Topic models
(Latent Dirichlet Allocation)

Klinton Bicknell

borrowing from Roger Levy, Tom Griffiths, David Blei

Topic modeling - Motivation

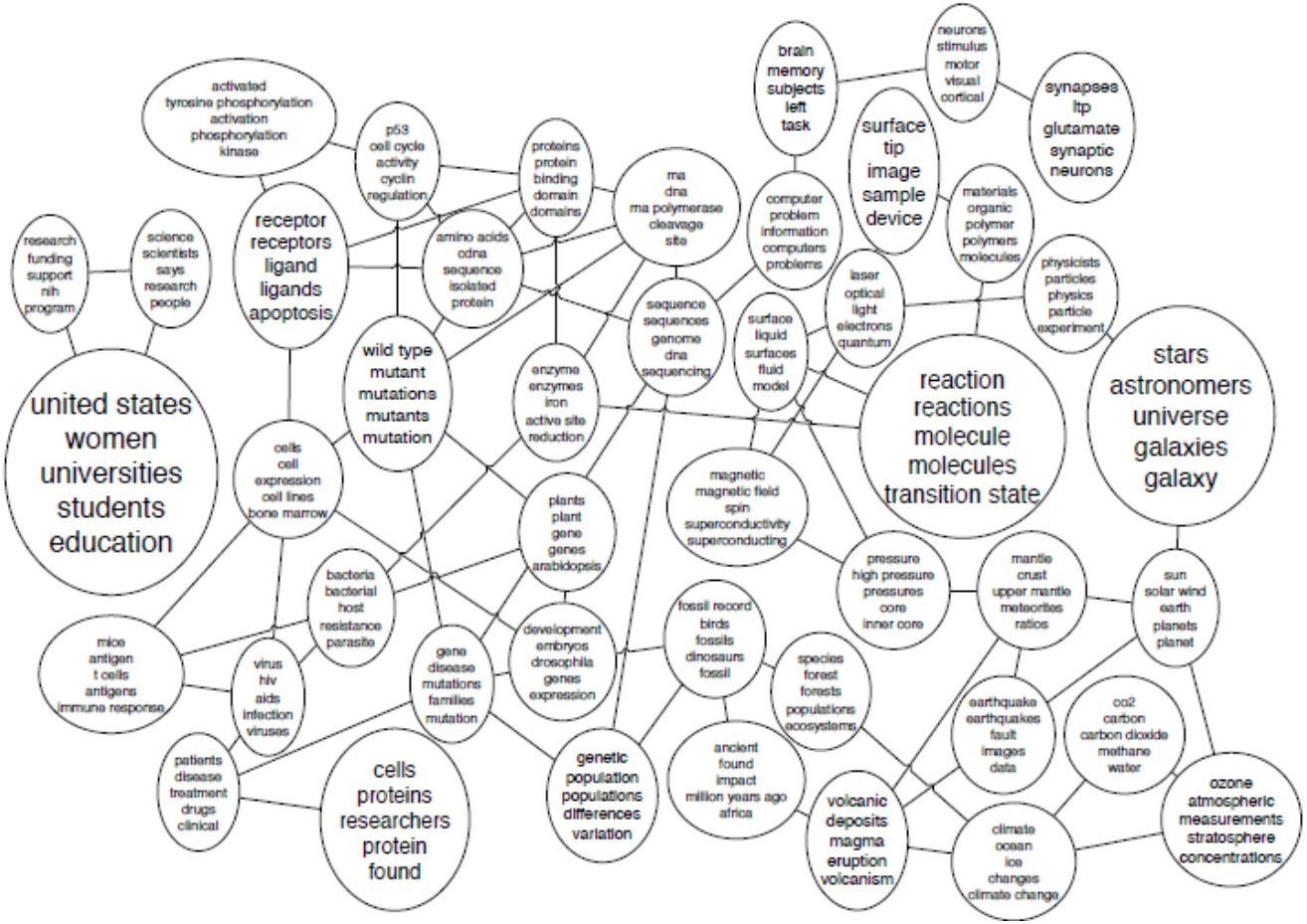
Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

1. Uncover the hidden topical patterns that pervade the collection.
2. Annotate the documents according to those topics.
3. Use the annotations to organize, summarize, and search the texts.

