_{i=1}^{n} q(z_i | \phi_i).$$

- Each instance of each variable has its own distribution.

- Each component is in the same family as the model conditional,
  $$p(\beta | z, x) = h(\beta) \exp\{\eta_g(z, x)^\top \beta - a(\eta_g(z, x))\}$$
  $$q(\beta | \lambda) = h(\beta) \exp\{\lambda^\top \beta - a(\lambda)\}$$

(And, same for the local variational parameters.)
Mean-field variational inference

- We optimize the ELBO with respect to these parameters,

\[ \mathcal{L}(\lambda, \phi_{1:n}) = E_q[\log p(\beta, Z, x)] - E_q[\log q(\beta, Z)]. \]

- Same as finding the \( q(\beta, z) \) that is closest in KL divergence to \( p(\beta, z | x) \)

- The ELBO links the observations/model to the variational distribution.
Coordinate ascent: Iteratively update each parameter, holding others fixed.

With respect to the global parameter, the gradient is

\[ \nabla_\lambda \mathcal{L} = a''(\lambda)(E_\phi[\eta_g(Z, x)] - \lambda). \]

This leads to a simple coordinate update

\[ \lambda^* = E_\phi[\eta_g(Z, x)]. \]

The local parameter is analogous.
Mean-field variational inference

Initialize $\lambda$ randomly.
Repeat until the ELBO converges

1. For each data point, update the local variational parameters:
   \[
   \phi_i^{(t)} = E_{\lambda^{(t-1)}} [\eta_\ell (\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}.
   \]

2. Update the global variational parameters:
   \[
   \lambda^{(t)} = E_{\phi^{(t)}} [\eta_g (Z_{1:n}, x_{1:n})].
   \]
Mean-field variational inference for LDA

- Document variables: Topic proportions \( \theta \) and topic assignments \( z_{1:N} \).
- Corpus variables: Topics \( \beta_{1:K} \)
- The variational distribution is

\[
q(\beta, \theta, z) = \prod_{k=1}^{K} q(\beta_k | \lambda_k) \prod_{d=1}^{D} q(\theta_d | \gamma_d) \prod_{n=1}^{N} q(z_{d,n} | \phi_{d,n})
\]
In the “local step” we iteratively update the parameters for each document, holding the topic parameters fixed.

\[
\gamma^{(t+1)} = \alpha + \sum_{n=1}^{N} \phi_n^{(t)}
\]

\[
\phi_{n}^{(t+1)} \propto \exp\{\mathbb{E}_q[\log \theta] + \mathbb{E}_q[\log \beta_{:,w_n}]\}.
\]
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, even 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.
In the “global step” we aggregate the parameters computed from the local step and update the parameters for the topics,

\[ \lambda_k = \eta + \sum_d \sum_n w_{d,n} \phi_{d,n}. \]
Mean-field variational inference for LDA

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

1: Initialize topics randomly.
2: repeat
3: for each document do
4:   repeat
5:     Update the topic assignment variational parameters.
6:     Update the topic proportions variational parameters.
7:   until document objective converges
8: end for
9: Update the topics from aggregated per-document parameters.
10: until corpus objective converges.
Initialize $\lambda$ randomly.
Repeat until the ELBO converges

1. Update the local variational parameters for each data point,
   \[ \phi_i^{(t)} = \mathbb{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}. \]

2. Update the global variational parameters,
   \[ \lambda^{(t)} = \mathbb{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})]. \]

- Note the relationship to existing algorithms like EM and Gibbs sampling.
- But we must analyze the whole data set before completing one iteration.
Initialize $\lambda$ randomly.
Repeat until the ELBO converges

1. Update the local variational parameters for each data point,
   
   $$\phi_i^{(t)} = E_{\lambda^{(t-1)}}[\eta_{\ell}(\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}.$$ 

2. Update the global variational parameters,
   
   $$\lambda^{(t)} = E_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})].$$

To make this more efficient, we need two ideas:

- Natural gradients
- Stochastic optimization
• In natural gradient ascent, we premultiply the gradient by the inverse of a Riemannian metric. Amari (1998) showed this is the steepest direction.
• For distributions, the Riemannian metric is the Fisher information.
• In the exponential family, the Fisher information is the second derivative of the log normalizer,

\[ G = a''(\lambda). \]

• So, the natural gradient of the ELBO is

\[ \hat{\nabla}_\lambda \mathcal{L} = E_\phi [\eta_g(Z, x)] - \lambda. \]

• We can compute the natural gradient by computing the coordinate updates in parallel and subtracting the current variational parameters.
Why waste time with the real gradient, when a cheaper noisy estimate of the gradient will do (Robbins and Monro, 1951)?

