Probabilistic Topic Models

David M. Blei

Department of Computer Science
Princeton University

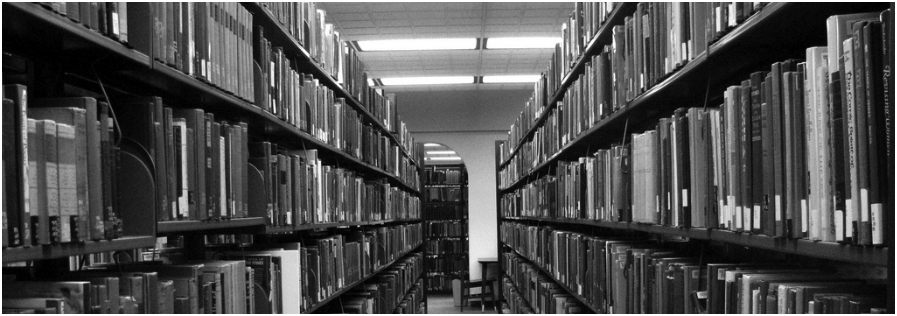
September 26, 2013

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.

Probabilistic topic models



Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

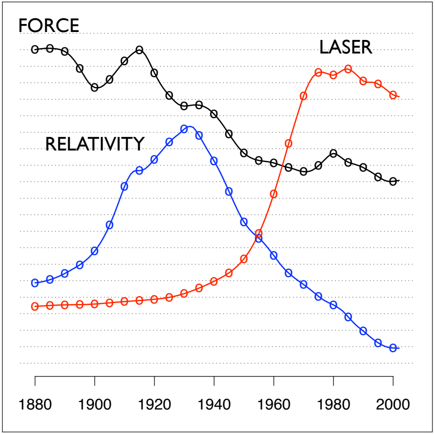
1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use annotations to organize, summarize, search, form predictions.

Probabilistic topic models

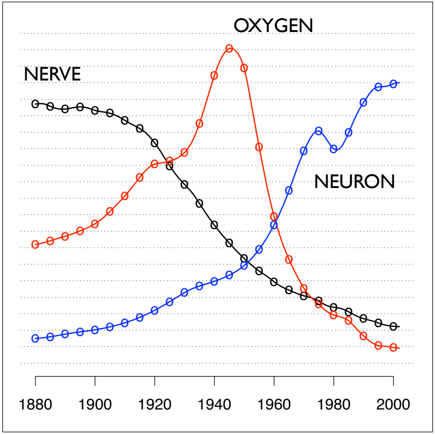
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living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

Probabilistic topic models

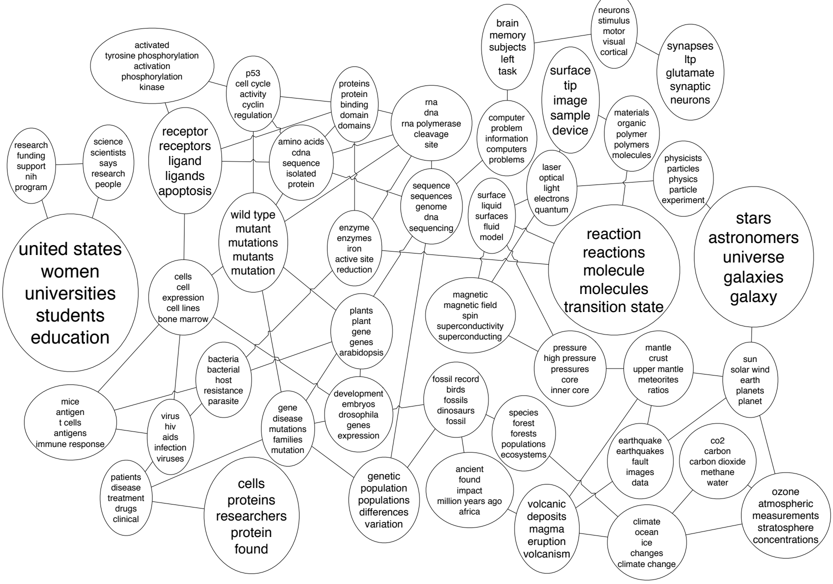
"Theoretical Physics"



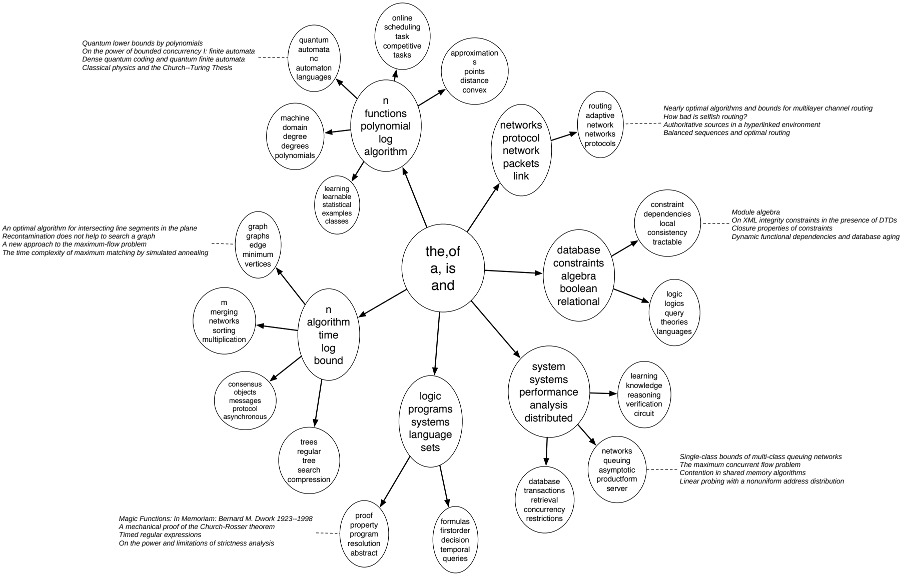
"Neuroscience"



Probabilistic topic models



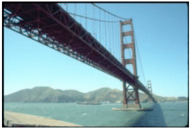
Probabilistic topic models



Probabilistic topic models

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE



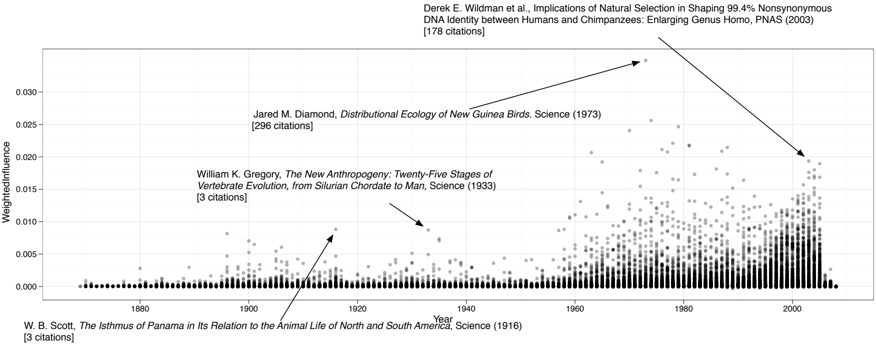
SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

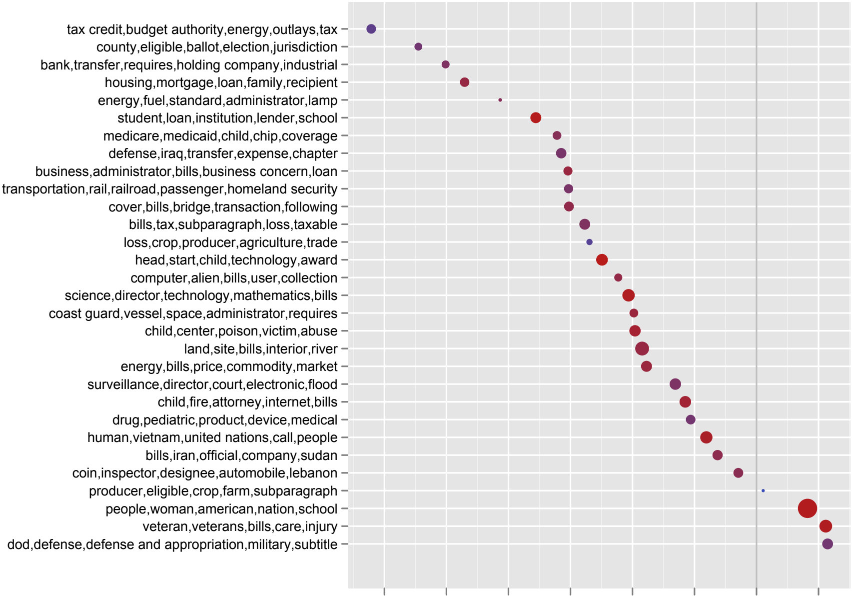
Probabilistic topic models



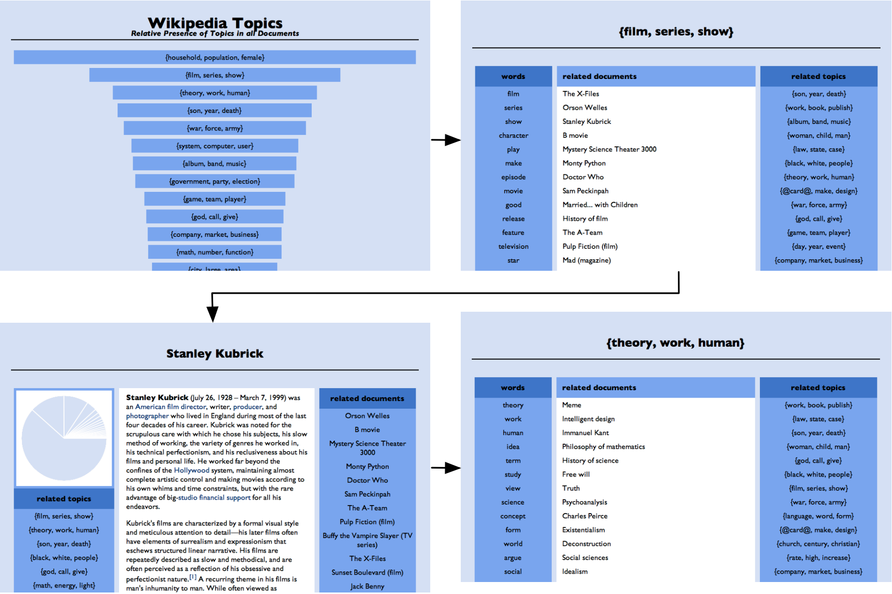
Probabilistic topic models

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

Probabilistic topic models



Probabilistic topic models



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?**

Probabilistic models

- 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:

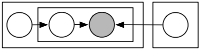
  "[By Bayesian inference,] I simply mean the method of statistical inference that draws conclusions by calculating conditional distributions of unknown quantities given (a) known quantities and (b) model specifications."
  (Rubin, 1984)
- (The Bayesian versus Frequentist debate is not relevant to this 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

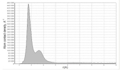
Probabilistic models

Make assumptions

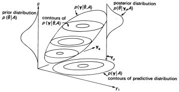


Collect data

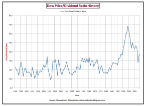
Infer the posterior



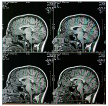
Check



Predict



Explore



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**

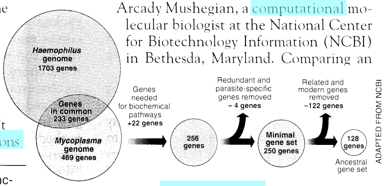
Introduction to Topic Modeling

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

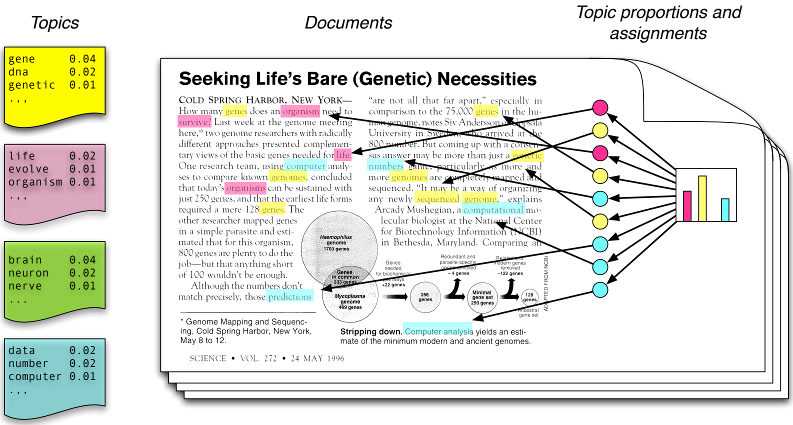


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

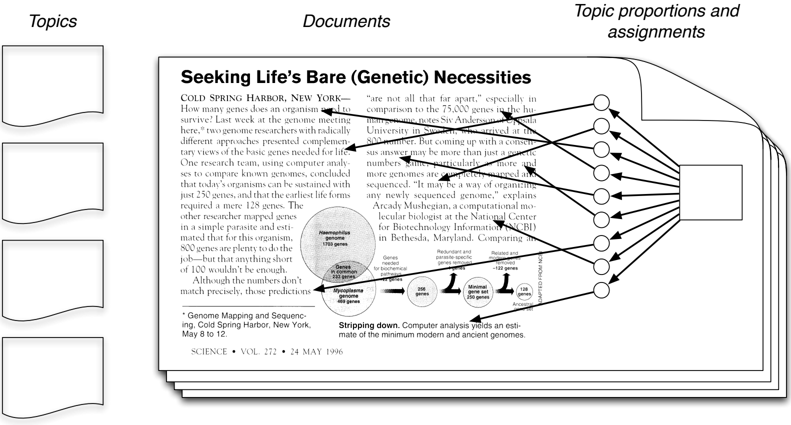
Simple intuition: Documents exhibit multiple topics.

Latent Dirichlet allocation (LDA)



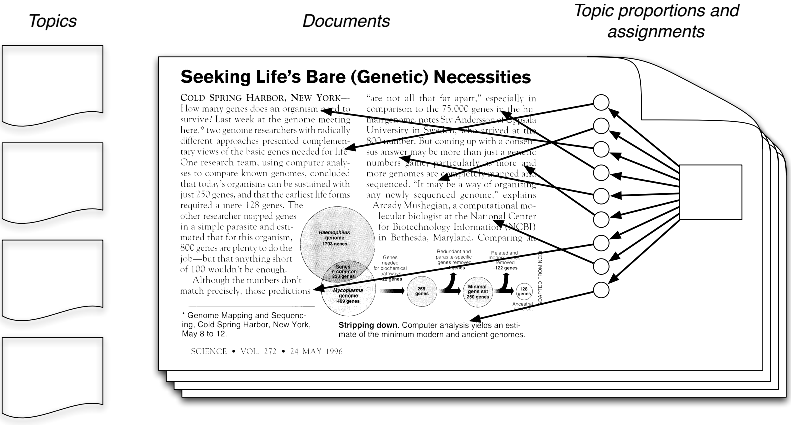
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Latent Dirichlet allocation (LDA)



- In reality, we only observe the documents
- The other structure are **hidden variables**
- Topic modeling algorithms **infer** these variables from data.

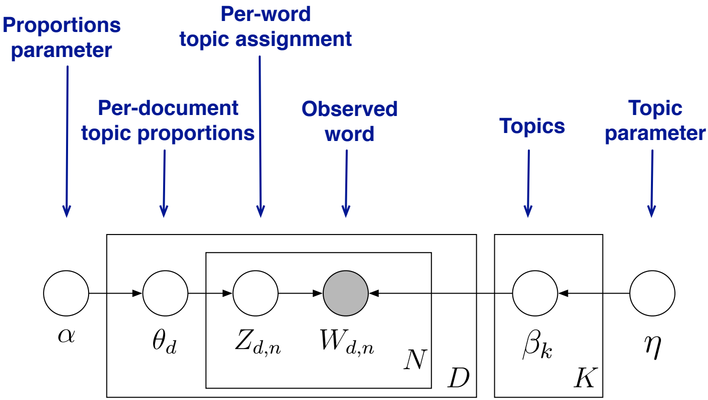
Latent Dirichlet allocation (LDA)



- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

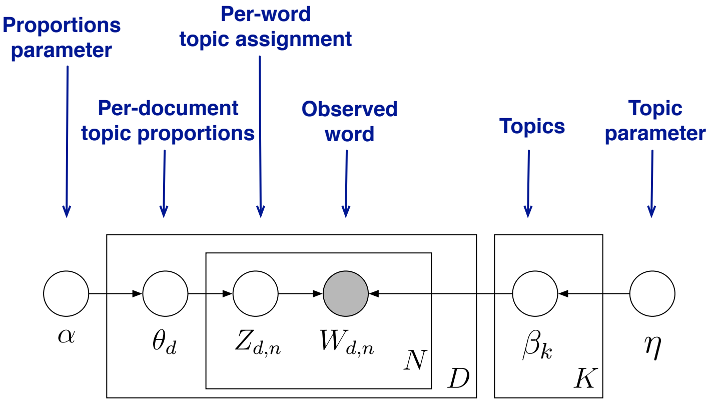
$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

LDA as a graphical model



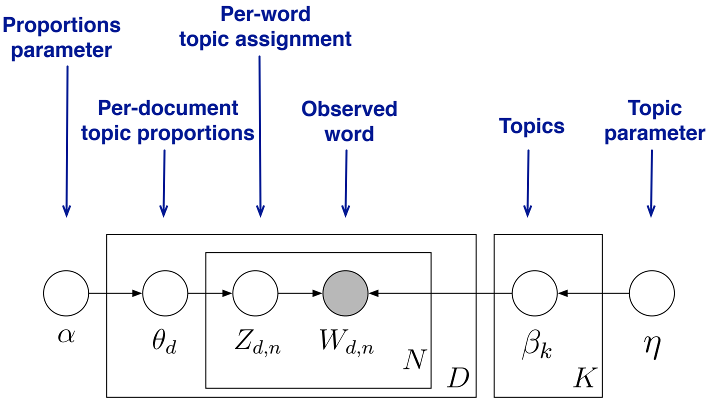
- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

LDA as a graphical model



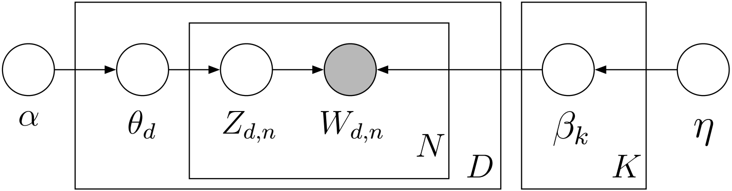
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.

LDA as a graphical model



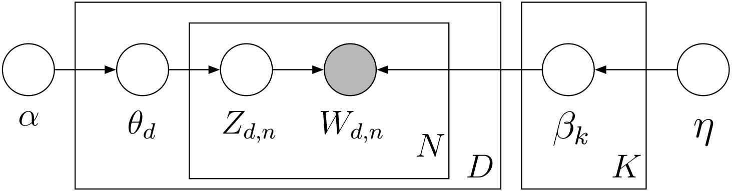
$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i \,|\, \eta) \right) \left(\prod_{d=1}^{D} p(\theta_d \,|\, \alpha) \prod_{n=1}^{N} p(z_{d,n} \,|\, \theta_d) p(w_{d,n} \,|\, \beta_{1:K}, z_{d,n}) \right)$$

LDA as a graphical model



- This joint defines a posterior, $p(\theta, z, \beta \mid w)$.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.