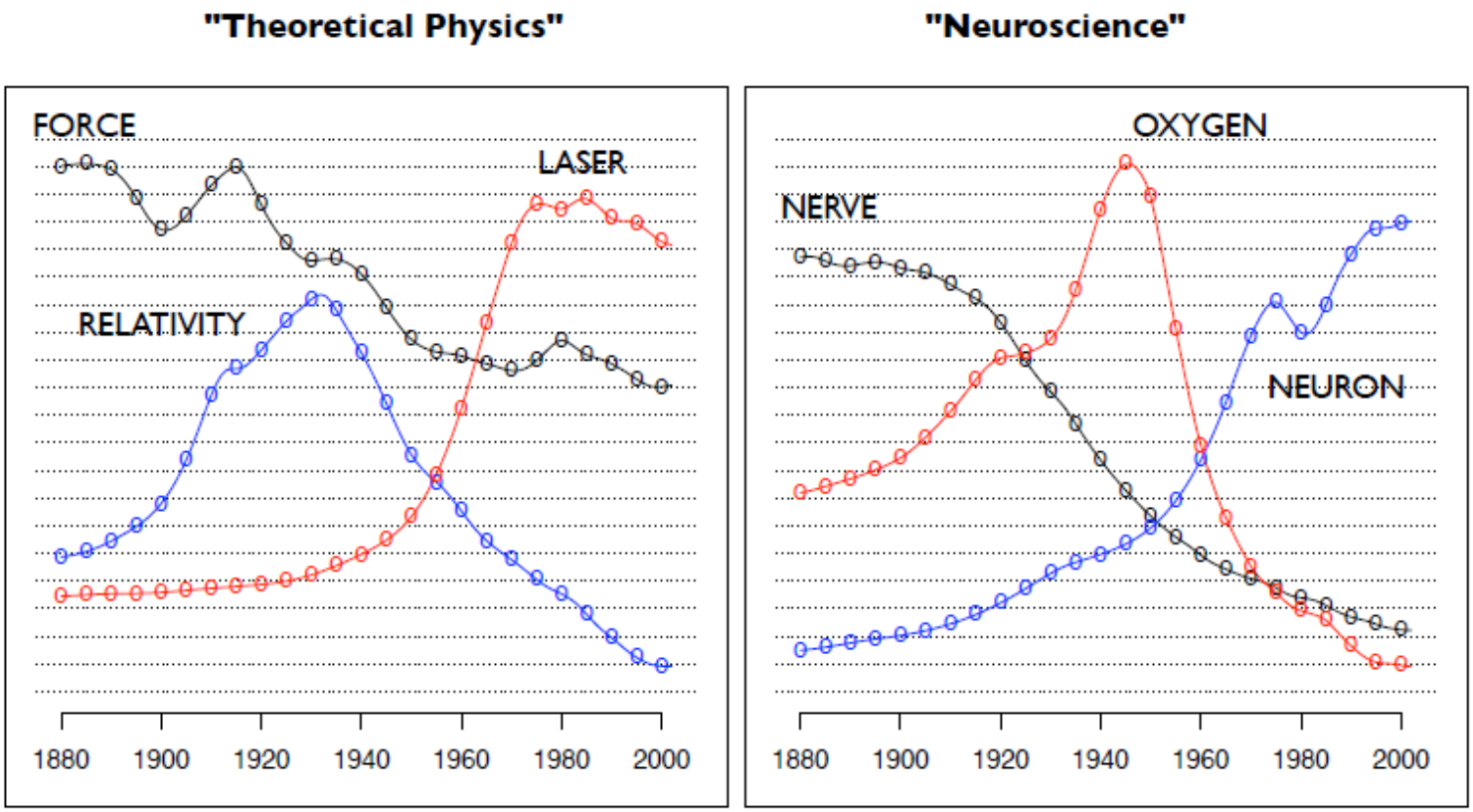
Discover topics from a corpus

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

Model connections between topics



Model the evolution of topics over time (or other relevant variable)