Idea: Follow a noisy estimate of the gradient with a step-size.

By decreasing the step-size according to a certain schedule, we guarantee convergence to a local optimum.
We will use stochastic optimization for global variables.

Let $\nabla_\lambda \mathcal{L}_t$ be a realization of a random variable whose expectation is $\mathbb{E}[\nabla_\lambda \mathcal{L}]$.

Iteratively set

$$\lambda(t) = \lambda(t-1) + \epsilon_t \nabla_\lambda \mathcal{L}_t$$

This leads to a local optimum when

$$\sum_{t=1}^\infty \epsilon_t = \infty \quad \text{and} \quad \sum_{t=1}^\infty \epsilon_t^2 < \infty$$

Next step: Form a noisy gradient.
We need to look more closely at the conditional distribution of the global hidden variable given the local hidden variables and observations.

The form of the local joint distribution is
\[
p(z_i, x_i \mid \beta) = h(z_i, x_i) \exp\{\beta^\top f(z_i, x_i) - a(\beta)\}.
\]

This means the conditional parameter of \(\beta\) is
\[
\eta_g(z_{1:n}, x_{1:n}) = \langle \alpha_1 + \sum_{i=1}^n f(z_i, x_i), \alpha_2 + n \rangle.
\]

See the discussion of conjugacy in Bernardo and Smith (1994).
A noisy natural gradient

- With local and global variables, we decompose the ELBO
  \[ L = \mathbb{E}[\log p(\beta)] - \mathbb{E}[\log q(\beta)] + \sum_{i=1}^{n} \mathbb{E}[\log p(z_i, x_i | \beta)] - \mathbb{E}[\log q(z_i)] \]

- Sample a single data point \( t \) uniformly from the data and define
  \[ L_t = \mathbb{E}[\log p(\beta)] - \mathbb{E}[\log q(\beta)] + n(\mathbb{E}[\log p(z_t, x_t | \beta)] - \mathbb{E}[\log q(z_t)]). \]

1. The ELBO is the expectation of \( L_t \) with respect to the sample.
2. The gradient of the \( t \)-ELBO is a noisy gradient of the ELBO.
3. The \( t \)-ELBO is like an ELBO where we saw \( x_t \) repeatedly.
A noisy natural gradient

- Define the conditional as though our whole data set is $n$ replications of $x_t$,
  \[ \eta_t(z_t, x_t) = \langle \alpha_1 + n \cdot f(z_t, x_t), \alpha_2 + n \rangle \]

- The noisy natural gradient of the ELBO is
  \[ \nabla_\lambda \hat{\mathcal{L}}_t = E_{\phi_t}[\eta_t(Z_t, x_t)] - \lambda. \]

- This only requires the local variational parameters of one data point.

- In contrast, the full natural gradient requires all local parameters.
Initialize global parameters $\lambda$ randomly.
Set the step-size schedule $\epsilon_t$ appropriately.
Repeat forever

1. Sample a data point uniformly,
   \[ x_t \sim \text{Uniform}(x_1, \ldots, x_n). \]

2. Compute its local variational parameter,
   \[ \phi = E_{\lambda(t-1)}[\eta_\ell(\beta, x_t)]. \]

3. Pretend it's the only data point in the data set,
   \[ \hat{\lambda} = E_{\phi}[\eta_t(Z_t, x_t)]. \]

4. Update the current global variational parameter,
   \[ \lambda(t) = (1 - \epsilon_t)\lambda^{(t-1)} + \epsilon_t \hat{\lambda}. \]
1. Sample a document
2. Estimate the local variational parameters using the current topics
3. Form “fake topics” from those local parameters
4. Update the topics to be a weighted average of “fake” and current topics
1: Define $\rho_t \overset{\Delta}{=} (\tau_0 + t)^{-\kappa}$
2: Initialize $\lambda$ randomly.
3: for $t = 0$ to $\infty$ do
4:     Choose a random document $w_t$
5:     Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
6:     repeat
7:         Set $\phi_{t,n} \propto \exp\{\mathbb{E}_q[\log \theta_t] + \mathbb{E}_q[\log \beta_{.,w_n}]\}$
8:         Set $\gamma_t = \alpha + \sum_n \phi_{t,n}$
9:     until $\frac{1}{K} \sum_k |\text{change in } \gamma_{t,k}| < \epsilon$
10:    Compute $\tilde{\lambda}_k = \eta + D \sum_n w_{t,n} \phi_{t,n}$
11:    Set $\lambda_k = (1 - \rho_t) \lambda_k + \rho_t \tilde{\lambda}_k$.
12: end for
Stochastic variational inference in LDA