LDA as a graphical model



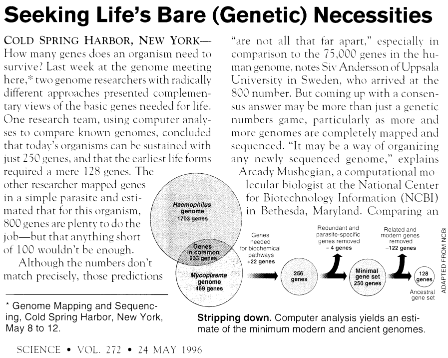
Approximate posterior inference algorithms

- 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

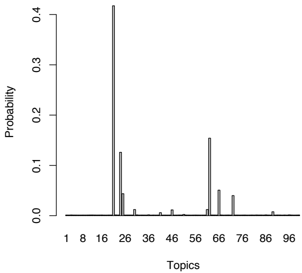
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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 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.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996



Example inference

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

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

6	7	8	9	10
end	time	materials	dna	disease
article	data	surface	rna	cancer
start	two	high	transcription	patients
science	model	structure	protein	human
readers	fig	temperature	site	gene
service	system	molecules	binding	medical
news	number	chemical	sequence	studies
card	different	molecular	proteins	drug
circle	results	fig	specific	normal
letters	rate	university	sequences	drugs

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

16	17	18	19	20
fax	cells	energy	research	neurons
manager	cell	electron	science	brain
science	gene	state	national	cells
aaas	genes	light	scientific	activity
advertising	expression	quantum	scientists	fig
sales	development	physics	new	channels
member	mutant	electrons	states	university
recruitment	mice	high	university	cortex
associate	fig	laser	united	neuronal
washington	biology	magnetic	health	visual

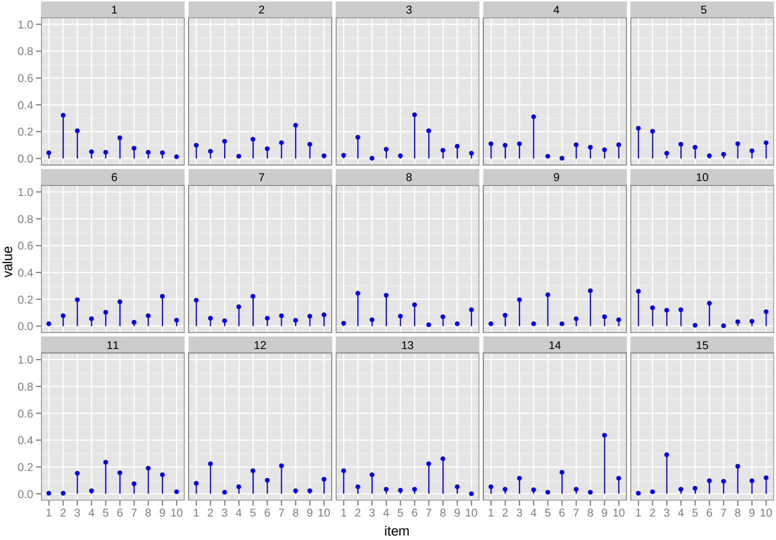
Aside: The Dirichlet distribution

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

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

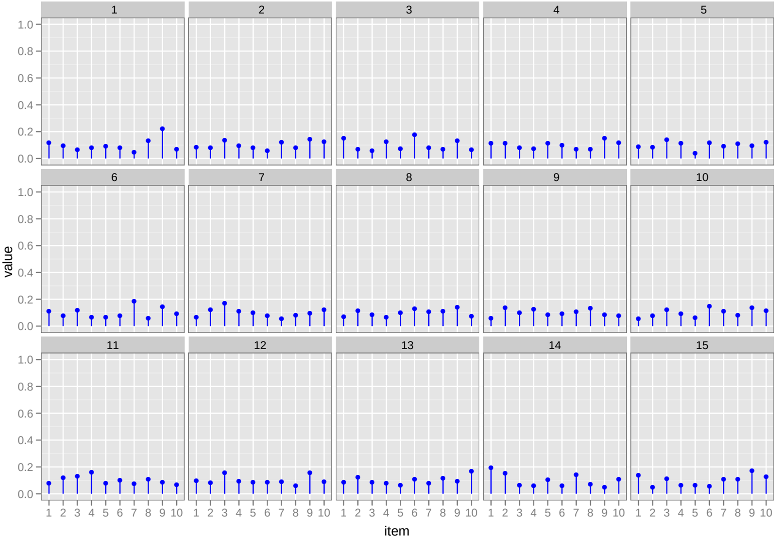
- It is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- The parameter α controls the mean shape and sparsity of θ.
- The topic proportions are a K dimensional Dirichlet.
  The topics are a V dimensional Dirichlet.

$\alpha = 1$



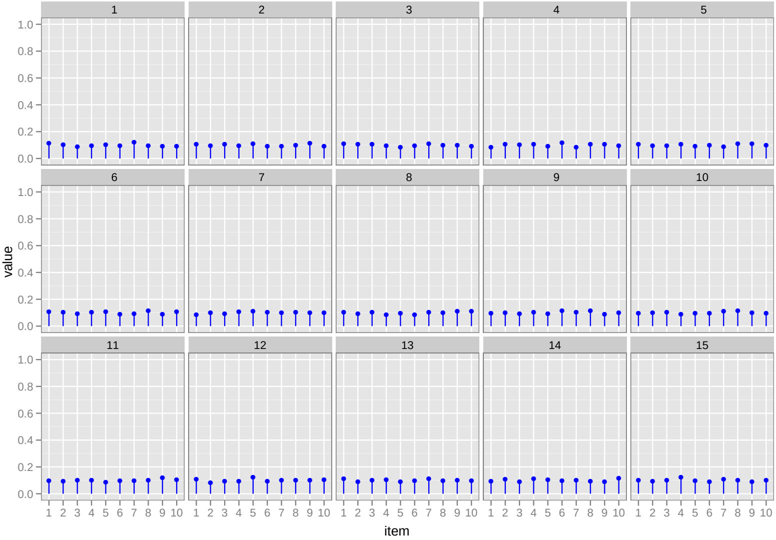
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1 2 3 4 5 6 7 8 9 10

$\alpha = 10$



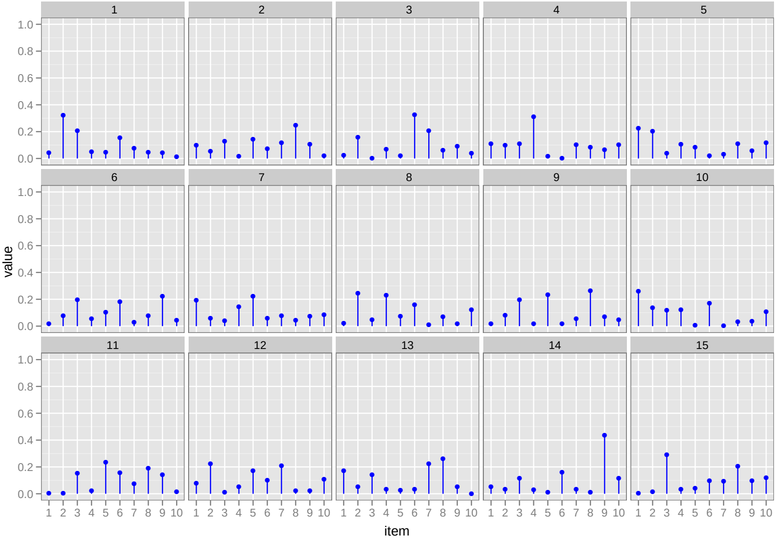
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10
item

$\alpha = 100$

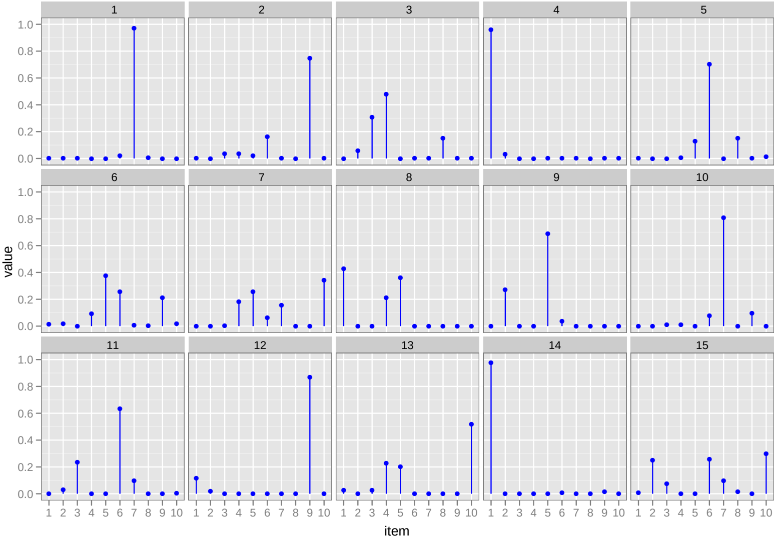


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
item
1 2 3 4 5 6 7 8 9 10

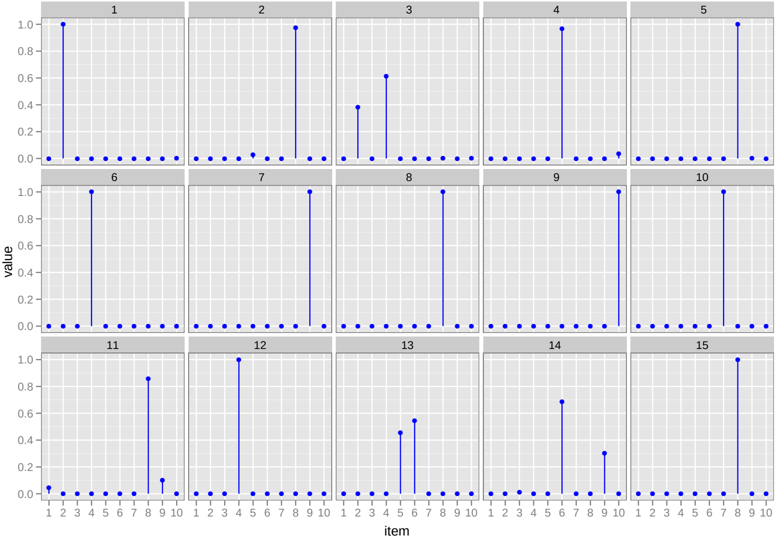
$\alpha = 1$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10
item

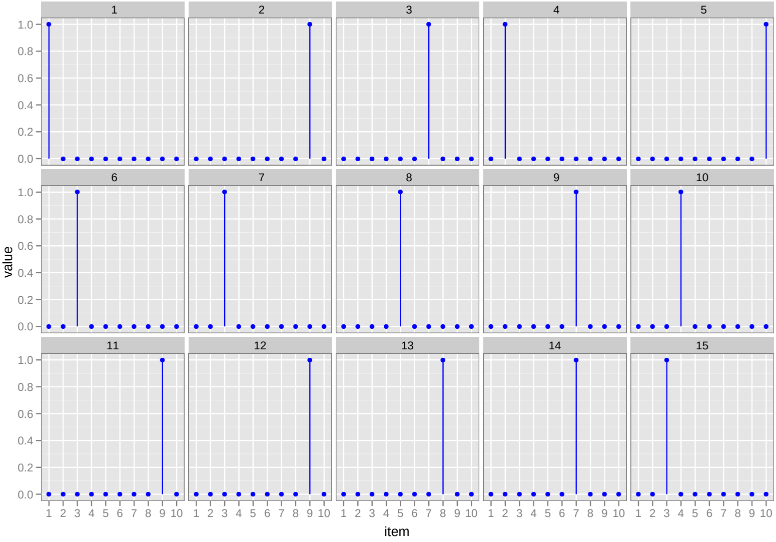
$\alpha = 0.1$



$\alpha = 0.01$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

$\alpha = 0.001$



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1.0
0.8
0.6
0.4
0.2
0.0
value
1 2 3 4 5 6 7 8 9 10
item

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.

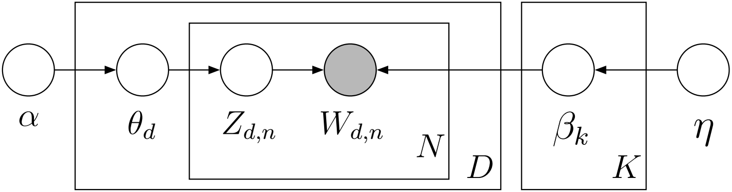
- These goals are at odds.

  - 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.

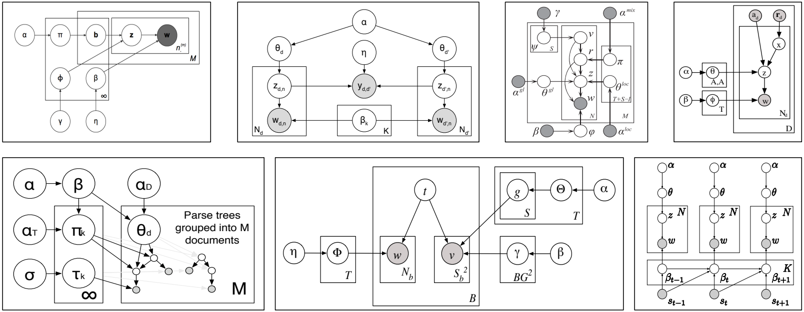
- Trading off these goals finds groups of tightly co-occurring words.

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, humanties, industry, and culture.
- Further, algorithmic improvements let us fit models to massive data.

Example: LDA in R (Jonathan Chang)

perspective identifying tumor suppressor genes in human...
letters global warming report leslie roberts article global....
research news a small revolution gets under way the 1990s....
a continuing series the reign of trial and error draws to a close...
making deep earthquakes in the laboratory lab experimenters...
quick fix for freeways thanks to a team of fast working...
feathers fly in grouse population dispute researchers...
....

```
245 1897:1 1467:1 1351:1 731:2 800:5 682:1 315:6 3668:1 14:1
260 4261:2 518:1 271:6 2734:1 2662:1 2432:1 683:2 1631:7
279 2724:1 107:3 518:1 141:3 3208:1 32:1 2444:1 182:1 250:1
266 2552:1 1993:1 116:1 539:1 1630:1 855:1 1422:1 182:3 2432:1
233 1372:1 1351:1 261:1 501:1 1938:1 32:1 14:1 4067:1 98:2
148 4384:1 1339:1 32:1 4107:1 2300:1 229:1 529:1 521:1 2231:1
193 569:1 3617:1 3781:2 14:1 98:1 3596:1 3037:1 1482:12 665:2
....
```

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

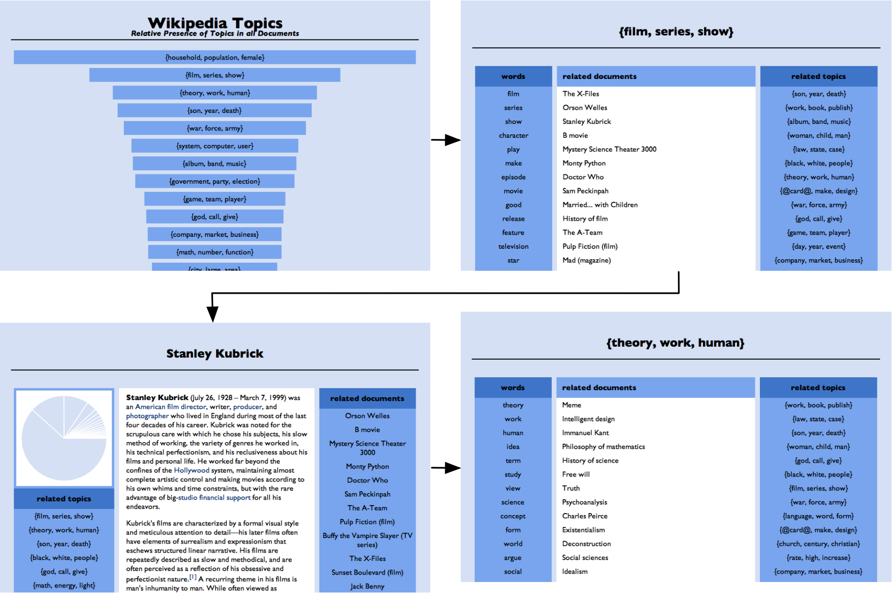
1	2	3	4	5
dna	protein	water	says	mantle
gene	cell	climate	researchers	high
sequence	cells	atmospheric	new	earth
genes	proteins	temperature	university	pressure
sequences	receptor	global	just	seismic
human	fig	surface	science	crust
genome	binding	ocean	like	temperature
genetic	activity	carbon	work	earths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

6	7	8	9	10
end	time	materials	dna	disease
article	data	surface	rna	cancer
start	two	high	transcription	patients
science	model	structure	protein	human
readers	fig	temperature	site	gene
service	system	molecules	binding	medical
news	number	chemical	sequence	studies
card	different	molecular	proteins	drug
circle	results	fig	specific	normal
letters	rate	university	sequences	drugs

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

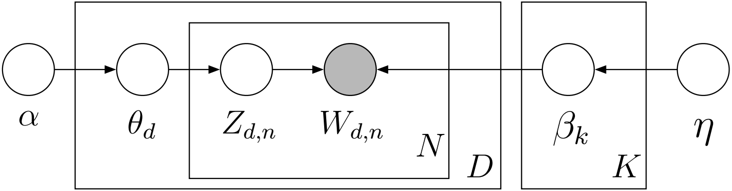
16	17	18	19	20
fax	cells	energy	research	neurons
manager	cell	electron	science	brain
science	gene	state	national	cells
aaas	genes	light	scientific	activity
advertising	expression	quantum	scientists	fig
sales	development	physics	new	channels
member	mutant	electrons	states	university
recruitment	mice	high	university	cortex
associate	fig	laser	united	neuronal
washington	biology	magnetic	health	visual

Open source document browser (with Allison Chaney)



Beyond Latent Dirichlet Allocation

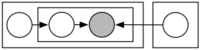
Extending LDA



- 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?

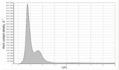
Extending LDA

Make assumptions

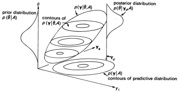


Collect data

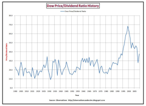
Infer the posterior



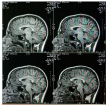
Check



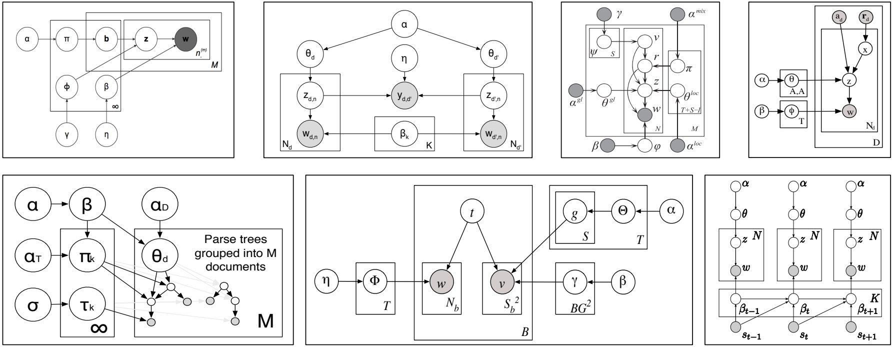
Predict



Explore

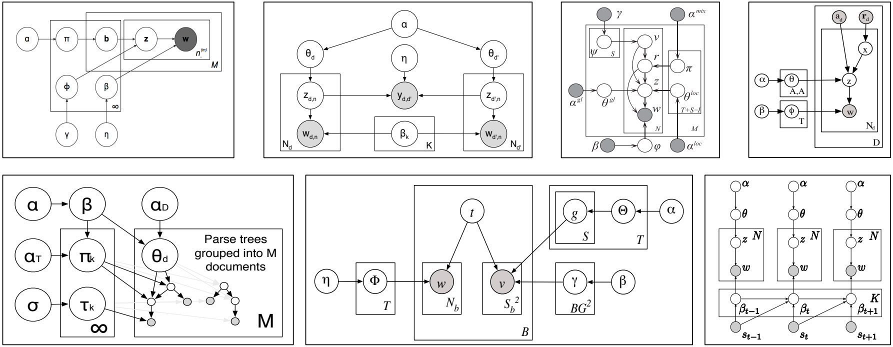


Extending LDA



- 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.