Connection to ML research

From a machine learning perspective, topic modeling is a case study in applying hierarchical Bayesian models to grouped data, like documents or images. Topic modeling research touches on

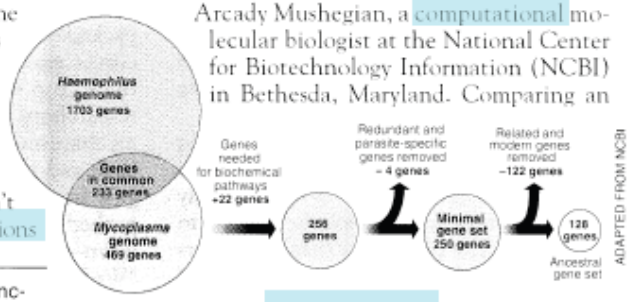
- Directed graphical models
- Conjugate priors and nonconjugate priors
- Time series modeling
- Modeling with graphs
- Hierarchical Bayesian methods
- Fast approximate posterior inference (MCMC, variational methods)
- Exploratory data analysis
- Model selection and nonparametric Bayesian methods
- Mixed membership models

Intuition behind 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



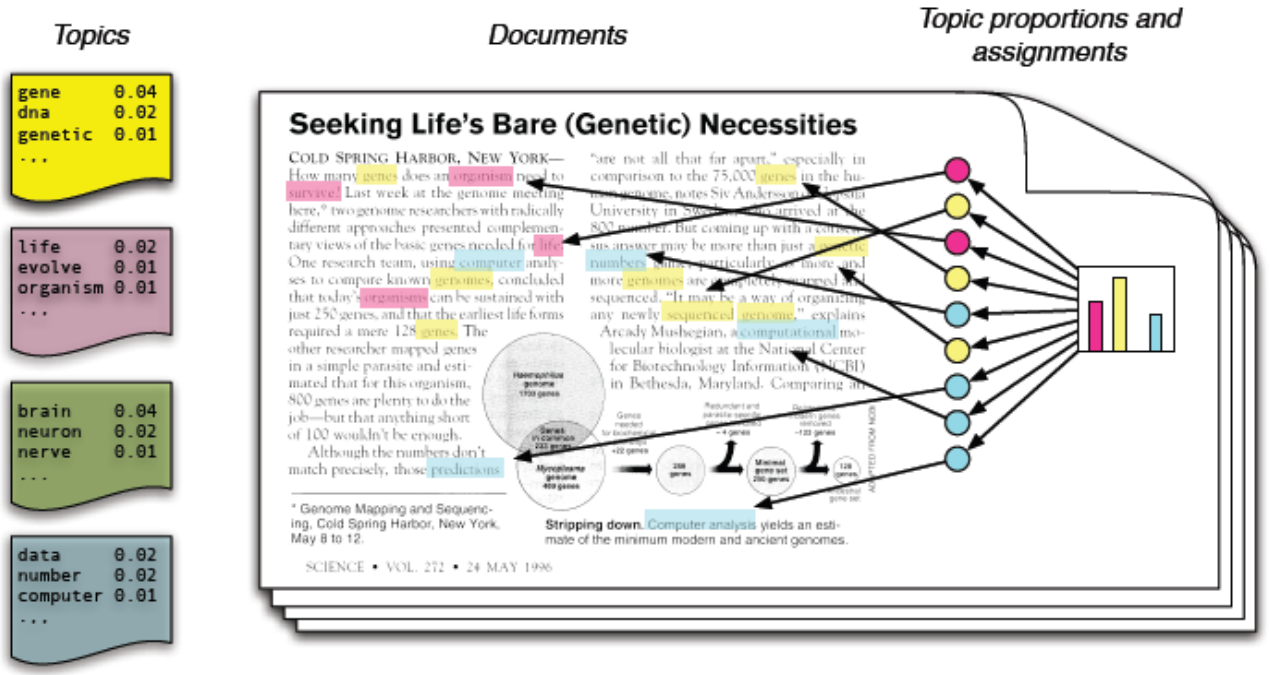
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996