![Graph showing perplexity against documents seen (log scale) with three lines: Online 98K, Batch 98K, Online 3.3M.](image)

**Documents seen (log scale):**
- Online 98K
- Batch 98K
- Online 3.3M

**Perplexity:**
- 900
- 850
- 800
- 750
- 700
- 650
- 600

**Documents analyzed:**
- 2048 systems, road, made, service
- 4096 systems, health, communication, service
- 8192 service, systems, health, companies, market
- 12288 service, systems, companies, business, company
- 16384 service, companies, systems, business, company
- 32768 business, service, companies, industry, company
- 49152 business, service, companies, industry, management
- 65536 business, service, companies, industry, management

**Top eight words:**
- Systems, road, made, service
- Systems, health, communication, service
- Service, systems, health, companies, market
- Service, systems, companies, business, company
- Service, companies, systems, business, company
- Business, service, companies, industry, company
- Business, service, companies, industry, management
- Business, service, companies, industry, management
We defined a generic algorithm for scalable variational inference.

- Bayesian mixture models
- Time series models (variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization (e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression (linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models (LDA and some variants)
• See Hoffman et al. (2010) for LDA (and code).
• See Wang et al. (2010) for Bayesian nonparametric models (and code).
• See Sato (2001) for the original stochastic variational inference.
• See Honkela et al. (2010) for natural gradients and variational inference.
Many applications posit a model, condition on data, and use the posterior.

We can now apply this kind of data analysis to very large data sets.
• The class of conditionally conjugate models is very flexible.

• However, some models—like the CTM and DTM—do not fit in.

• In the past, researchers developed tailored optimization procedures for fitting the variational objective.

• We recently developed a more general approach that subsumes many of these strategies.
Bishop (2006) showed that the optimal mean-field variational distribution is

\[ q^*(z) \propto \exp \left\{ E_{q(\beta)} \left[ \log p(z | \beta, x) \right] \right\} \]
\[ q^*(\beta) \propto \exp \left\{ E_{q(z)} \left[ \log p(\beta | z, x) \right] \right\} \]

- In conjugate models, we can compute these expectations. This determines the form of the optimal variational distribution.
- In nonconjugate models we can’t compute the expectations.
- But, under certain conditions, we can use Taylor approximations. This leads to Gaussian variational distributions.
Using and Checking Topic Models
• We have collected data, selected a model, and inferred the posterior.

• How do we use the topic model?
Using a model means doing something with the posterior inference.

- E.g., visualization, prediction, assessing document similarity, using the representation in a downstream task (like IR)
Questions we ask when evaluating a model:
  - Does my model work? Is it better than another model?
  - Which topic model should I choose? Should I make a new one?
- These questions are tied up in the application at hand.
- Sometimes evaluation is straightforward, especially in prediction tasks.
But a promise of topic models is that they give good exploratory tools. Evaluation is complicated, e.g., is this a good navigator of my collection?

And this leads to more questions:

- How do I interpret a topic model?
- What quantities help me understand what it says about the data?
Using and checking topic models

- How to interpret and evaluate topic models is an active area of research.
  - Visualizing topic models
  - Naming topics
  - Matching topic models to human judgements
  - Matching topic models to external ontologies
  - Computing held out likelihoods in different ways

- I will discuss two components:
  - **Predictive scores** for evaluating topic models
  - **Posterior predictive checks** for topic modeling
### The predictive score

- Assess how well a model can predict **future data**

- In text, a natural setting is one where we observe part of a new document and want to predict the remainder.