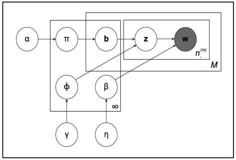
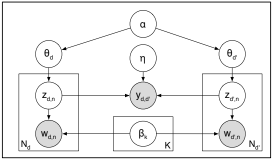
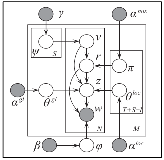
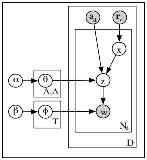
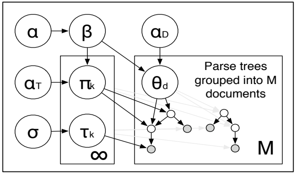
Extending LDA

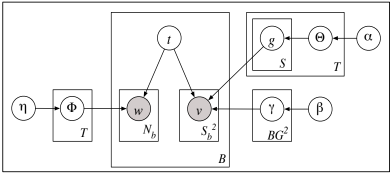
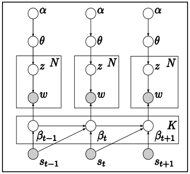


- The **data generating distribution** can be changed. We can apply mixed-membership assumptions to many kinds of data.

- E.g., we can build models of images, social networks, music, purchase histories, computer code, genetic data, and other types.

Extending LDA

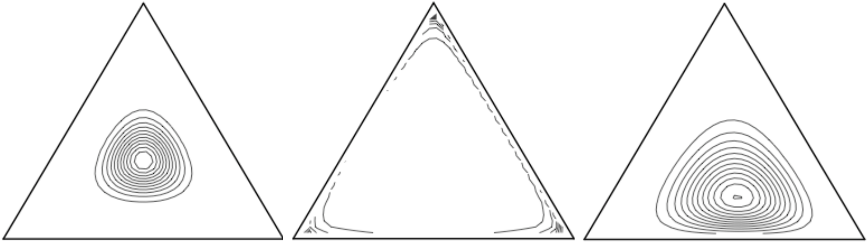
- The **posterior** can be used in creative ways.
- E.g., we can use inferences in information retrieval, recommendation, similarity, visualization, summarization, and other applications.

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**

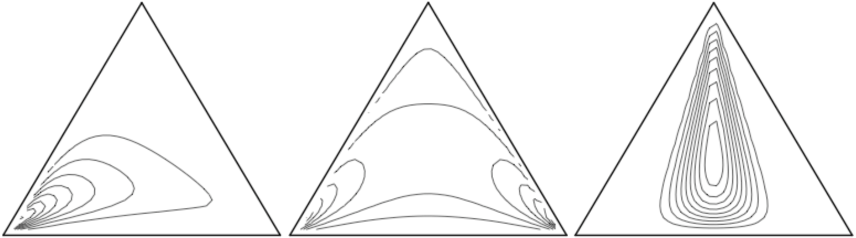
Correlated and Dynamic Topic Models

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*.

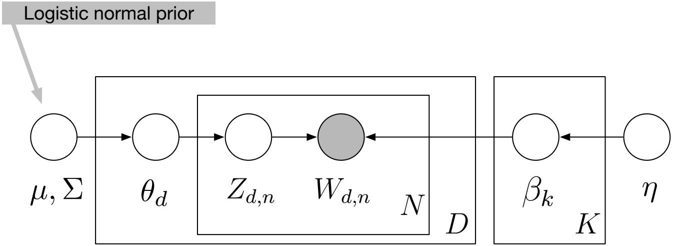
Correlated topic models



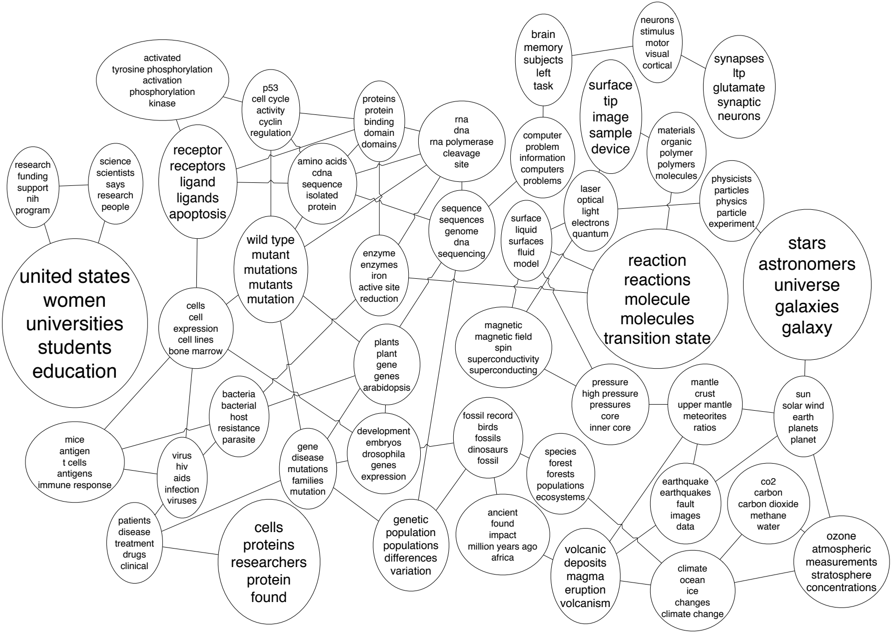
- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

$$
\begin{aligned}
X &\sim \mathcal{N}_K(\mu, \Sigma) \\
\theta_i &\propto \exp\{x_i\}.
\end{aligned}
$$

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



activated tyrosine phosphorylation activation phosphorylation kinase
p53 cell cycle activity cyclin regulation
proteins protein binding domain domains
rna dna rna polymerase cleavage site
brain memory subjects left task
neurons stimulus motor visual cortical
synapses ltp glutamate synaptic neurons
surface tip image sample device
materials organic polymer polymers molecules
computer problem information computers problems
research funding support nih program
science scientists says research people
receptor receptors ligand ligands apoptosis
amino acids cdna sequence isolated protein
sequence sequences genome dna sequencing
laser optical light electrons quantum
physicists particles physics particle experiment
surface liquid surfaces fluid model
wild type mutant mutations mutants mutation
enzyme enzymes iron active site reduction
reaction reactions molecule molecules transition state
stars astronomers universe galaxies galaxy
united states women universities students education
cells cell expression cell lines bone marrow
magnetic magnetic field spin superconductivity superconducting
plants plant gene genes arabidopsis
pressure high pressure pressures core inner core
mantle crust upper mantle meteorites ratios
sun solar wind earth planets planet
bacteria bacterial host resistance parasite
development embryos drosophila genes expression
fossil record birds fossils dinosaurs fossil
species forest forests populations ecosystems
mice antigen t cells antigens immune response
virus hiv aids infection viruses
gene disease mutations families mutation
earthquake earthquakes fault images data
co2 carbon carbon dioxide methane water
patients disease treatment drugs clinical
cells proteins researchers protein found
genetic population populations differences variation
ancient found impact million years ago africa
volcanic deposits magma eruption volcanism
climate ocean ice changes climate change
ozone atmospheric measurements stratosphere concentrations

Dynamic topic models

1789

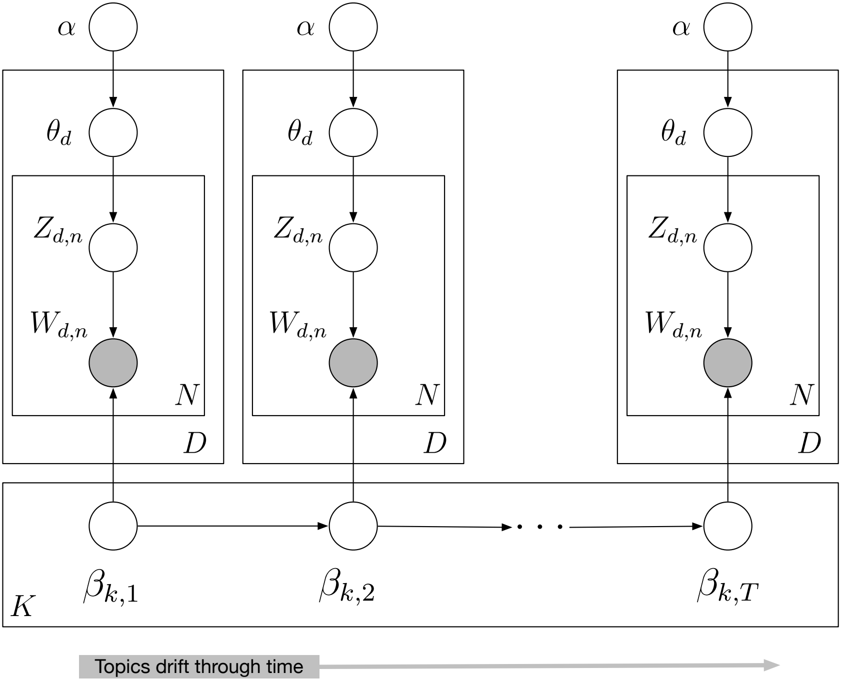
My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

Inaugural addresses

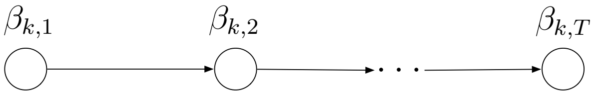
2009

AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

- 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 *drift* in a sequence.


α
θ_d
$Z_{d,n}$
$W_{d,n}$
N
D
$\beta_{k,1}$
$\beta_{k,2}$
$\beta_{k,T}$
K
Topics drift through time

Dynamic topic models



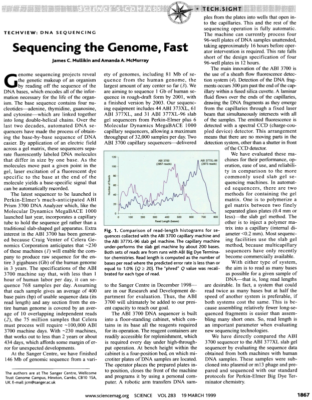
- Use a logistic normal distribution to model topics evolving over time.
- Embed it in a state-space model on the log of the topic distribution

$$\begin{aligned} \beta_{t,k} \mid \beta_{t-1,k} &\sim \mathcal{N}(\beta_{t-1,k}, I\sigma^2) \\ p(w \mid \beta_{t,k}) &\propto \exp\{\beta_{t,k}\} \end{aligned}$$

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

Dynamic topic models

Original article



SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 μm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

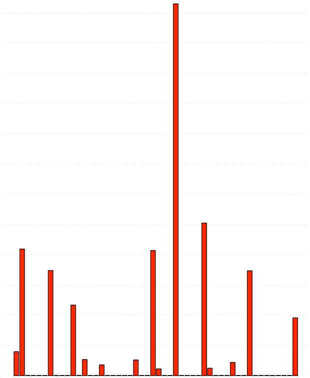
With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

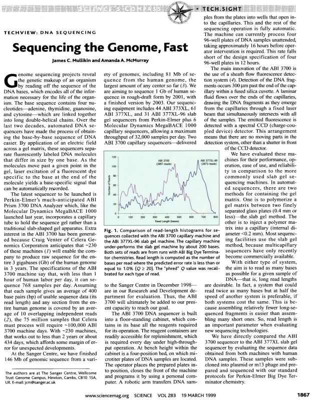
The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

Topic proportions



Dynamic topic models

Original article



SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 μm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

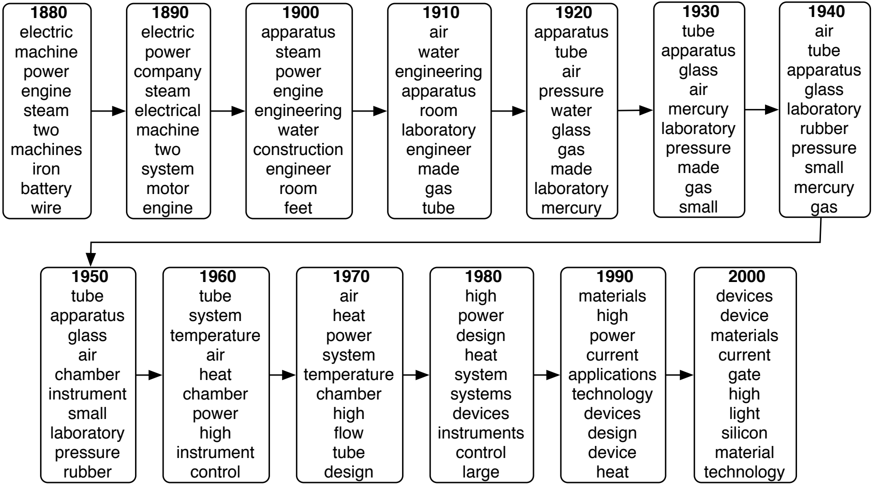
Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

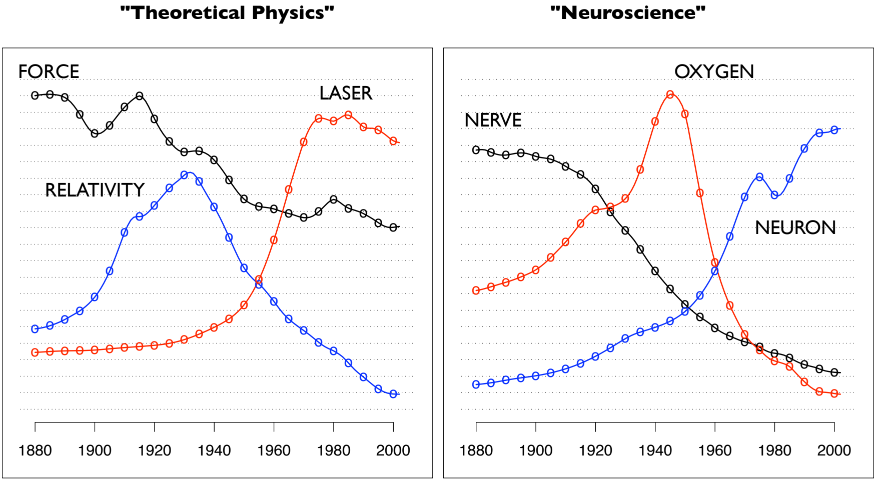
Most likely words from top topics

sequence	devices	data
genome	device	information
genes	materials	network
sequences	current	web
human	high	computer
gene	gate	language
dna	light	networks
sequencing	silicon	time
chromosome	material	software
regions	technology	system
analysis	electrical	words
data	fiber	algorithm
genomic	power	number
number	based	internet

Dynamic topic models



Dynamic topic models



Dynamic topic models

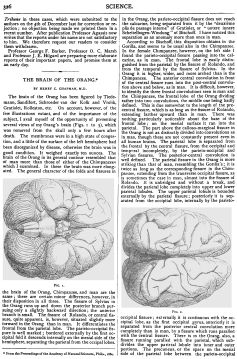
- **Time-corrected similarity** shows a new way of using the posterior.
- Consider the expected Hellinger distance between the topic proportions of two documents,

$$d_{ij} = \mathrm{E}\left[\sum_{k=1}^{K}(\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}})^2 \,|\, \mathbf{w}_i, \mathbf{w}_j\right]$$

- 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

Dynamic topic models

The Brain of the Orang (1880)



Tribune in these cases, which were submitted to the authors on the 4th of December last for correction or rejection; no objection being made we printed them in a recent number. After publication Professor Agassiz now writes that the reports under his name are not satisfactory to him. We therefore request our readers to consider them withdrawn.

Professor George F. Barker, Professor O. C. Marsh and Professor J. E. Hilgard are preparing more elaborate reports of their important papers, and promise them at an early day.

THE BRAIN OF THE ORANG.*

BY HENRY C. CHAPMAN, M.D.

The brain of the Orang has been figured by Tiedemann, Sandifort, Schroeder van der Kolk and Vrolik, Gratiolet, Rolleston, etc. On account, however, of the few illustrations extant, and of the importance of the subject, I avail myself of the opportunity of presenting several views of my Orang's brain (Figs. 1 to 5), which was removed from the skull only a few hours after death. The membranes were in a high state of congestion, and a little of the surface of the left hemisphere had been disorganized by disease, otherwise the brain was in good condition. It weighed exactly ten ounces. The brain of the Orang in its general contour resembled that of man more than those of either of the Chimpanzees which I examined. In these the brain was more elongated. The general character of the folds and fissures in

Fig. 1.

the brain of the Orang, Chimpanzee, and man are the same; there are certain minor differences, however, in their disposition in all three. The fissure of Sylvius in the Orang runs up and down the posterior branch pursuing only a slightly backward direction; the anterior branch is small. The fissure of Rolando, or central fissure, quite apparent, is, however, situated slightly more forward in the Orang than in man. It differentiates the frontal from the parietal lobe. The parieto-occipital fispure is well marked; bordered externally by the first occipital fold it descends internally on the mesial side of the hemisphere, separating the parietal from the occipal lobes.

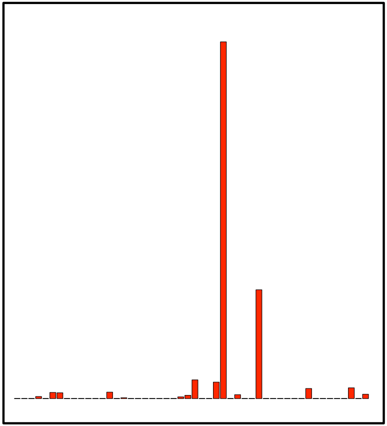
in the Orang, the parieto-occipital fissure does not reach the calcarine, being separated from it by the "deuxième plis de passage interne" of Gratiolet, or "untere innere Scheitelbogen-Windung" of Bischoff. I have noticed this separation as an anomaly more than once in man.

According to Bischoff, this disposition obtains in the Gorilla, and seems to be usual also in the Chimpanzee. In the female Chimpanzee, however, on the left side I found the parieto-occipital fissure passing into the calcarine, as in man. The frontal lobe is easily distinguished from the parietal by the fissure of Rolando, and from the temporal by the fissure of Sylvius. In the Orang it is higher, wider, and more arched than in the Chimpanzee. The anterior central convolution in front of the central fissure runs into the post-central convolution above and below, as in man. It is difficult, however, to identify the three frontal convolutions seen in man and the Chimpanzee, the frontal lobe of the Orang dividing rather into two convolutions, the middle one being badly defined. This is due somewhat to the length of the precentral fissure, which is as long as the fissure of Rolando, extending farther upward than in man. There was nothing particularly noticeable about the base of the frontal lobe; on the mesial surface it ran into the parietal. The part above the calloso-marginal fissure in the Orang is not as distinctly divided into convolutions as in man, though these are not constantly present even in all human brains. The parietal lobe is separated from the frontal by the central fissure, from the occipital and temporal incompletely, by the parieto-occipital and Sylvian fissures. The posterior-central convolution is well defined. The parietal fissure in the Orang is more striking than that of man, resembling the Gorilla's; it is twice as long as the corresponding fissure in the Chimpanzee, extending from the transverse occipital fissure, as is sometimes the case in man, almost into the fissure of Rolando. It is unbridged and without a break, and divides the parietal lobe completely into upper and lower parietal lobules. The upper parietal lobule is bounded externally by the parietal fissure; posteriorly it is separated from the occipital lobe, internally by the parieto-

Fig. 2.