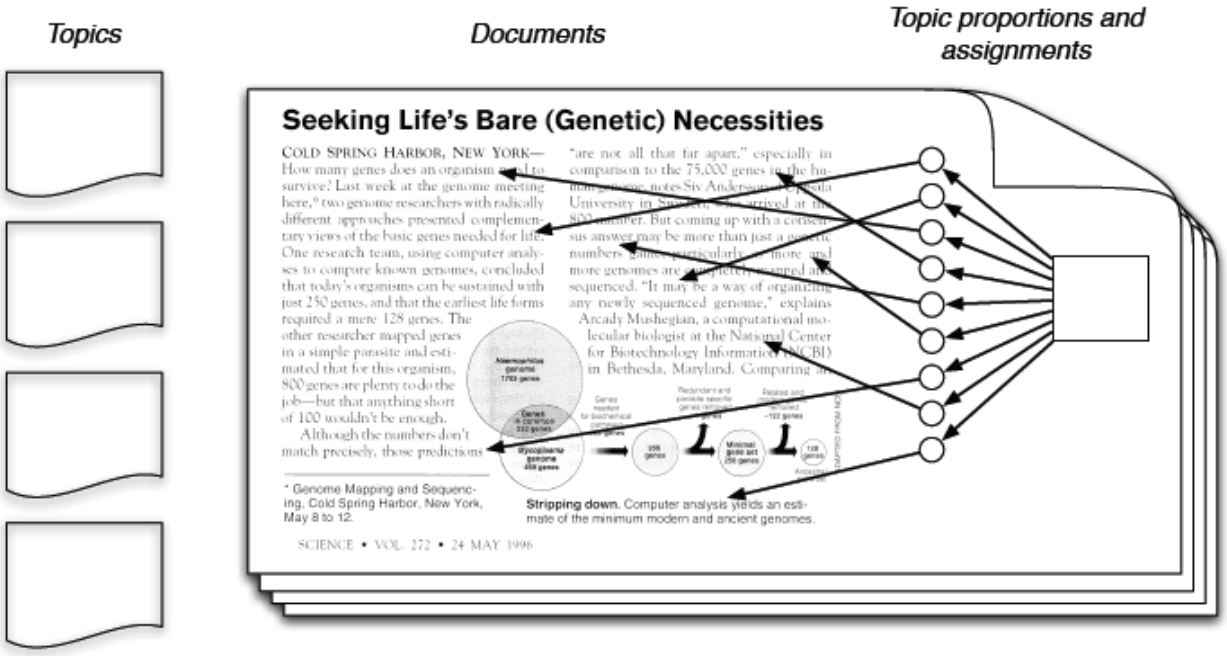
Simple intuition: Documents exhibit multiple topics.