- The **predictive distribution** is a distribution conditioned on the corpus and the partial document,

\[
p(w \mid D, w_{\text{obs}}) = \int_\beta \int_\theta \left( \sum_{k=1}^{K} \theta_k \beta_{k,w} \right) p(\theta \mid w_{\text{obs}}, \beta) p(\beta \mid D) \\
\approx \int_\beta \int_\theta \left( \sum_{k=1}^{K} \theta_k \beta_{k,w} \right) q(\theta) q(\beta) \\
= E_q[\theta \mid w_{\text{obs}}]^T E_q[\beta_{,w} \mid D].
\]
The predictive score

- The **predictive score** evaluates the remainder of the document independently under this distribution.

\[
s = \sum_{w \in \mathcal{W}_{\text{held out}}} \log p(w | \mathcal{D}, \mathbf{w}_{\text{obs}})
\]  

- In the predictive distribution, \( q \) is any approximate posterior. This puts various models and inference procedures on the same scale.

- (In contrast, perplexity of entire held out documents requires different approximations for each inference method.)
## The predictive score

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA 100</td>
<td>-7.26</td>
<td>-7.66</td>
<td>-7.41</td>
</tr>
<tr>
<td>LDA 200</td>
<td>-7.50</td>
<td>-7.78</td>
<td>-7.64</td>
</tr>
<tr>
<td>LDA 300</td>
<td>-7.86</td>
<td>-7.98</td>
<td>-7.74</td>
</tr>
<tr>
<td>HDP</td>
<td>-6.97</td>
<td>-7.38</td>
<td>-7.07</td>
</tr>
</tbody>
</table>

The predictive score on large corpora using stochastic variational inference
Posterior predictive checks

- The predictive score and other model selection criteria are good for choosing among several models.

- But they don’t help with the model building process; they don’t tell us how a model is misfit. (E.g. should I go from LDA to a DTM or LDA to a CTM?)

- Further, prediction is not always important in exploratory or descriptive tasks. We may want models that capture other aspects of the data.

- **Posterior predictive checks** are a technique from Bayesian statistics that help with these issues.
This feels even more relevant today. I think of modeling as piecing together various modules, rather than choosing among a population of models.

Machine learning has given us many new building blocks, but has little to say about how to diagnose models.

This is especially important in exploratory analysis, e.g., to form hypotheses or organize data. Many exploratory tasks do not have clear measures of quality.

• Automating model building is a tall order. Even BNP methods do not automate it.

– They help define flexible models, but it is up to the modeler to define likelihood functions, dependencies between the observed data and latent variables, etc.

The predictive check

• Box (1980) describes a predictive check, which tells the story. (Though this story will be refined in a posterior predictive check.)

• All the intuitions about how to assess a model are in this picture:

• The set up from Box (1980) is the following.

– The data are $y$; the hidden variables are $\theta$; the model is $M$.

– Each point of the hidden variable $\theta$ yields a distribution of data.

– The joint distribution combines the prior and the likelihood

$$p(y, \theta | M) = p(y|\theta)p(\theta | M)$$

This is a predictive check from Box (1980).
Posterior predictive checks

- Three stages to model building: estimation, criticism, and revision.

- In **criticism**, the model “confronts” our data.

- Suppose we observe a data set $y$. The predictive distribution is the distribution of data *if the model is true*:

  $$ p(y|\mathcal{M}) = \int_{\theta} p(y|\theta)p(\theta) $$

- Locating $y$ in the predictive distribution indicates if we can “trust” the model.

- Or, locating a **discrepancy function** $g(y)$ in its predictive distribution indicates if what is important to us is captured in the model.
Rubin (1984) located the data $y$ in the posterior $p(y | y, M)$.

Gelman, Meng, Stern (1996) expanded this idea to “realized discrepancies” that include hidden variables $g(y, z)$.

We might make modeling decisions based on a variety of simplifying considerations (e.g., algorithmic). But we can design the realized discrepancy function to capture what we really care about.

Further, realized discrepancies let us consider which parts of the model fit well and which parts don’t. This is apt in exploratory tasks.
Consider a decomposition of a corpus into topics, i.e., \( \{w_{d,n}, z_{d,n}\} \). Note that \( z_{d,n} \) is a latent variable.

For all the observations assigned to a topic, consider the variable \( \{w_{d,n}, d\} \). This is the observed word and the document it appeared in.

One measure of how well a topic model fits the LDA assumptions is to look at the **per-topic mutual information** between \( w \) and \( d \).

If the words from the topic are independently generated then we expect lower mutual information.