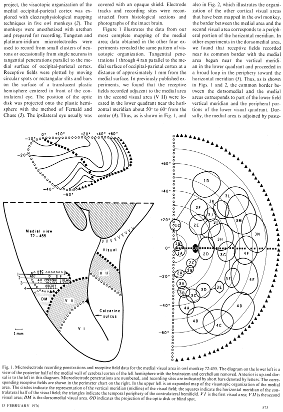
occipital fissure; externally it is continuous with the occipital lobe, as the first occipital gyrus, anteriorly it is separated from the posterior central convolution more completely than in man, by a fissure which runs paralled with the central fissure. There is in the Orang, also, a fissure running paralled with the parietal, which subdivides the upper parietal lobule into inner and outer portions. The precuneus, or the space on the mesial side of the parietal lobe between the parieto-occipital

* From the Proceedings of the Academy of Natural Sciences, Phila., 1880.



Dynamic topic models

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

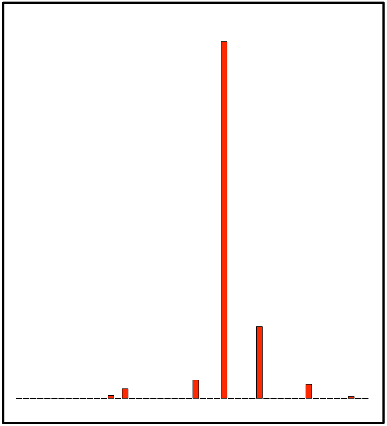


project, the visuotopic organization of the medial occipital-parietal cortex was explored with electrophysiological mapping techniques in five owl monkeys (*2*). The monkeys were anesthetized with urethan and prepared for recording. Tungsten and platinum-iridium microelectrodes were used to record from small clusters of neurons or occasionally from single neurons in tangential penetrations parallel to the medial surface of occipital-parietal cortex. Receptive fields were plotted by moving circular spots or rectangular slits and bars on the surface of a translucent plastic hemisphere centered in front of the contralateral eye. The position of the optic disk was projected onto the plastic hemisphere with the method of Fernald and Chase (*3*). The ipsilateral eye usually was covered with an opaque shield. Electrode tracks and recording sites were reconstructed from histological sections and photographs of the intact brain.

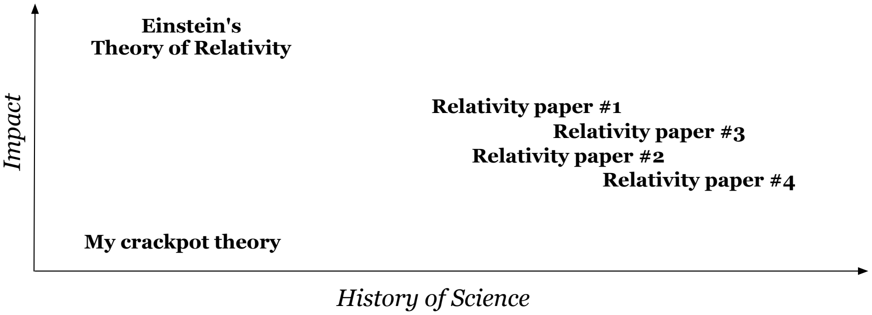
Figure 1 illustrates the data from our most complete mapping of the medial area; data obtained in the other four experiments revealed the same pattern of visuotopic organization. Tangential penetrations 1 through 4 ran parallel to the medial surface of occipital-parietal cortex at a distance of approximately 1 mm from the medial surface. In previously published experiments, we found that the receptive fields recorded adjacent to the medial area in the second visual area (V II) were located in the lower quadrant near the horizontal meridian about 50° to 60° from the center (*4*). Thus, as is shown in Fig. 1, and also in Fig. 2, which illustrates the organization of the other cortical visual areas that have been mapped in the owl monkey, the border between the medial area and the second visual area corresponds to a peripheral portion of the horizontal meridian. In other experiments in the dorsomedial area, we found that receptive fields recorded near its common border with the medial area began near the vertical meridian in the lower quadrant and proceeded in a broad loop in the periphery toward the horizontal meridian (*5*). Thus, as is shown in Figs. 1 and 2, the common border between the dorsomedial and the medial areas corresponds to part of the lower field vertical meridian and the peripheral portions of the lower visual quadrant. Dorsally, the medial area is adjoined by poste-

Fig. 1. Microelectrode recording penetrations and receptive field data for the medial visual area in owl monkey 72-455. The diagram on the lower left is a view of the posterior half of the medial wall of cerebral cortex of the left hemisphere with the brainstem and cerebellum removed. Anterior is up and dorsal is to the left in this diagram. Microelectrode penetrations are numbered, and recording sites are indicated by short bars denoted by letters. The corresponding receptive fields are shown in the perimeter chart on the right. In the upper left is an expanded map of the visuotopic organization of the medial area. The circles indicate the representation of the vertical meridian (midline) of the visual field; the squares indicate the horizontal meridian of the contralateral half of the visual field; the triangles indicate the temporal periphery of the contralateral hemifield. *V I* is the first visual area; *V II* is the second visual area; *DM* is the dorsomedial visual area. *OD* indicates the projection of the optic disk or blind spot.

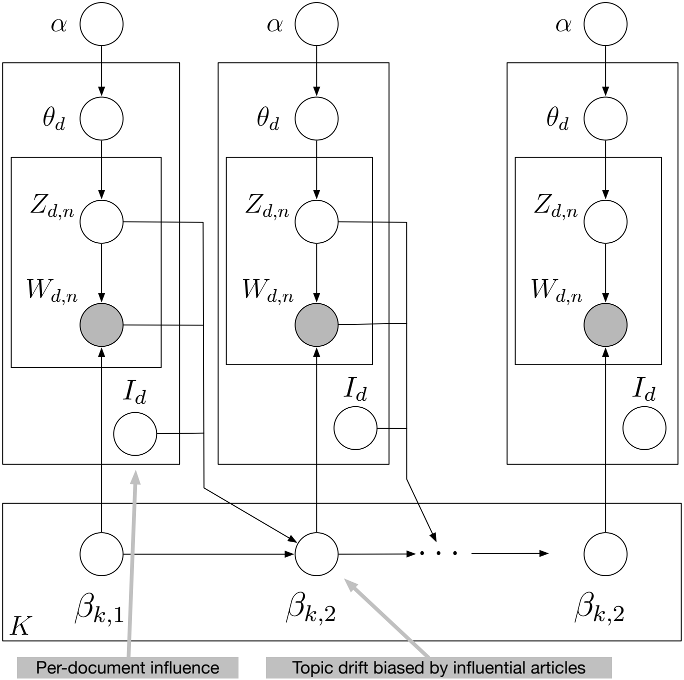
13 FEBRUARY 1976 573



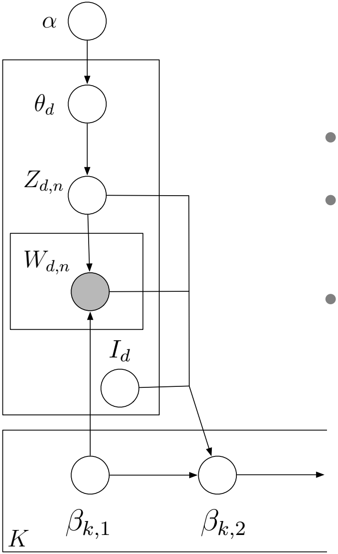
Measuring scholarly impact



- 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.

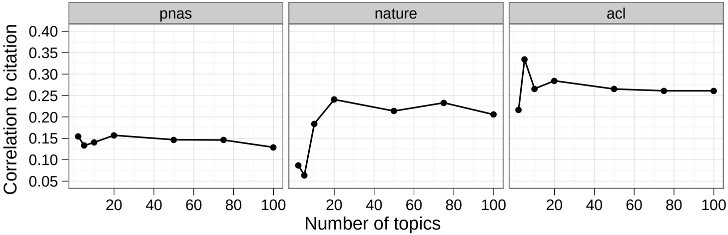

α
θ_d
$Z_{d,n}$
$W_{d,n}$
I_d
$\beta_{k,1}$
$\beta_{k,2}$
$\beta_{k,2}$
K
Per-document influence
Topic drift biased by influential articles

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

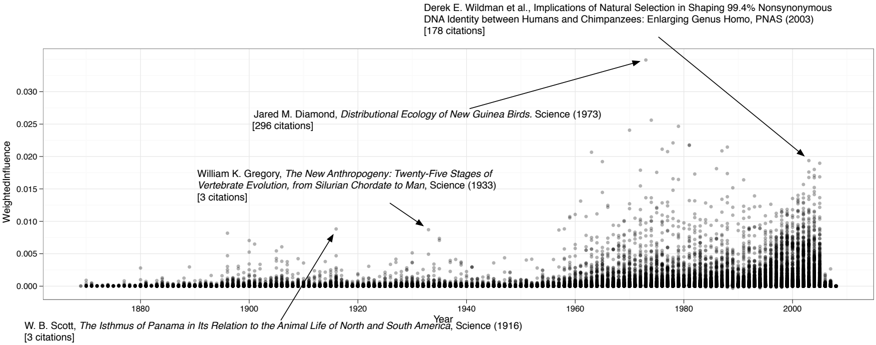


- This measure of impact only uses the words of the documents.
  It correlates strongly with citation counts.

- High impact, high citation: “The Mathematics of Statistical Machine Translation: Parameter Estimation” (Brown et al., 1993)

- “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

Summary: Correlated and dynamic topic models

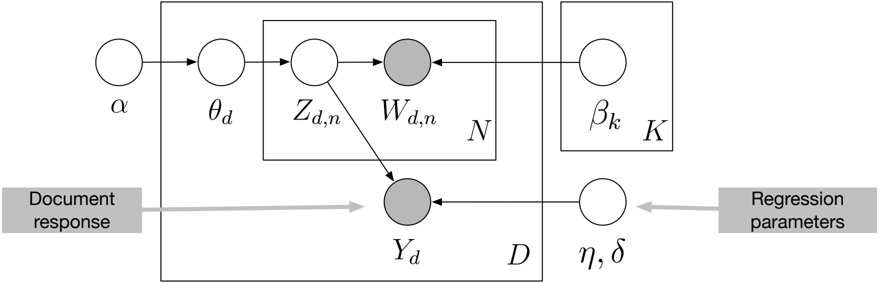
- The Dirichlet assumption on topics and topic proportions makes strong conditional independence assumptions about the data.
- The **correlated topic model** uses a logistic normal on the topic proportions to find patterns in how topics tend to co-occur.
- The **dynamic topic model** uses a logistic 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.

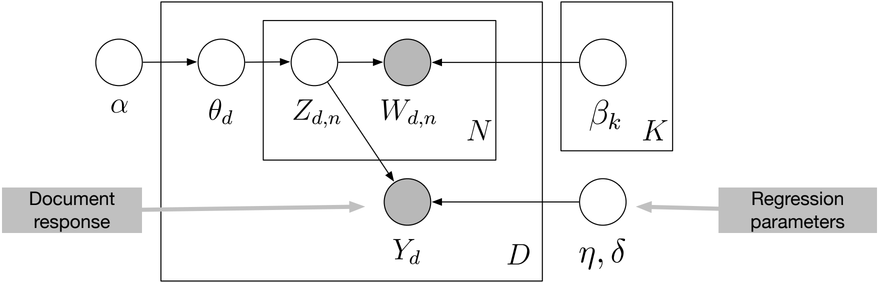
Supervised LDA



1. Draw topic proportions $\theta \,|\, \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \,|\, \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \,|\, z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \,|\, z_{1:N}, \eta, \sigma^2 \sim \mathrm{N}\left(\eta^\top \bar{z}, \sigma^2\right)$, where

$$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

Supervised LDA

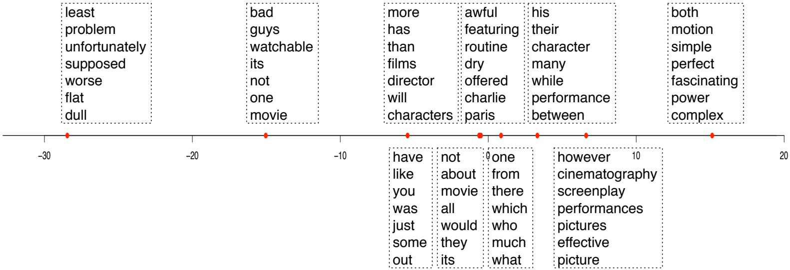


- Fit sLDA parameters to documents and responses.
  This gives: topics $\beta_{1:K}$ and coefficients $\eta_{1:K}$.
- Given a new document, predict its response using the expected value:

$$\mathrm{E}\left[Y \,|\, w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2\right] = \eta^\top \mathrm{E}\left[\bar{Z} \,|\, w_{1:N}\right]$$

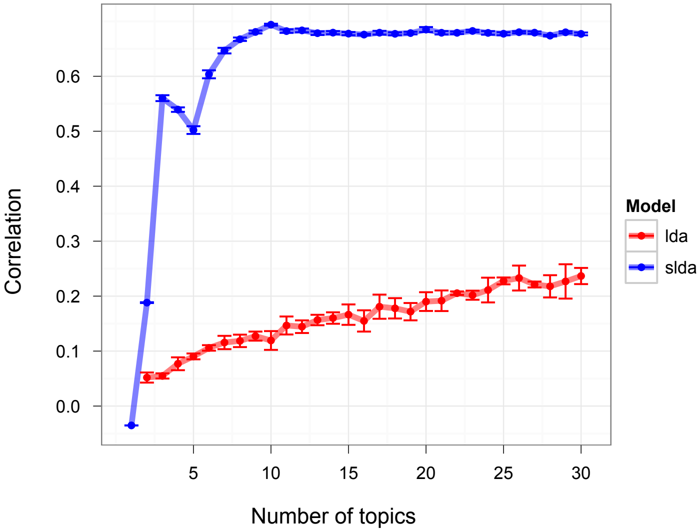
- This blends generative and discriminative modeling.

Supervised LDA

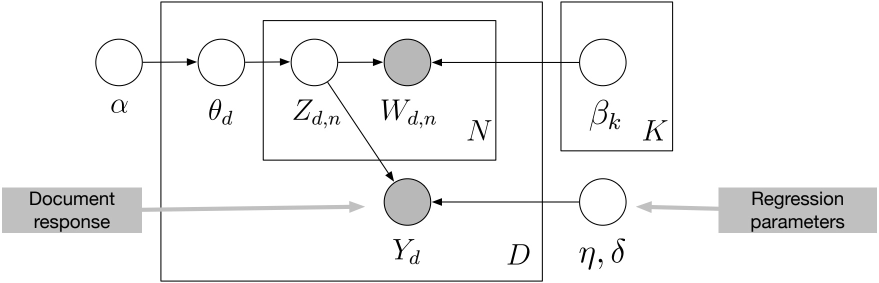


- 10-topic sLDA model on movie reviews (Pang and Lee, 2005).
- Response: number of stars associated with each review
- Each component of coefficient vector η is associated with a topic.

Supervised LDA

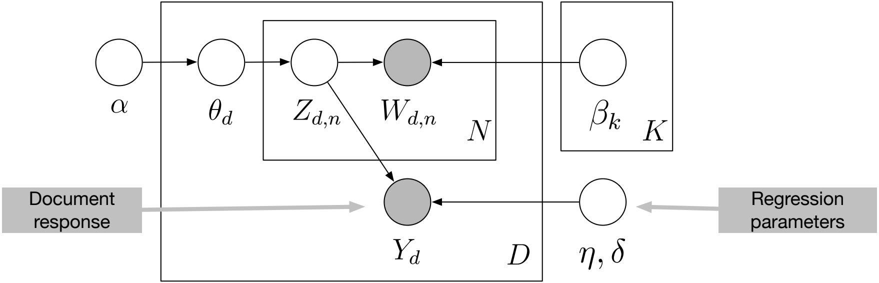


Supervised LDA



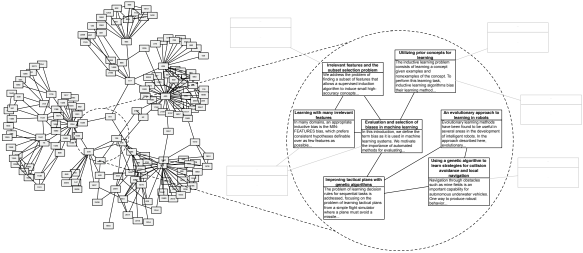
- SLDA enables model-based regression where the predictor is a document.
- It can easily be used wherever LDA is used in an unsupervised fashion (e.g., images, genes, music).
- SLDA is a supervised dimension-reduction technique, whereas LDA performs unsupervised dimension reduction.

Supervised LDA



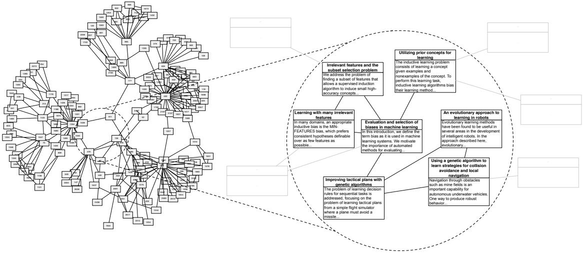
- SLDA has been extended to generalized linear models, e.g., for image classification and other non-continuous responses.
- We will discuss two extensions of sLDA
  - **Relational topic models**: Models of networks and text
  - **Ideal point topic models**: Models of legislative voting behavior

Relational topic models



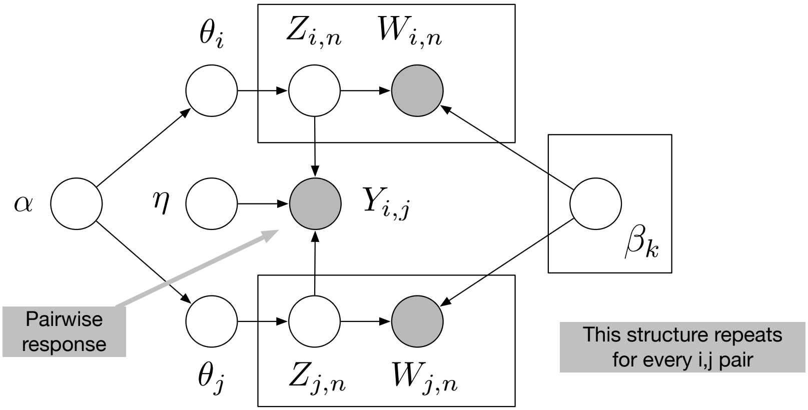
- Many data sets contain **connected observations.**
- For example:
  - Citation networks of documents
  - Hyperlinked networks of web-pages.
  - Friend-connected social network profiles

Relational topic models



- 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.

Relational topic models



- 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

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

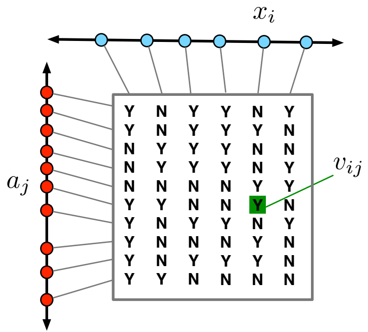
Given a new document, which documents is it likely to link to?