Generative model



- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

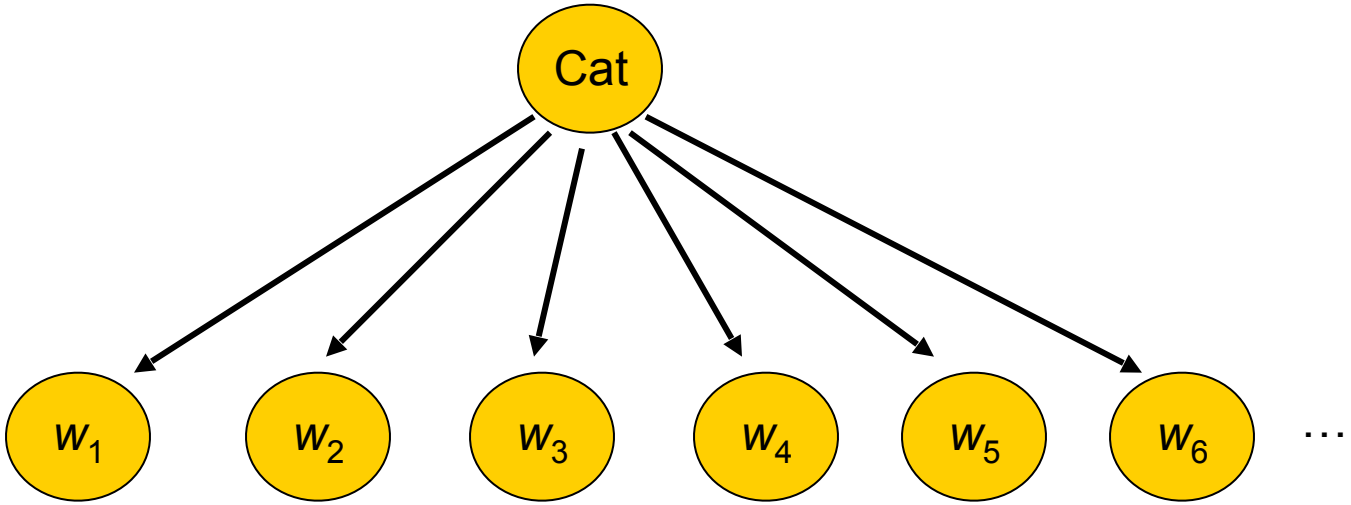
The posterior distribution



- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

Previously

- ***Supervised*** text categorization through Naïve Bayes
- Generative model: first generate a document category, then words in the document (unigram model)



- Inference: obtain posterior over document categories using Bayes rule (argmax to choose the category)

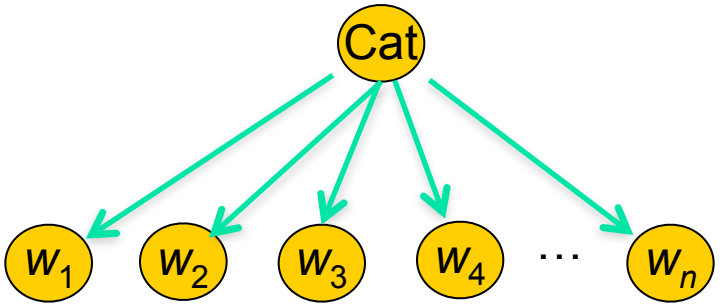
$$P(Cat \mid w_{1\ldots n}) = \frac{P(w_{1\ldots n} \mid Cat)P(Cat)}{P(w_{1\ldots n})}$$

What we're doing here

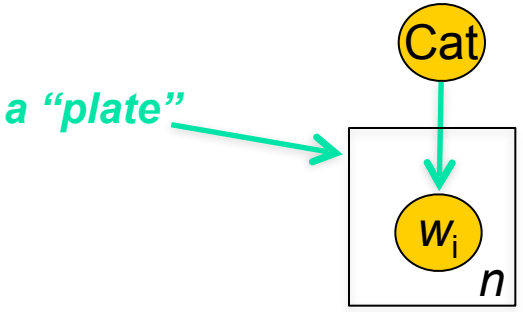
- Supervised categorization requires hand-labeling documents
- This can be extremely time-consuming
- Unlabeled documents are cheap
- So we'd really like to do ***unsupervised*** text categorization
- Now we'll look at unsupervised learning within the Naïve Bayes model

Compact graphical model representations

- We're going to lean heavily on graphical model representations here.