What is “low”? To answer that, we can shuffle the words and recompute. This gives values of the MI when the words are independent.
### Posterior predictive checks in topic models

<table>
<thead>
<tr>
<th>Tax</th>
<th>Income</th>
<th>Labor</th>
<th>Women</th>
<th>Contract</th>
</tr>
</thead>
<tbody>
<tr>
<td>income</td>
<td>taxation</td>
<td>workers</td>
<td>sexual</td>
<td>liability</td>
</tr>
<tr>
<td>taxes</td>
<td>revenue</td>
<td>employees</td>
<td>men</td>
<td>parties</td>
</tr>
<tr>
<td>estate</td>
<td>subsidies</td>
<td>employment</td>
<td>sex</td>
<td>contracts</td>
</tr>
<tr>
<td>exemption</td>
<td>organizations</td>
<td>work</td>
<td>child</td>
<td>party</td>
</tr>
<tr>
<td>year</td>
<td>taxpayers</td>
<td>employee</td>
<td>family</td>
<td>creditors</td>
</tr>
<tr>
<td>treasury</td>
<td>consumption</td>
<td>job</td>
<td>children</td>
<td>agreement</td>
</tr>
<tr>
<td></td>
<td>earnings</td>
<td>bargaining</td>
<td>gender</td>
<td>breach</td>
</tr>
<tr>
<td></td>
<td>funds</td>
<td>unions</td>
<td>woman</td>
<td>contractual</td>
</tr>
<tr>
<td></td>
<td></td>
<td>worker</td>
<td>marriage</td>
<td>terms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>collective</td>
<td>discrimination</td>
<td>bargaining</td>
</tr>
<tr>
<td></td>
<td></td>
<td>industrial</td>
<td>male</td>
<td>contracting</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>social</td>
<td>debt</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>female</td>
<td>exchange</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>parents</td>
<td>limited</td>
</tr>
</tbody>
</table>

- This realized discrepancy measures model fitness.
- Can use it to measure model fitness per topic.
- Helps us explore parts of the model that fit well.
Discussion
Probabilistic topic models

- What are topic models?
- What kinds of things can they do?
- How do I compute with a topic model?
- How do I evaluate and check a topic model?
- What are some unanswered questions in this field?
- How can I learn more?
Introduction to topic modeling

- LDA assumes that there are $K$ topics shared by the collection.
- Each document exhibits the topics with different proportions.
- Each word is drawn from one topic.
- We discover the structure that best explain a corpus.

**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 by 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, 880 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


**Topics**

- gene 0.04
- dna 0.02
- genetic 0.01
-...
- life 0.02
- evolve 0.01
- organism 0.01
-...
- brain 0.04
- neuron 0.02
- nerve 0.01
-...
- data 0.02
- number 0.02
- computer 0.01
-...

**Documents**

- are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Jim Anderson at the University in Seattle, who arrived at the 880 number. But coming up with a concrete answer may be more than just a technical numbers game, particularly as more and more genomes are completely sequenced. “It may be a way of organizing any newly sequenced genome,” explains Araceli Mushakjian, 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.

**Subject proportions and assignments**

- LDA assumes that there are $K$ topics shared by the collection.
- Each document exhibits the topics with different proportions.
- Each word is drawn from one topic.
- We discover the structure that best explain a corpus.
Extensions of LDA

Topic models can be adapted to many settings

- relax assumptions
- combine models
- model more complex data
Posterior inference is the central computational problem.
Stochastic variational inference is a scalable algorithm.
We can handle nonconjugacy with Laplace inference.
(Note: There are many types of inference we didn’t discuss.)
### Posterior predictive checks