Relational topic models

Competitive environments evolve better solutions for complex tasks	
Coevolving High Level Representations	RTM (ψ_e)
A Survey of Evolutionary Strategies	
Genetic Algorithms in Search, Optimization and Machine Learning	
Strongly typed genetic programming in evolving cooperation strategies	
Solving combinatorial problems using evolutionary algorithms	
A promising genetic algorithm approach to job-shop scheduling...	
Evolutionary Module Acquisition	
An Empirical Investigation of Multi-Parent Recombination Operators...	
A New Algorithm for DNA Sequence Assembly	**LDA + Regression**
Identification of protein coding regions in genomic DNA	
Solving combinatorial problems using evolutionary algorithms	
A promising genetic algorithm approach to job-shop scheduling...	
A genetic algorithm for passive management	
The Performance of a Genetic Algorithm on a Chaotic Objective Function	
Adaptive global optimization with local search	
Mutation rates as adaptations	

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

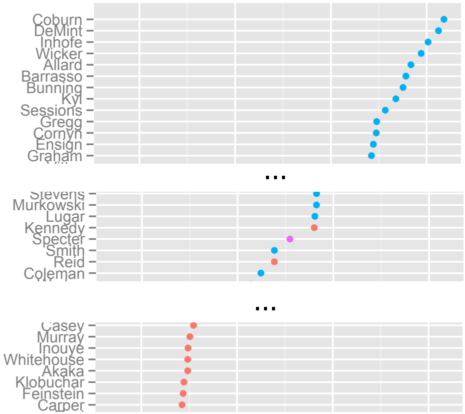
Ideal point topic models



$$p(v_{ij}) = f(d(x_i, a_j))$$

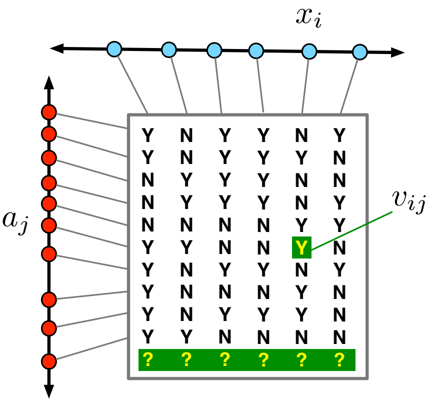
- The **ideal point model** uncovers voting patterns in legislative data
- We observe roll call data v_{ij}.
- Bills attached to discrimination parameters a_j.
  Senators attached to ideal points x_i.

Ideal point topic models



- Posterior inference reveals the political spectrum of senators
- Widely used in quantitative political science.

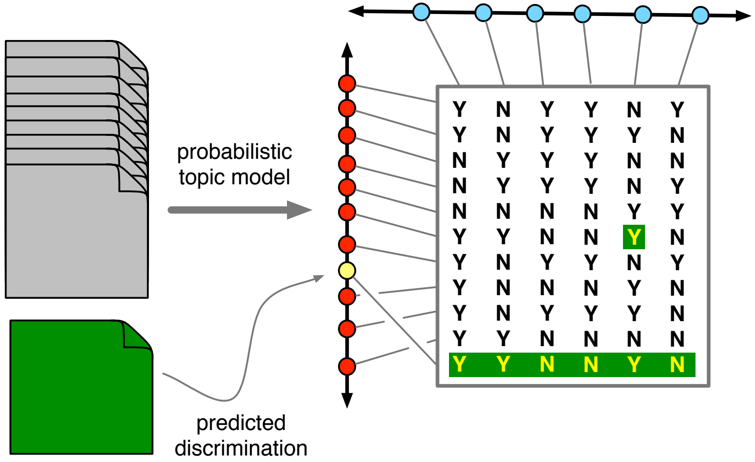
Ideal point topic models



$$p(v_{ij}) = f(d(x_i, a_j))$$

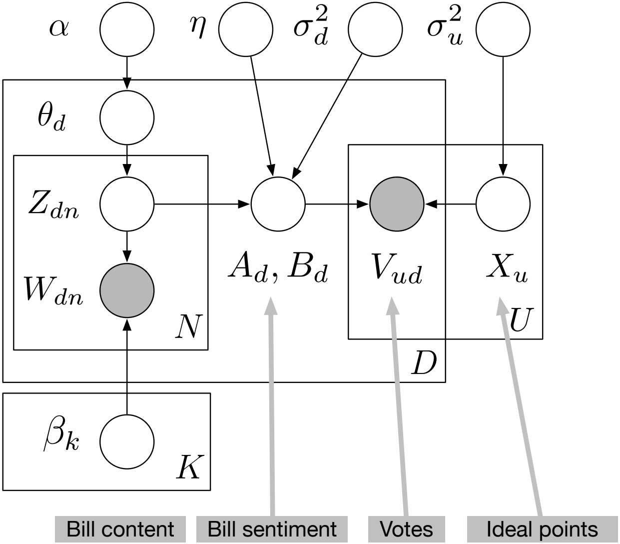
- We can predict a missing vote.
- But we cannot predict all the missing votes from a bill.
- Cf. the limitations of collaborative filtering

Ideal point topic models

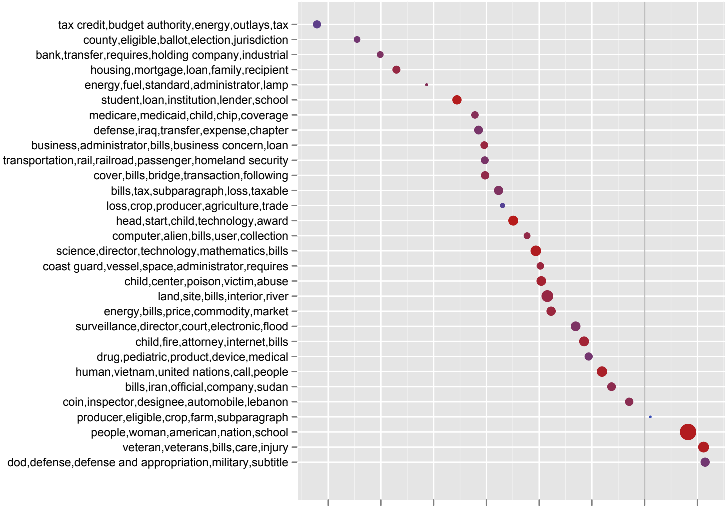


- Use supervised LDA to predict bill discrimination from bill text.
- But this is a **latent response.**

Ideal point topic models



Ideal point topic models



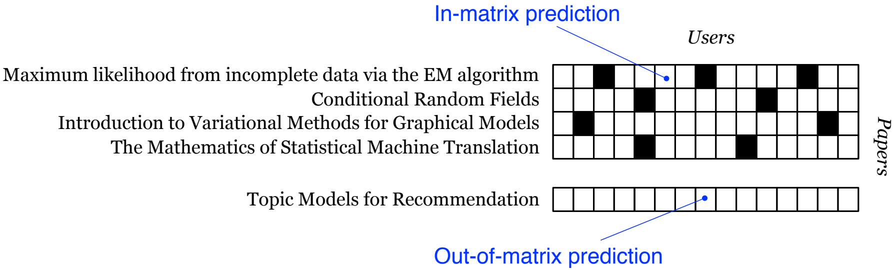
In addition to senators and bills, IPTM places **topics** on the spectrum.

Summary: Supervised topic models

- 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.”)

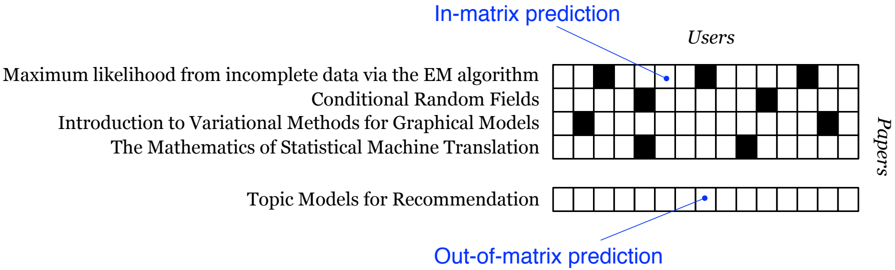
Modeling User Data and Text

Topic models for recommendation (Wang and Blei, 2011)



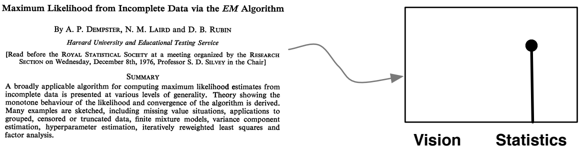
- In many settings, we have information about **how people use documents.**
- With new models, this can be used to
  - Help people find documents that they are interested in
  - Learn about what the documents mean to the people reading them
  - Learn about the people reading (or voting on) the documents.
- (We also saw this in ideal point topic models.)

Topic models for recommendation (Wang and Blei, 2011)

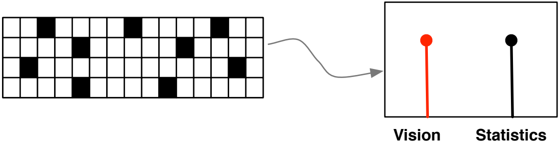


- Online communities of scientists' allow for new ways of connecting researchers to the research literature.
- With **collaborative topic models**, we recommend scientific articles based both on other scientists' preferences and their content.
- We can form both "in-matrix" and "out-of-matrix" predictions. We can learn about which articles are important, and which are interdisciplinary.

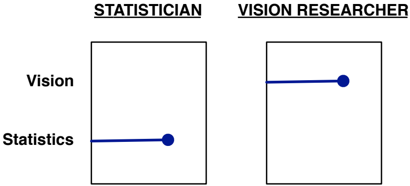
- Consider EM (Dempster et al., 1977). The text lets us estimate its topics:



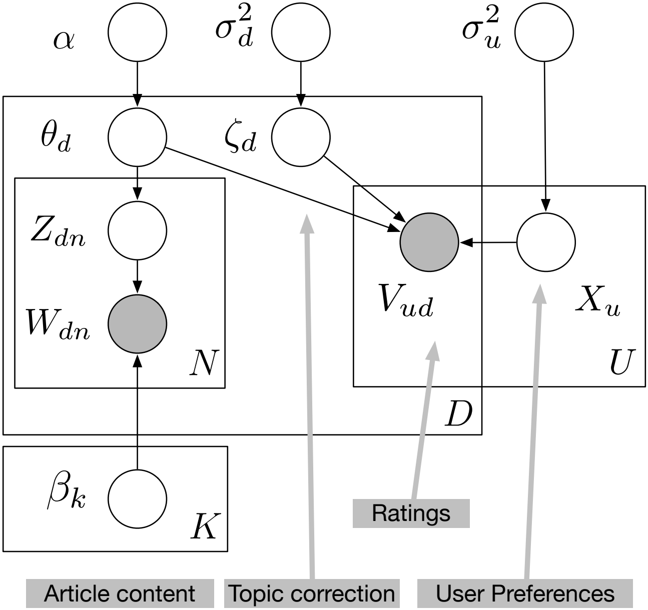
- With user data, we adjust the topics to account for who liked it:



- We can then recommend to users:



Topic models for recommendation



Topic models for recommendation

MENDELEY

- 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

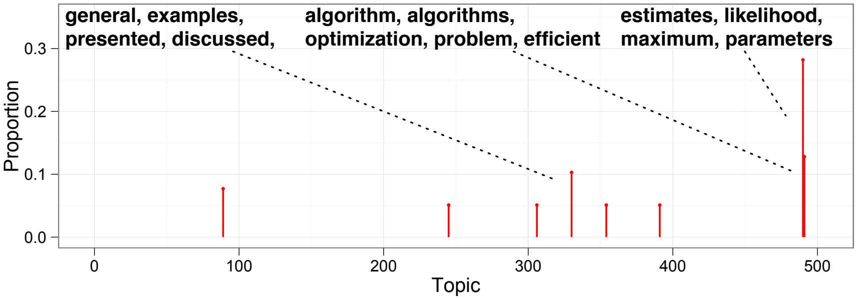
By A. P. DEMPSTER, N. M. LAIRD and D. B. RUBIN

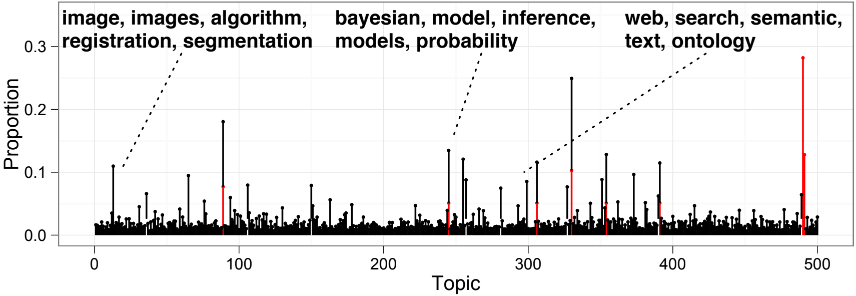
Harvard University and Educational Testing Service

[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.


general, examples,
presented, discussed,
algorithm, algorithms,
optimization, problem, efficient
estimates, likelihood,
maximum, parameters
0.3
0.2
0.1
0.0
Proportion
0
100
200
300
400
500
Topic

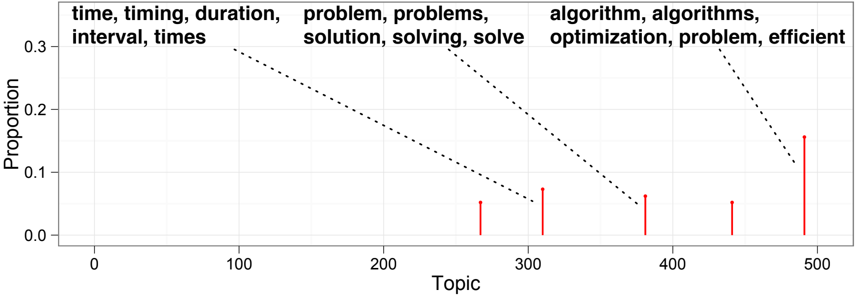


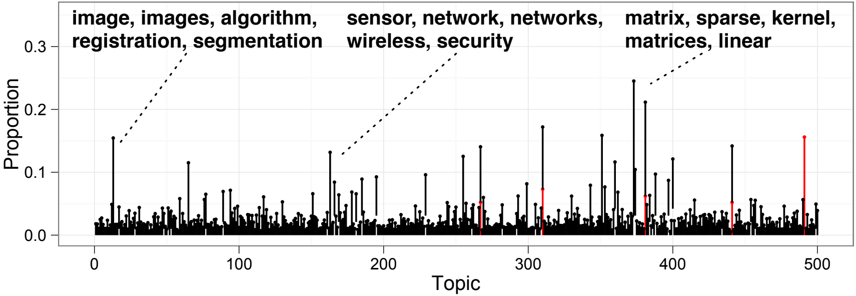
image, images, algorithm, registration, segmentation
bayesian, model, inference, models, probability
web, search, semantic, text, ontology
0.0
0.1
0.2
0.3
Proportion
0
100
200
300
400
500
Topic

Stephen Boyd and
Lieven Vandenberghe

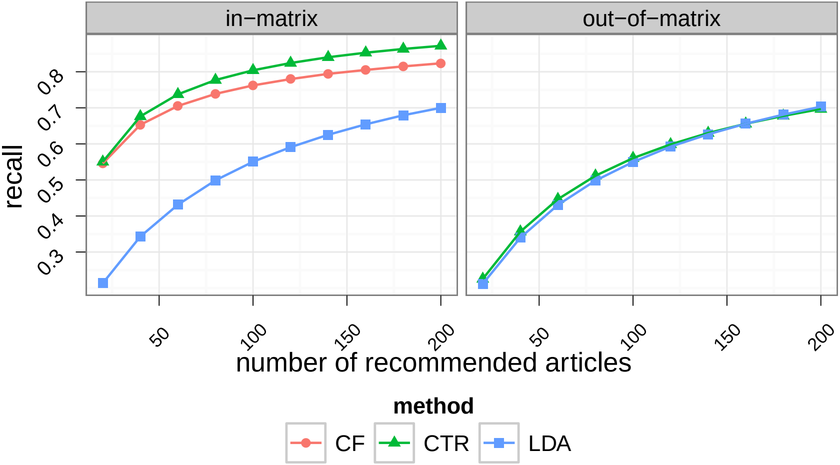
Convex Optimization

CAMBRIDGE


time, timing, duration,
interval, times
problem, problems,
solution, solving, solve
algorithm, algorithms,
optimization, problem, efficient
Proportion
0.3
0.2
0.1
0.0
0
100
200
300
400
500
Topic

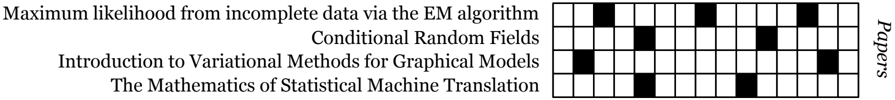

image, images, algorithm,
registration, segmentation
sensor, network, networks,
wireless, security
matrix, sparse, kernel,
matrices, linear
Proportion
0.0
0.1
0.2
0.3
Topic
0
100
200
300
400
500

Topic models for recommendation



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

Topic	estimates, likelihood, maximum, parameters, method
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*

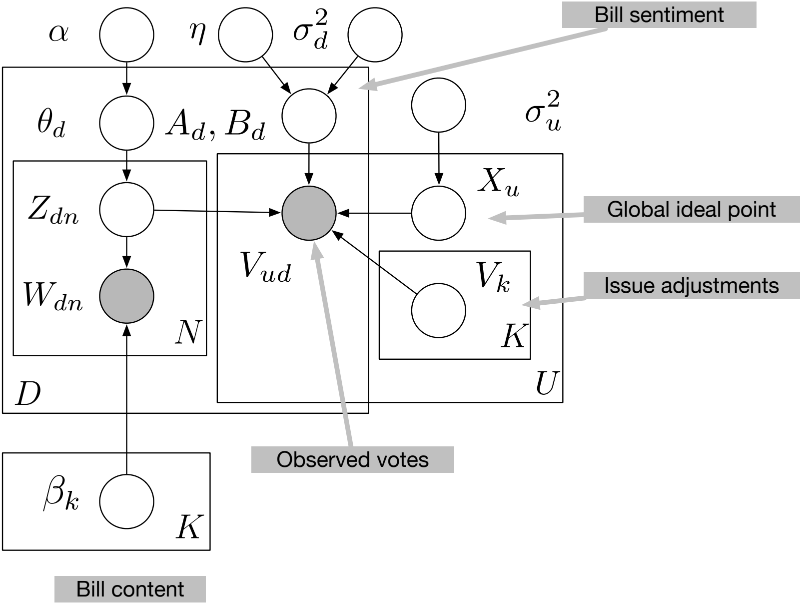
Network Science

Topic	networks, topology, connected, nodes, links, degree
In-topic, read in topic	*Assortative Mixing in Networks* *Characterizing the Dynamical Importance of Network Nodes and Links* *Subgraph Centrality in Complex Networks*
In-topic, read in other topics	*Assortative Mixing in Networks* *The Structure and Function of Complex Networks* *Statistical Mechanics of Complex Networks*
Out-of-topic, read in topic	*Power Law Distributions in Empirical Data* *Graph Structure in the Web* *The Orgins of Bursts and Heavy Tails in Human Dynamics*

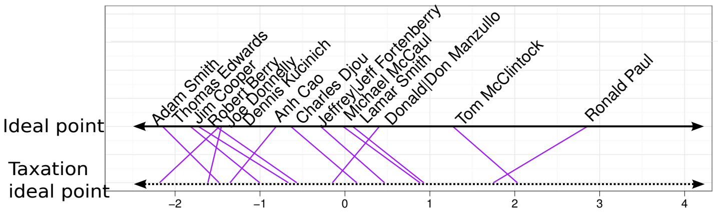
Issue-adjusted ideal points

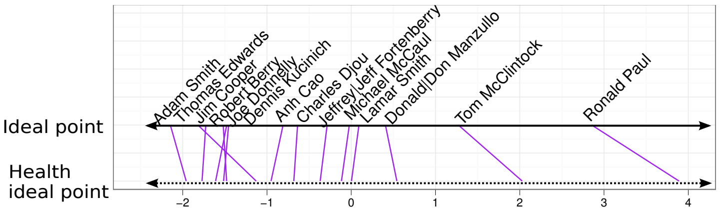
- Our earlier ideal point model uses topics to predict votes from new bills.
- Alternatively, we can use the text to characterize how legislators diverge from their usual ideal points.
- For example: A senator might be left wing, but vote conservatively when it comes to economic matters.

Issue-adjusted ideal points



Issue-adjusted ideal points





Extending LDA

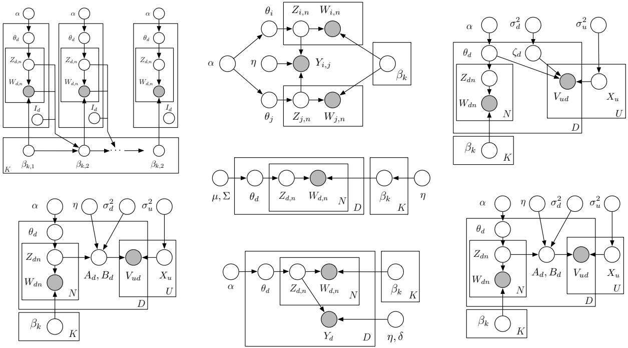
New applications—

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

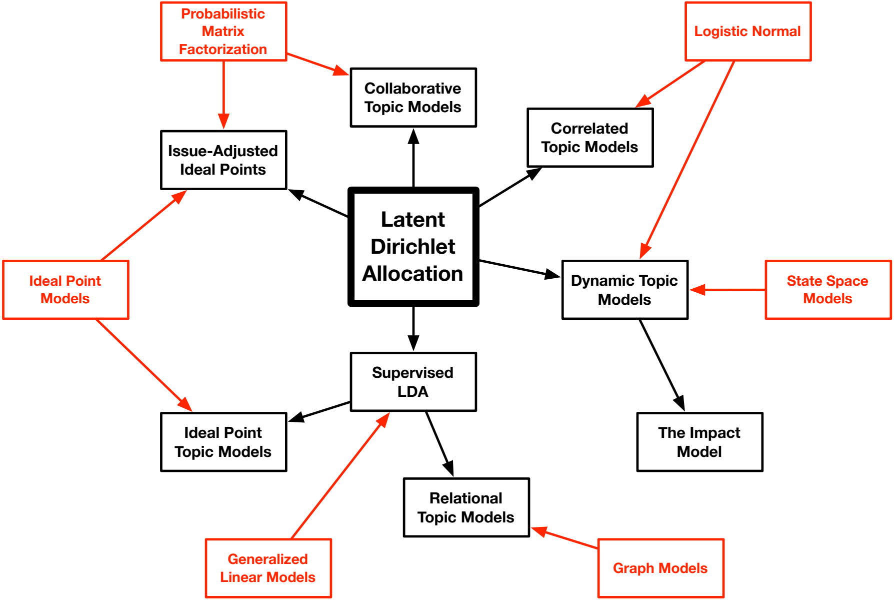
Testing and relaxing assumptions—

- Spike and slab priors
- Models of word contagion
- N-gram topic models

Extending LDA



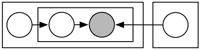
- Each of these models is tailored to solve a problem.
  - Some problems arise from new kinds of data.
  - Others arise from an issue with existing models.
- Probabilistic modeling is a *flexible and modular language for designing solutions to specific problems*.



Probabilistic Matrix Factorization
Logistic Normal
Collaborative Topic Models
Correlated Topic Models
Issue-Adjusted Ideal Points
Latent Dirichlet Allocation
Ideal Point Models
Dynamic Topic Models
State Space Models
Supervised LDA
Ideal Point Topic Models
The Impact Model
Relational Topic Models
Generalized Linear Models
Graph Models

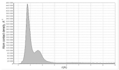
Extending LDA

Make assumptions

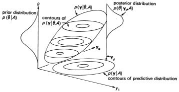


Collect data

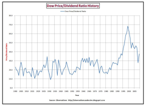
Infer the posterior



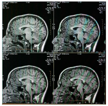
Check



Predict



Explore



Bayesian Nonparametric Models

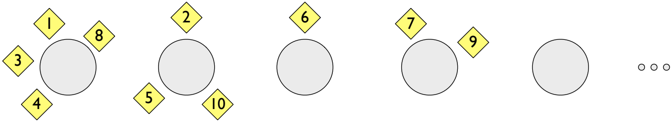
Bayesian nonparametric models

- 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)



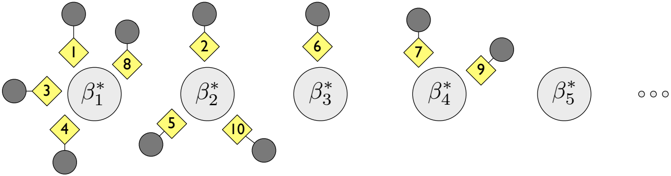
- N customers arrive to an infinite-table restaurant. Each sits down according to how many people are sitting at each table,

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

- The resulting seating plan provides a partition

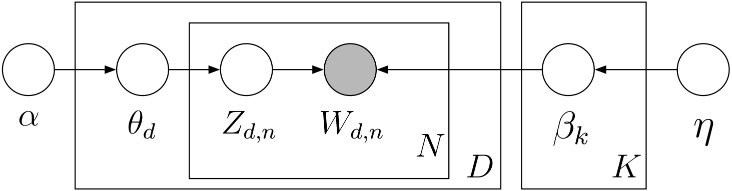
- This distribution is **exchangeable**: Seating plan probabilities are the same regardless of the order of customers (Pitman, 2002).

CRP mixture models



- Associate each table with a topic (β^*).
  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

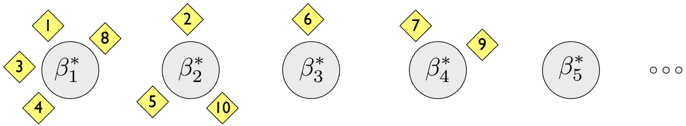


- Mixture models draw each data point from one component.
- The advantage of LDA is that it's a **mixed-membership model**.
- This is addressed by the **Chinese restaurant franchise**.

The Chinese restaurant franchise (Teh et al., 2006)

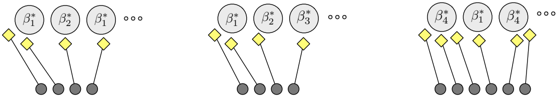
Corpus level restaurant

At the corpus level, topics are drawn from a prior.



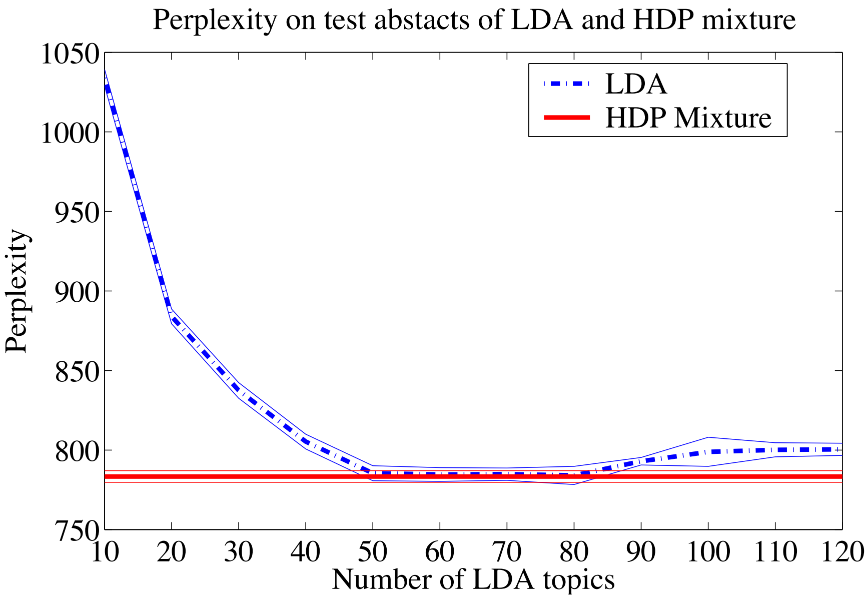
Document level restaurants

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

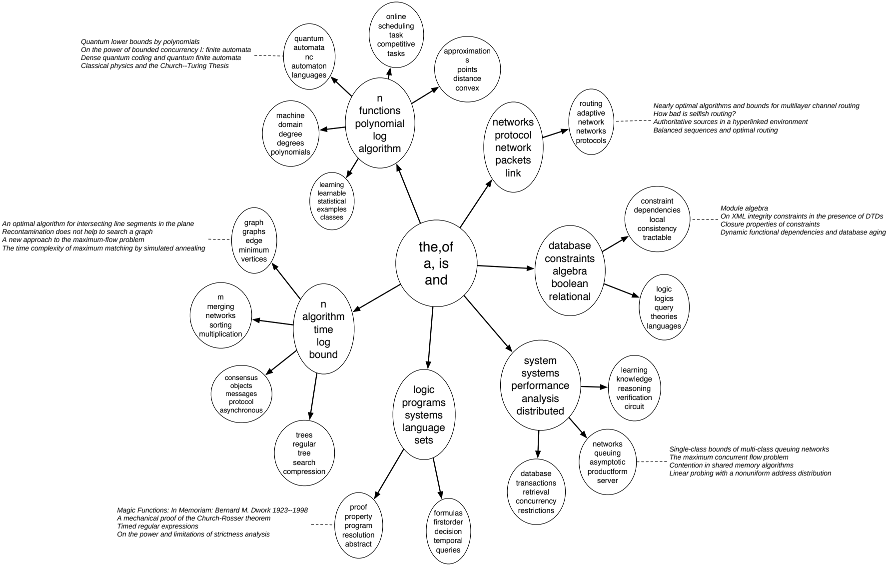


Each word is associated with a customer at the document's restuarant. It is drawn from the topic that its table is associated with.

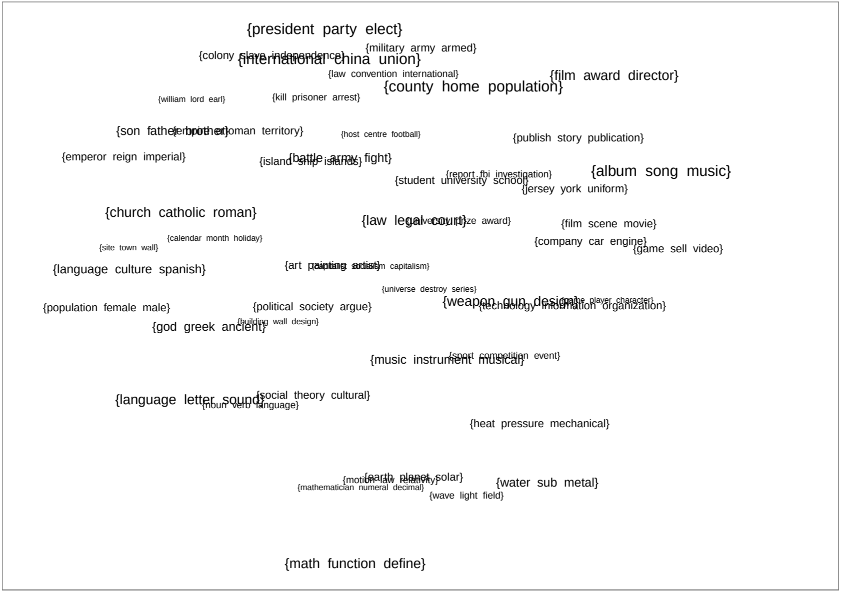
The CRF selects the "right" number of topics (Teh et al., 2006)



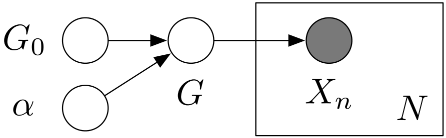
Extended to find hierarchies (Blei et al., 2010)



BNP correlated topic model (Paisley et al., 2011)



Random measures

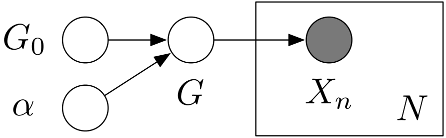


- The CRP metaphors are the best first way to understand BNP methods.
- BNP models were originally developed as **random measure models**.
- E.g., data drawn independently from a random distribution:

$$
\begin{aligned}
G &\sim \mathrm{DP}(\alpha G_0) \\
X_n &\sim G
\end{aligned}
$$

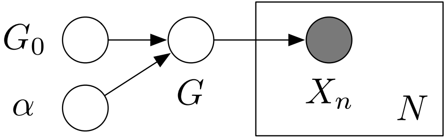
- The random measure perspective helps with certain applications (such as the BNP correlated topic model) and for some approaches to inference.

The Dirichlet process (Ferguson, 1973)



- The Dirichlet process is a distribution of distributions, $G \sim \mathrm{DP}(\alpha, G_0)$
  - *concentration parameter* α (a positive scalar)
  - *base distribution* G_0.
- It produces distributions defined on the same space as its base distribution.

The Dirichlet process (Ferguson, 1973)

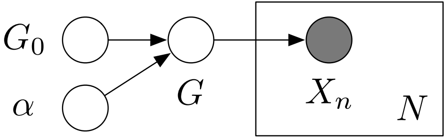


- Consider a partition of the probability space $(A_1, \ldots, A_K)$.
- Ferguson: If for all partitions,

$$\langle G(A_1), \ldots, G(A_k) \rangle \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$

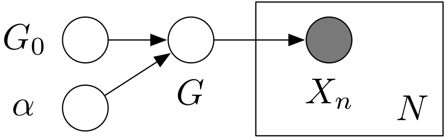
  then G is distributed with a Dirichlet process.
- Note: In this process, the random variables $G(A_k)$ are indexed by the Borel sets of the probability space.

The Dirichlet process (Ferguson, 1973)



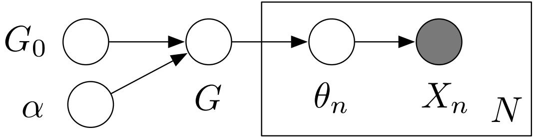
- G is discrete; it places its mass on a countably infinite set of atoms.
- The distribution of the locations is the base distribution G_0.
- As α gets large, G looks more like G_0.
- The conditional $P(G \mid x_{1:N})$ is a Dirichlet process.

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 $\mathrm{CRP}(\alpha)$.

The Dirichlet process mixture (Antoniak, 1974)

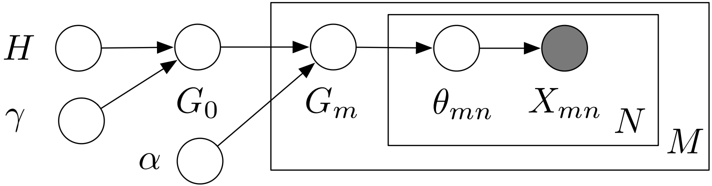


- The draw from G can be a latent parameter to an observed variable:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0) \\ \theta_n &\sim G \\ x_n &\sim p(\cdot \,|\, \theta_n). \end{aligned}$$

- This smooths the random discrete distribution to a *DP mixture*.
- Because of the clustering property, marginalizing out G reveals that this model is the same as a CRP mixture.

Hierarchical Dirichlet processes (Teh et al., 2006)

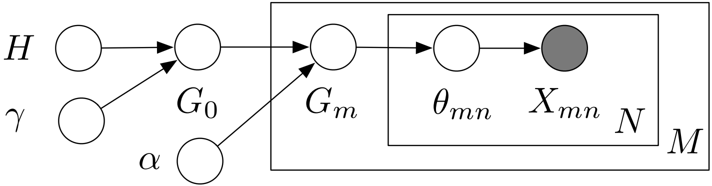


- The hierarchical Dirichlet process (HDP) models *grouped data*.

$$
\begin{aligned}
G_0 &\sim \mathrm{DP}(\gamma, H) \\
G_m &\sim \mathrm{DP}(\alpha, G_0) \\
\theta_{mn} &\sim G_m \\
x_{mn} &\sim p(\cdot \mid \theta_{mn})
\end{aligned}
$$

- 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 θ_{mn} is a topic drawn from G_m.
  - The observation x_{mn} is a word drawn from the topic θ_{mn}.
- Note that in the original topic modeling story, we worked with pointers to topics. Here the θ_{mn} variables are distributions over words.

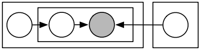
Summary: Bayesian nonparametrics

- 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

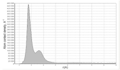
Posterior inference

Make assumptions

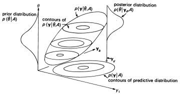


Collect data

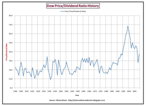
Infer the posterior



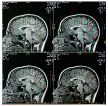
Check



Predict

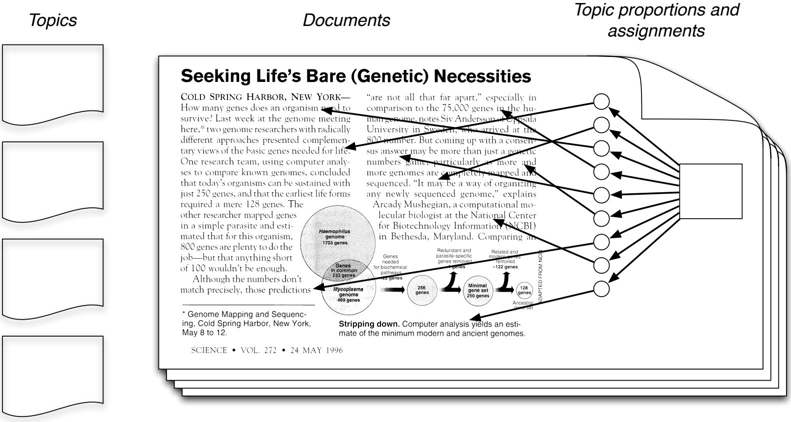


Explore



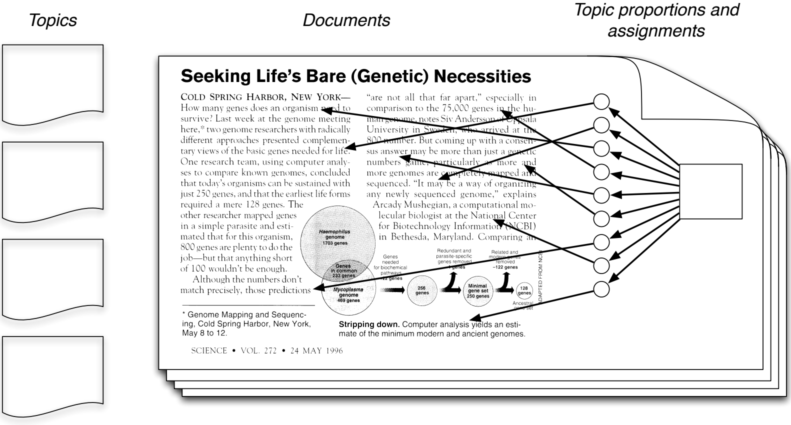
- We can express many kinds of assumptions.
- How can we analyze the collection under those assumptions?