<table>
<thead>
<tr>
<th>4</th>
<th>10</th>
<th>3</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>tax</td>
<td>labor</td>
<td>women</td>
<td>contract</td>
</tr>
<tr>
<td>income</td>
<td>workers</td>
<td>sexual</td>
<td>liability</td>
</tr>
<tr>
<td>taxation</td>
<td>employees</td>
<td>men</td>
<td>parties</td>
</tr>
<tr>
<td>taxes</td>
<td>union</td>
<td>sex</td>
<td>contracts</td>
</tr>
<tr>
<td>revenue</td>
<td>employer</td>
<td>child</td>
<td>party</td>
</tr>
<tr>
<td>estate</td>
<td>employers</td>
<td>family</td>
<td>creditors</td>
</tr>
<tr>
<td>subsidies</td>
<td>employment</td>
<td>children</td>
<td>agreement</td>
</tr>
<tr>
<td>exemption</td>
<td>work</td>
<td>gender</td>
<td>breach</td>
</tr>
<tr>
<td>organizations</td>
<td>employee</td>
<td>woman</td>
<td>contractual</td>
</tr>
<tr>
<td>year</td>
<td>job</td>
<td>marriage</td>
<td>terms</td>
</tr>
<tr>
<td>treasury</td>
<td>bargaining</td>
<td>discrimination</td>
<td>contractual</td>
</tr>
<tr>
<td>consumption</td>
<td>unions</td>
<td>male</td>
<td>debt</td>
</tr>
<tr>
<td>taxpayers</td>
<td>worker</td>
<td>social</td>
<td>exchange</td>
</tr>
<tr>
<td>earnings</td>
<td>collective</td>
<td>female</td>
<td>limited</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6</th>
<th>15</th>
<th>1</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>jury</td>
<td>speech</td>
<td>firms</td>
<td>constitutional</td>
</tr>
<tr>
<td>trial</td>
<td>free</td>
<td>price</td>
<td>political</td>
</tr>
<tr>
<td>crime</td>
<td>amendment</td>
<td>corporate</td>
<td>constitution</td>
</tr>
<tr>
<td>defendant</td>
<td>freedom</td>
<td>firm</td>
<td>government</td>
</tr>
<tr>
<td>defendants</td>
<td>expression</td>
<td>value</td>
<td>amendment</td>
</tr>
<tr>
<td>sentencing</td>
<td>protected</td>
<td>market</td>
<td>justice</td>
</tr>
<tr>
<td>judges</td>
<td>culture</td>
<td>cost</td>
<td>history</td>
</tr>
<tr>
<td>punishment</td>
<td>context</td>
<td>capital</td>
<td>people</td>
</tr>
<tr>
<td>judge</td>
<td>equality</td>
<td>shareholders</td>
<td>legislative</td>
</tr>
<tr>
<td>crimes</td>
<td>values</td>
<td>stock</td>
<td>opinion</td>
</tr>
<tr>
<td>evidence</td>
<td>conduct</td>
<td>insurance</td>
<td>fourteenth</td>
</tr>
<tr>
<td>sentence</td>
<td>ideas</td>
<td>efficient</td>
<td>article</td>
</tr>
<tr>
<td>jurors</td>
<td>information</td>
<td>assets</td>
<td>majority</td>
</tr>
<tr>
<td>offense</td>
<td>protect</td>
<td>offer</td>
<td>citizens</td>
</tr>
<tr>
<td>guilty</td>
<td>content</td>
<td>share</td>
<td>republican</td>
</tr>
</tbody>
</table>
Probabilistic models

Make assumptions

Collect data

Infer the posterior

Check

Predict

Explore
There are many available implementations of topic modeling. Here is an incomplete list—

- **LDA-C** A C implementation of LDA
- **HDP** A C implementation of the HDP ("infinite LDA")
- **Online LDA** A python package for LDA on massive data
- **LDA in R** Package in R for many topic models
- **LingPipe** Java toolkit for NLP and computational linguistics
- **Mallet** Java toolkit for statistical NLP
- **TMVE** A python package to build browsers from topic models

* available at www.cs.princeton.edu/~blei/
Research opportunities in topic modeling

- **New applications of topic modeling**
  What methods should we develop to solve problems in the computational social sciences? The digital humanities? Digital medical records?

- **Interfaces and downstream applications of topic modeling**
  What can I do with an annotated corpus? How can I incorporate latent variables into a user interface? How should I visualize a topic model?

- **Model interpretation and model checking**
  Which model should I choose for which task? What does the model tell me about my corpus?
Research opportunities in topic modeling

- **Incorporating corpus, discourse, or linguistic structure**
  How can our knowledge of language help inform better topic models?

- **Prediction from text**
  What is the best way to link topics to prediction?

- **Theoretical understanding of approximate inference**
  What do we know about variational inference? Can we analyze it from either the statistical or learning perspective? What are the relative advantages of the many inference methods?

- **And many specific problems**
  E.g., sensitivity to the vocabulary, modeling word contagion, modeling complex trends in dynamic models, robust topic modeling, combining graph models with relational models, ...
“We should seek out unfamiliar summaries of observational material, and establish their useful properties... And still more novelty can come from finding, and evading, still deeper lying constraints.”

“Despite all the computations, you could just dance to the rock ’n’ roll station.”