Posterior inference



- Posterior inference is the main computational problem.
- Inference links observed data to statistical assumptions.
- Inference on large data is crucial for topic modeling applications.

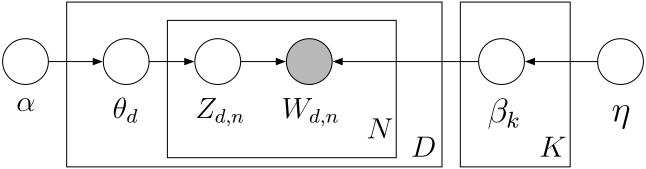
Posterior inference



- Our goal is to compute the distribution of the hidden variables conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

Posterior inference for LDA



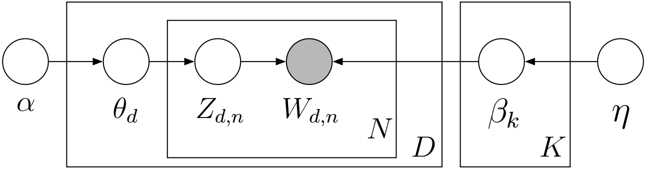
- The joint distribution of the latent variables and documents is

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

- The posterior of the latent variables given the documents is

$$p(\beta, \theta, \mathbf{z} \mid \mathbf{w}).$$

Posterior inference for LDA

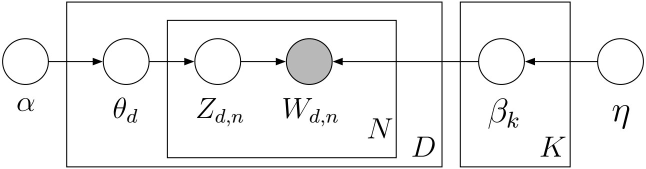


- This is equal to

$$\frac{p(\beta, \theta, \mathbf{z}, \mathbf{w})}{\int_\beta \int_\theta \sum_{\mathbf{z}} p(\beta, \theta, \mathbf{z}, \mathbf{w})}.$$

- We can't compute the denominator, the marginal $p(\mathbf{w})$.
- This is the crux of the inference problem.

Posterior inference for LDA



- 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**.

Stochastic variational inference

- 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.

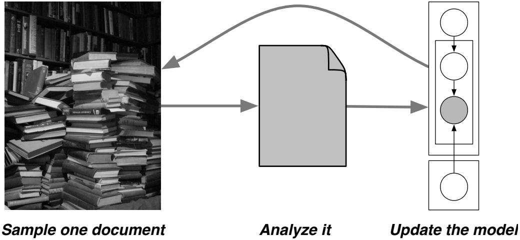
Stochastic variational inference

- 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

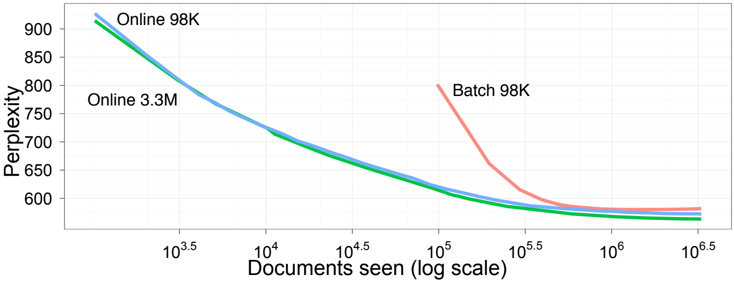
- The structure of the algorithm is:
  - Subsample the data—one data point or a small batch.
  - Infer local variables for the subsample.
  - Update the current estimate of the posterior of the global variables.
  - Repeat.
- This is **efficient**—we need only process one data point at a time.
- We will show: Just as easy as "classical" variational inference

Stochastic variational inference for LDA



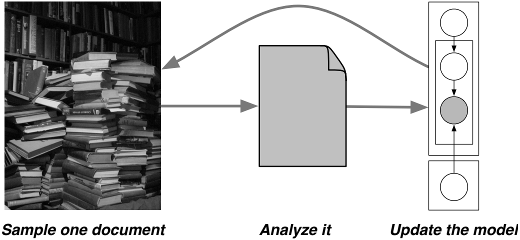
1. Sample a document w_d from the collection
2. Infer how w_d exhibits the current topics
3. Create intermediate topics, formed as though the w_d is the only document.
4. Adjust the current topics according to the intermediate topics.
5. Repeat.

Stochastic variational inference for LDA



Documents analyzed	**2048**	**4096**	**8192**	**12288**	**16384**	**32768**	**49152**	**65536**
Top eight words	systems road made service announced national west language	systems health communication service billion language care road	service systems health companies market communication company billion	service systems companies business company billion health industry	service companies systems business company industry market billion	business service companies industry company management systems services	business service companies industry services company management public	business industry service companies services company management public

Stochastic variational inference for LDA



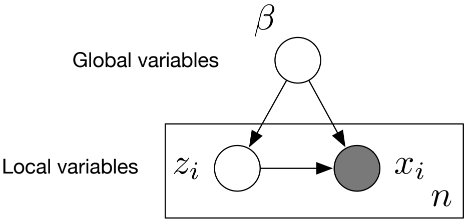
We have developed stochastic variational inference algorithms for

- 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

Organization

- 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

Organization

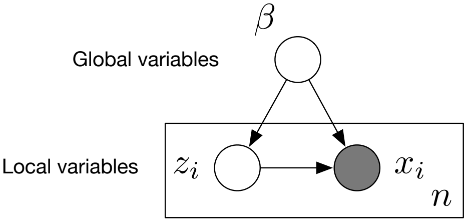


- We consider a **generic model**.
  - Hidden variables are local or global.

- 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.

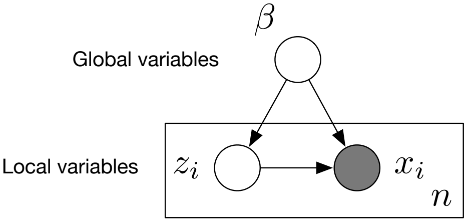
Generic model



$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i \mid \beta) p(x_i \mid z_i, \beta)$$

- The observations are $x = x_{1:n}$.
- The **local** variables are $z = z_{1:n}$.
- Th **global** variables are β.
- The ith data point x_i only depends on z_i and β.
- Our goal is to compute $p(\beta, z \mid x)$.

Generic model

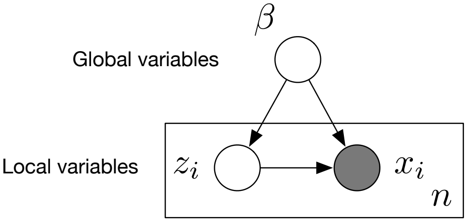


$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i \mid \beta) p(x_i \mid z_i, \beta)$$

- 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,

$$\begin{aligned} p(z_i \mid \beta, x_i) &= h(z_i) \exp\{\eta_\ell(\beta, x_i)^\top z_i - a(\eta_\ell(\beta, x_i))\} \\ p(\beta \mid z, x) &= h(\beta) \exp\{\eta_g(z, x)^\top \beta - a(\eta_g(z, x))\}. \end{aligned}$$

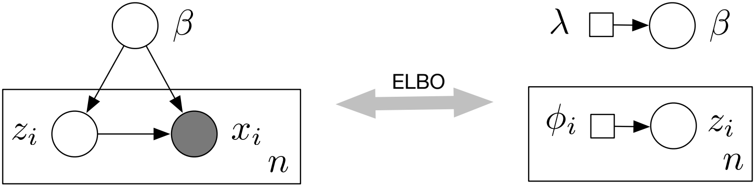
Generic model



$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta) \prod_{i=1}^{n} p(z_i \mid \beta) p(x_i \mid z_i, \beta)$$

- 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

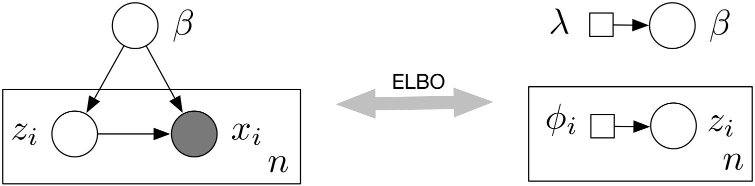


- Introduce a **variational distribution** over the latent variables $q(\beta, z)$.
- We optimize the **evidence lower bound** (ELBO) with respect to q,

$$\log p(x) \geq \mathrm{E}_q[\log p(\beta, Z, x)] - \mathrm{E}_q[\log q(\beta, Z)].$$

- Up to a constant, this is the negative KL between q and the posterior.

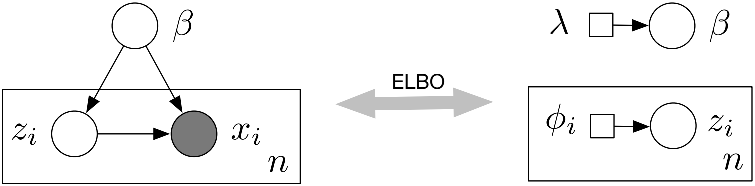
Mean-field variational inference



We can derive the ELBO with Jensen's inequality:

$$
\begin{aligned}
\log p(x) &= \log \int p(\beta, Z, x)\, dZ d\beta \\
&= \log \int p(\beta, Z, x) \frac{q(\beta, Z)}{q(\beta, Z)} dZ d\beta \\
&\geq \int q(\beta, Z) \log \frac{p(\beta, Z, x)}{q(\beta, Z)} dZ d\beta \\
&= \mathrm{E}_q[\log p(\beta, Z, x)] - \mathrm{E}_q[\log q(\beta, Z)].
\end{aligned}
$$

Mean-field variational inference



- We specify $q(\beta, z)$ to be a fully factored variational distribution,

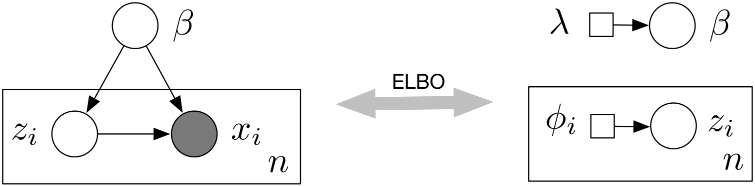
$$q(\beta, z) = q(\beta \,|\, \lambda) \prod_{i=1}^{n} q(z_i \,|\, \phi_i).$$

- Each instance of each variable has its own distribution.
- Each component is in the same family as the model conditional,

$$\begin{aligned} p(\beta \,|\, z, x) &= h(\beta) \exp\{\eta_g(z, x)^\top \beta - a(\eta_g(z, x))\} \\ q(\beta \,|\, \lambda) &= h(\beta) \exp\{\lambda^\top \beta - a(\lambda)\} \end{aligned}$$

(And, same for the local variational parameters.)

Mean-field variational inference

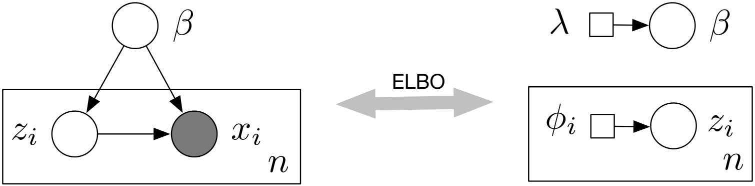


- We optimize the ELBO with respect to these parameters,

$$\mathscr{L}(\lambda, \phi_{1:n}) = \mathrm{E}_q[\log p(\beta, Z, x)] - \mathrm{E}_q[\log q(\beta, Z)].$$

- Same as finding the $q(\beta, z)$ that is closest in KL divergence to $p(\beta, z \mid x)$
- The ELBO links the observations/model to the variational distribution.

Mean-field variational inference



- Coordinate ascent: Iteratively update each parameter, holding others fixed.
- With respect to the global parameter, the gradient is

  $$\nabla_\lambda \mathscr{L} = a''(\lambda)(\mathrm{E}_\phi[\eta_g(Z,x)] - \lambda).$$

  This leads to a simple coordinate update

  $$\lambda^* = \mathrm{E}_\phi\left[\eta_g(Z,x)\right].$$

- The local parameter is analogous.

Mean-field variational inference

Initialize λ randomly.

Repeat until the ELBO converges

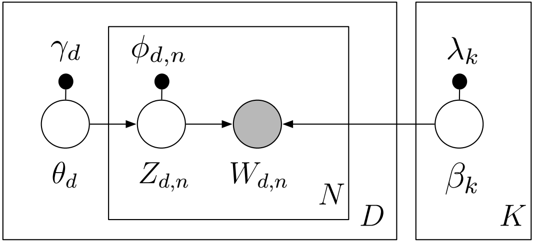
1. For each data point, update the local variational parameters:

$$\phi_i^{(t)} = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}.$$

2. Update the global variational parameters:

$$\lambda^{(t)} = \mathrm{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})].$$

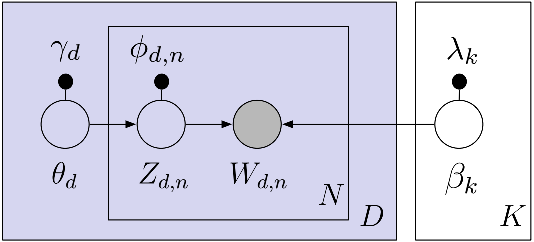
Mean-field variational inference for LDA



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

$$q(\beta, \theta, z) = \prod_{k=1}^{K} q(\beta_k \,|\, \lambda_k) \prod_{d=1}^{D} q(\theta_d \,|\, \gamma_d) \prod_{n=1}^{N} q(z_{d,n} \,|\, \phi_{d,n})$$

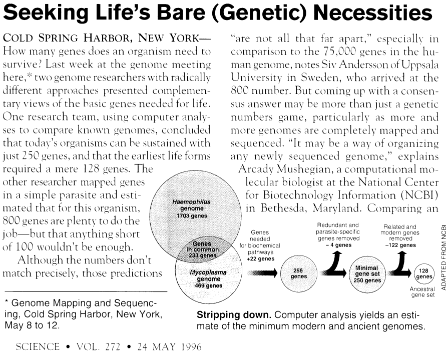
Mean-field variational inference for LDA



- In the "local step" we iteratively update the parameters for each document, holding the topic parameters fixed.

$$\begin{aligned} \gamma^{(t+1)} &= \alpha + \textstyle\sum_{n=1}^{N} \phi_n^{(t)} \\ \phi_n^{(t+1)} &\propto \exp\{\mathbb{E}_q[\log \theta] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}. \end{aligned}$$

Mean-field variational inference for LDA



Seeking Life's Bare (Genetic) Necessities

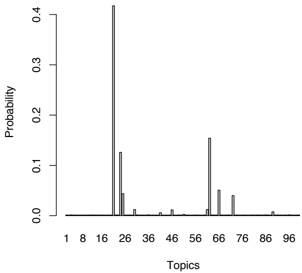
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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

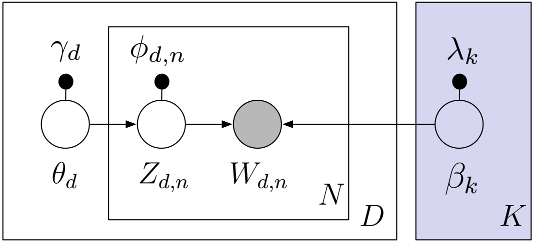
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996



Mean-field variational inference for LDA



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

Mean-field variational inference for LDA

```
Initialize topics randomly.
repeat
    for each document do
        repeat
            Update the topic assignment variational parameters.
            Update the topic proportions variational parameters.
        until document objective converges
    end for
    Update the topics from aggregated per-document parameters.
until corpus objective converges.
```

Mean-field variational inference

Initialize λ randomly.

Repeat until the ELBO converges

1. Update the local variational parameters for each data point,

$$\phi_i^{(t)} = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}.$$

2. Update the global variational parameters,

$$\lambda^{(t)} = \mathrm{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})].$$

- Note the relationship to existing algorithms like EM and Gibbs sampling.
- But we must analyze the whole data set before completing one iteration.

Mean-field variational inference

Initialize λ randomly.

Repeat until the ELBO converges

1. Update the local variational parameters for each data point,

$$\phi_i^{(t)} = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)] \quad \text{for } i \in \{1, \ldots, n\}.$$

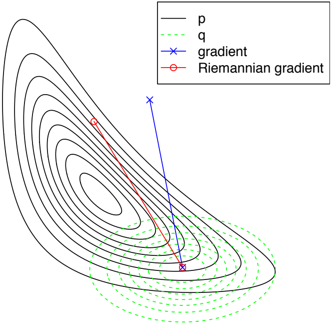
2. Update the global variational parameters,

$$\lambda^{(t)} = \mathrm{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})].$$

To make this more efficient, we need two ideas:

- Natural gradients
- Stochastic optimization

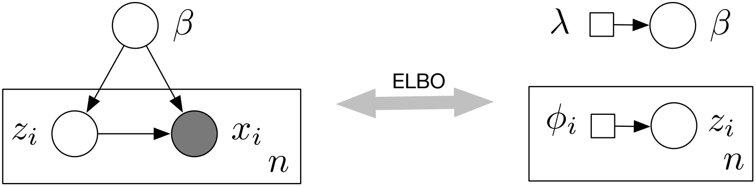
The natural gradient



(from Honkela et al., 2010)

- 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.

The natural gradient



- 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

$$\hat{\nabla}_\lambda \mathscr{L} = \mathrm{E}_\phi[\eta_g(Z, x)] - \lambda.$$

- We can compute the natural gradient by computing the coordinate updates in parallel and subtracting the current variational parameters.

Stochastic optimization

A STOCHASTIC APPROXIMATION METHOD[1]

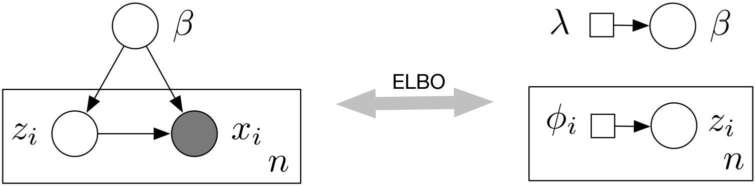
BY HERBERT ROBBINS AND SUTTON MONRO

University of North Carolina

1. Summary. Let $M(x)$ denote the expected value at level x of the response to a certain experiment. $M(x)$ is assumed to be a monotone function of x but is unknown to the experimenter, and it is desired to find the solution $x = \theta$ of the equation $M(x) = \alpha$, where α is a given constant. We give a method for making successive experiments at levels $x_1, x_2, \cdots$ in such a way that x_n will tend to θ in probability.

- 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.

Stochastic optimization



- We will use stochastic optimization for global variables.
- Let $\nabla_\lambda \mathscr{L}_t$ be a realization of a random variable whose expectation is $\nabla_\lambda \mathscr{L}$.
- Iteratively set

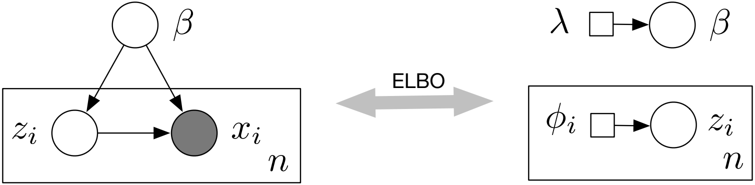
$$\lambda^{(t)} = \lambda^{(t-1)} + \epsilon_t \nabla_\lambda \mathscr{L}_t$$

- This leads to a local optimum when

$$\begin{aligned} \textstyle\sum_{t=1}^{\infty} \epsilon_t &= \infty \\ \textstyle\sum_{t=1}^{\infty} \epsilon_t^2 &< \infty \end{aligned}$$

- Next step: Form a noisy gradient.

A noisy natural 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 β is

$$\eta_g(z_{1:n}, x_{1:n}) = \langle \alpha_1 + \textstyle\sum_{i=1}^n f(z_i, x_i), \alpha_2 + n \rangle.$$

- See the discussion of conjugacy in Bernardo and Smith (1994).

A noisy natural gradient

- With local and global variables, we decompose the ELBO

$$\mathscr{L} = \mathrm{E}[\log p(\beta)] - \mathrm{E}[\log q(\beta)] + \sum_{i=1}^{n} \mathrm{E}[\log p(z_i, x_i \mid \beta)] - \mathrm{E}[\log q(z_i)]$$

- Sample a single data point t uniformly from the data and define

$$\mathscr{L}_t = \mathrm{E}[\log p(\beta)] - \mathrm{E}[\log q(\beta)] + n(\mathrm{E}[\log p(z_t, x_t \mid \beta)] - \mathrm{E}[\log q(z_t)]).$$

1. **The ELBO is the expectation of $\mathscr{L}_t$ 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

$$\nabla_\lambda \hat{\mathscr{L}}_t = \mathrm{E}_{\phi_t}[\eta_t(Z_t, x_t)] - \lambda.$$

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

- In contrast, the full natural gradient requires all local parameters.

Stochastic variational inference

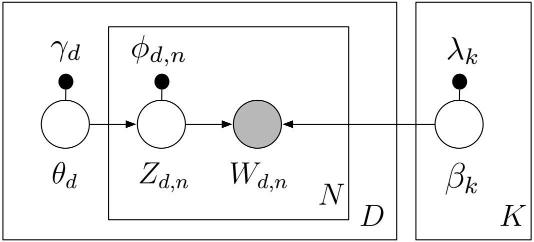
Initialize global parameters λ randomly.

Set the step-size schedule ϵ_t appropriately.

Repeat forever

1. Sample a data point uniformly,
$$x_t \sim \mathrm{Uniform}(x_1, \ldots, x_n).$$
2. Compute its local variational parameter,
$$\phi = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_t)].$$
3. Pretend its the only data point in the data set,
$$\hat{\lambda} = \mathrm{E}_\phi[\eta_t(Z_t, x_t)].$$
4. Update the current global variational parameter,
$$\lambda^{(t)} = (1 - \epsilon_t)\lambda^{(t-1)} + \epsilon_t \hat{\lambda}.$$

Stochastic variational inference in LDA



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

Stochastic variational inference in LDA

Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$

Initialize $\boldsymbol{\lambda}$ randomly.

for $t = 0$ to ∞ **do**

  Choose a random document w_t

  Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)

  repeat

    Set $\phi_{t,n} \propto \exp\{\mathbb{E}_q[\log \theta_t] + \mathbb{E}_q[\log \beta_{\cdot,w_n}]\}$

    Set $\gamma_t = \alpha + \sum_n \phi_{t,n}$

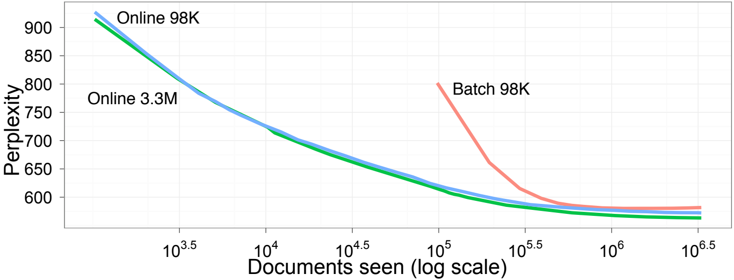
  until $\frac{1}{K}\sum_k |\text{change in } \gamma_{t,k}| < \epsilon$

  Compute $\tilde{\lambda}_k = \eta + D \sum_n w_{t,n} \phi_{t,n}$

  Set $\lambda_k = (1 - \rho_t)\lambda_k + \rho_t \tilde{\lambda}_k$.

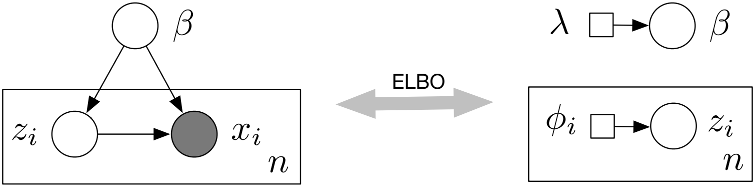
end for

Stochastic variational inference in LDA



Documents analyzed	**2048**	**4096**	**8192**	**12288**	**16384**	**32768**	**49152**	**65536**
Top eight words	systems road made service announced national west language	systems health communication service billion language care road	service systems health companies market communication company billion	service systems companies business company billion health industry	service companies systems business company industry market billion	business service companies industry company management systems services	business service companies industry services company management public	business industry service companies services company management public

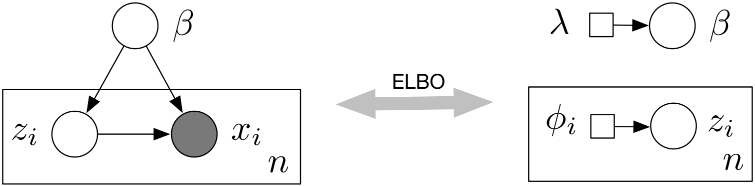
Stochastic variational inference



We defined a generic algorithm for scalable variational inference.

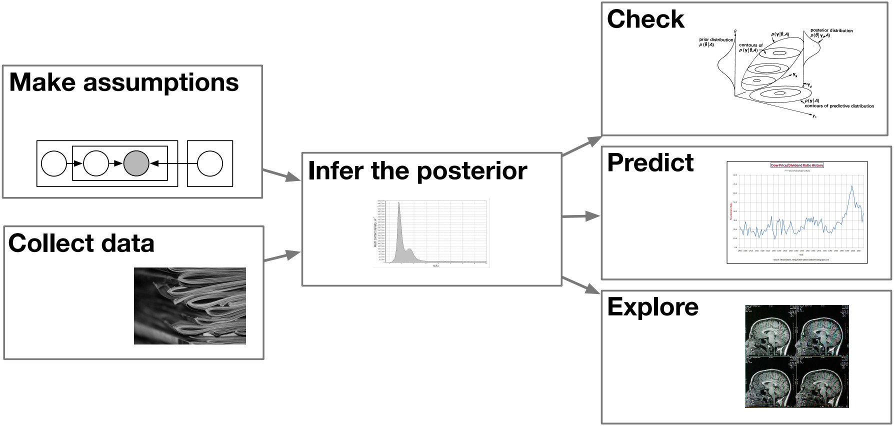
- Bayesian mixture models
- Time series models
  (variants of HMMs, Kalman filters)
- Factorial models
- Matrix factorization
  (e.g., factor analysis, PCA, CCA)
- Dirichlet process mixtures, HDPs
- Multilevel regression
  (linear, probit, Poisson)
- Stochastic blockmodels
- Mixed-membership models
  (LDA and some variants)

Stochastic variational inference



- 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.

Stochastic 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.

Nonconjugate variational inference

- 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.

Nonconjugate variational inference

- Bishop (2006) showed that the optimal mean-field variational distribution is

$$
\begin{aligned}
q^*(z) &\propto \exp\left\{\mathrm{E}_{q(\beta)}\left[\log p(z \,|\, \beta, x)\right]\right\} \\
q^*(\beta) &\propto \exp\left\{\mathrm{E}_{q(z)}\left[\log p(\beta \,|\, z, x)\right]\right\}
\end{aligned}
$$

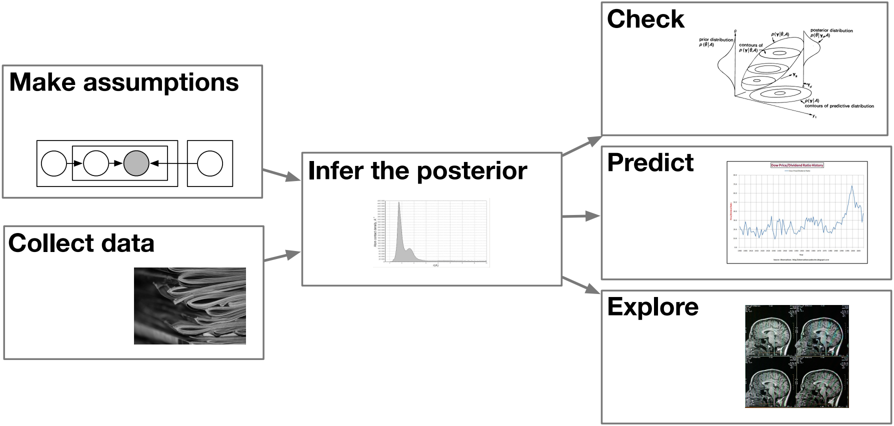
- In conjugate models, we can compute these expectations.
  This determines the form of the optimal variational distribution.

- In nonconjugate models we can't compute the expectations.

- But, under certain conditions, we can use Taylor approximations.
  This leads to Gaussian variational distributions.

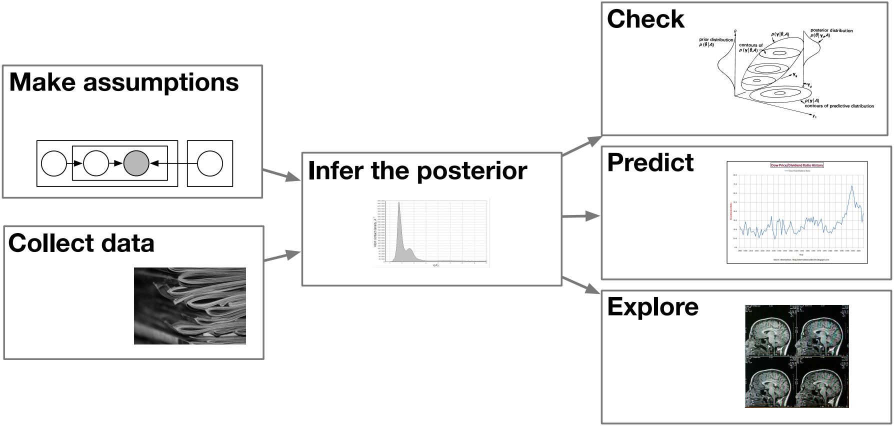
Using and Checking Topic Models

Using and checking topic models



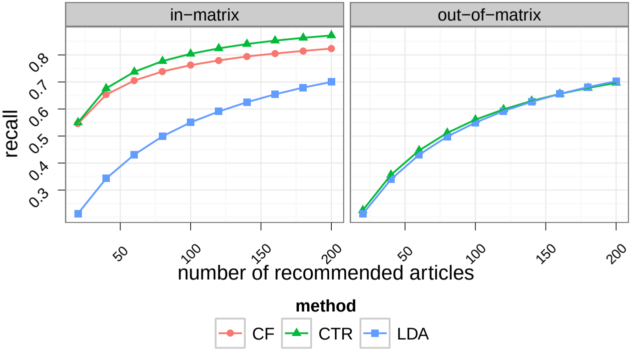
- We have collected data, selected a model, and inferred the posterior.
- How do we use the topic model?

Using and checking topic models



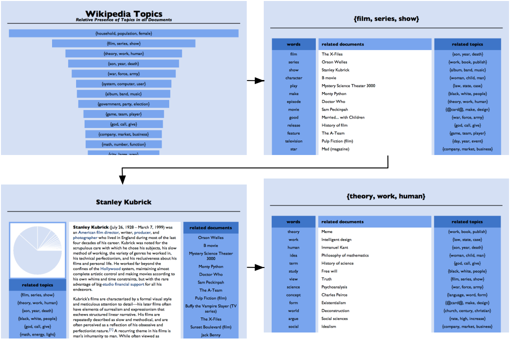
- Using a model means doing something with the posterior inference.
- E.g., visualization, prediction, assessing document similarity, using the representation in a downstream task (like IR)

Using and checking topic models



- 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.

Using and checking topic models



- 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,

$$
\begin{aligned}
p(w \mid \mathscr{D}, \mathbf{w}_{\mathrm{obs}}) &= \int_{\beta}\int_{\theta}\left(\textstyle\sum_{k=1}^{K} \theta_k \beta_{k,w}\right) p(\theta \mid \mathbf{w}_{\mathrm{obs}}, \beta) p(\beta \mid \mathscr{D}) \\
&\approx \int_{\beta}\int_{\theta}\left(\textstyle\sum_{k=1}^{K} \theta_k \beta_{k,w}\right) q(\theta) q(\beta) \\
&= \mathrm{E}_q[\theta \mid \mathbf{w}_{\mathrm{obs}}]^{\top} \mathrm{E}_q[\beta_{\cdot,w} \mid \mathscr{D}].
\end{aligned}
$$

The predictive score

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

$$s = \sum_{w \in \mathbf{w}_{\text{held out}}} \log p(w \mid \mathscr{D}, \mathbf{w}_{\text{obs}}) \tag{1}$$

- In the predictive distribution, q is any approximate poterior. 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

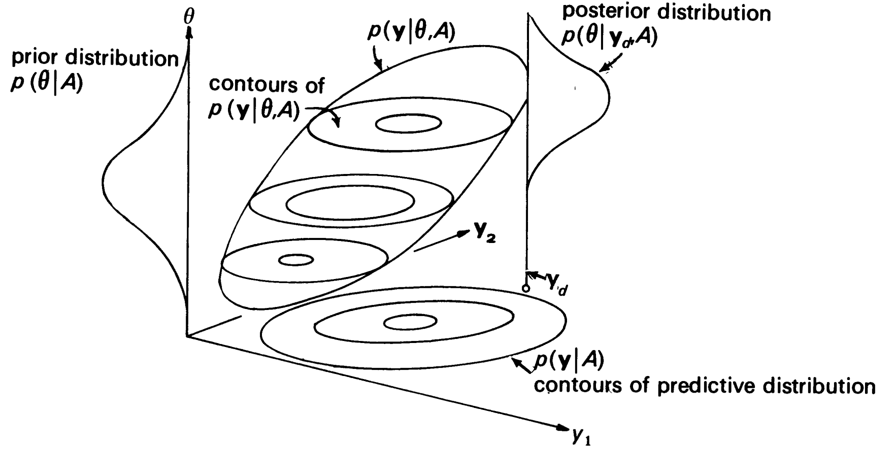
	Nature	*New York Times*	*Wikipedia*
LDA 100	-7.26	-7.66	-7.41
LDA 200	-7.50	-7.78	-7.64
LDA 300	-7.86	-7.98	-7.74
HDP	**-6.97**	**-7.38**	**-7.07**

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.

Posterior predictive checks



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*:

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

- Locating **y** in the predictive distribution indicates if we can "trust" the model.
- Or, locating a **discrepancy function** $g(\mathbf{y})$ in its predictive distribution indicates if what is important to us is captured in the model.

Posterior predictive checks

- Rubin (1984) located the data $\mathbf{y}$ in the **posterior** $p(y \mid \mathbf{y}, M)$.
- Gelman, Meng, Stern (1996) expanded this idea to “realized discrepancies” that include **hidden variables** $g(\mathbf{y}, \mathbf{z})$.
- 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.

Posterior predictive checks in topic models

- 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

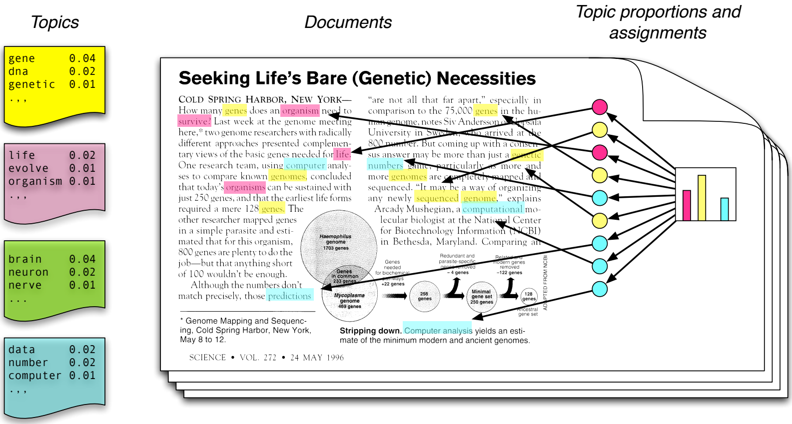
- This realized discrepancy measures model fitness
- Can use it to measure model fitness **per topic**.
- Helps us explore parts of the model that fit well.

Discussion

Probabilistic topic models

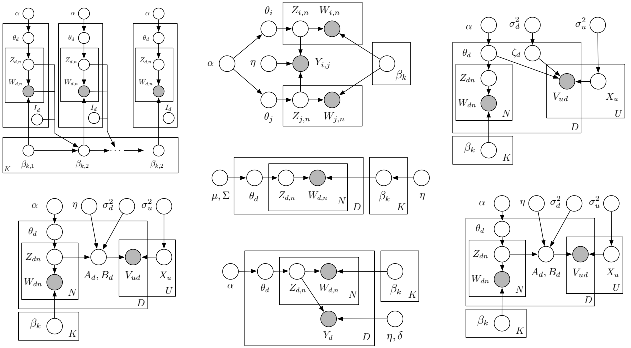
- What are topic models?
- What kinds of things can they do?
- How do I compute with a topic model?
- How do I evaluate and check a topic model?
- What are some unanswered questions in this field?
- How can I learn more?

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.

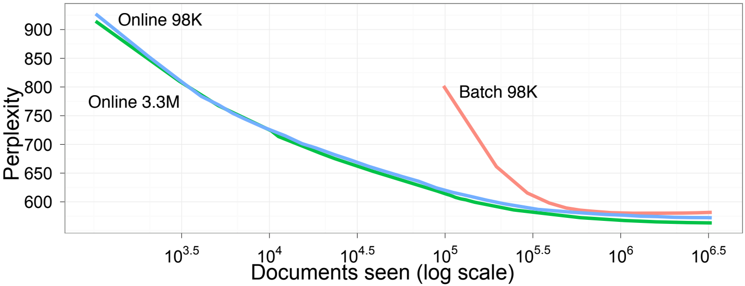
Extensions of LDA



Topic models can be adapted to many settings

- relax assumptions
- combine models
- model more complex data

Posterior inference

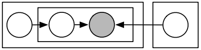


- Posterior inference is the central computational problem.
- Stochastic variational inference is a scalable algorithm.
- We can handle nonconjugacy with Laplace inference.
- (Note: There are many types of inference we didn't discuss.)

Posterior predictive checks

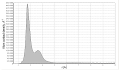
Probabilistic models

Make assumptions

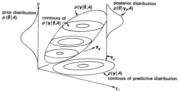


Collect data

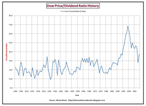
Infer the posterior



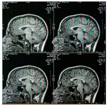
Check



Predict



Explore



Implementations of LDA

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 humanties? 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.”

(J. Tukey, *The Future of Data Analysis*, 1962)

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

(The Velvet Underground, *Rock & Roll*, 1969)