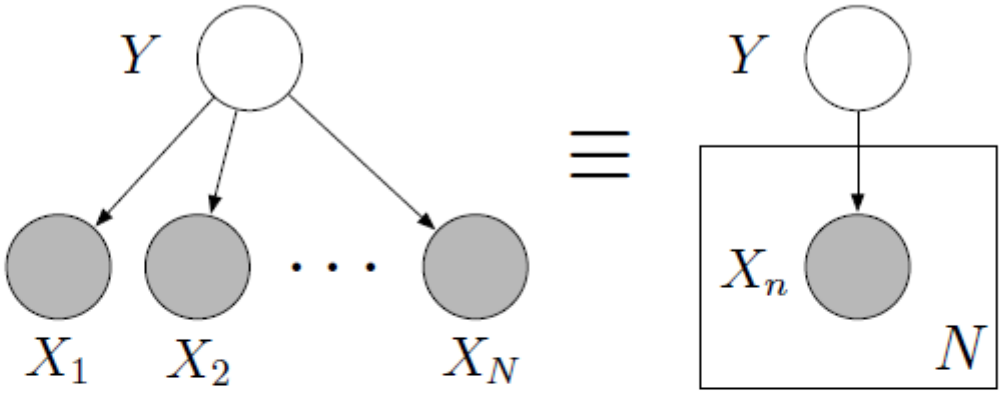


- We'll use a more compact notation:



"generate a word from Cat *n* times"

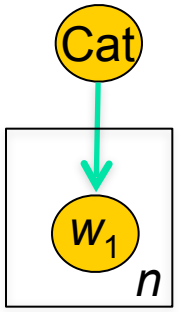
Graphical models (Aside)



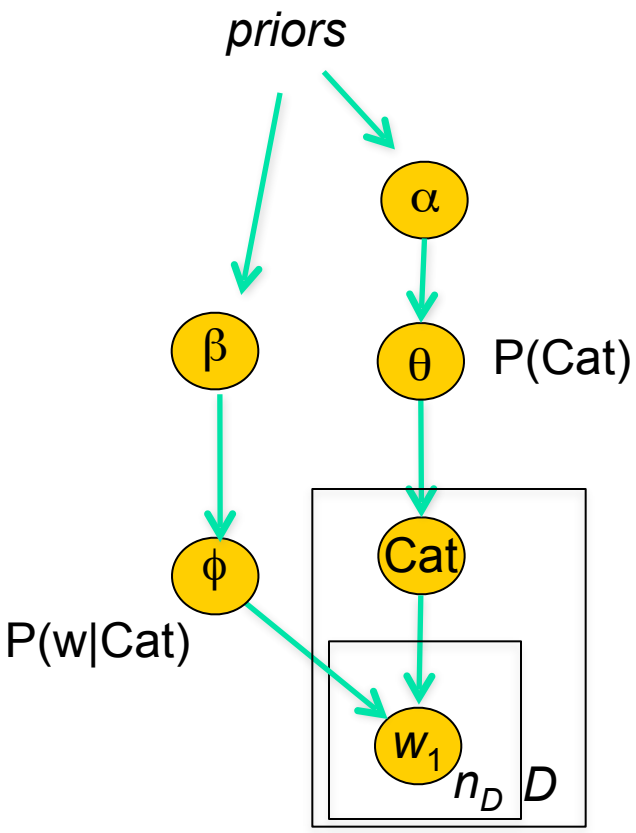
- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure
- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n \mid y)$$

- Now suppose that Cat isn't observed
- We need to learn two distributions:
  - P(Cat)
  - P(w|Cat)
- How do we do this?
  - We might use the method of maximum likelihood (MLE)
  - But it turns out that the likelihood surface is highly non-convex and lots of information isn't contained in a point estimate
  - Alternative: Bayesian methods



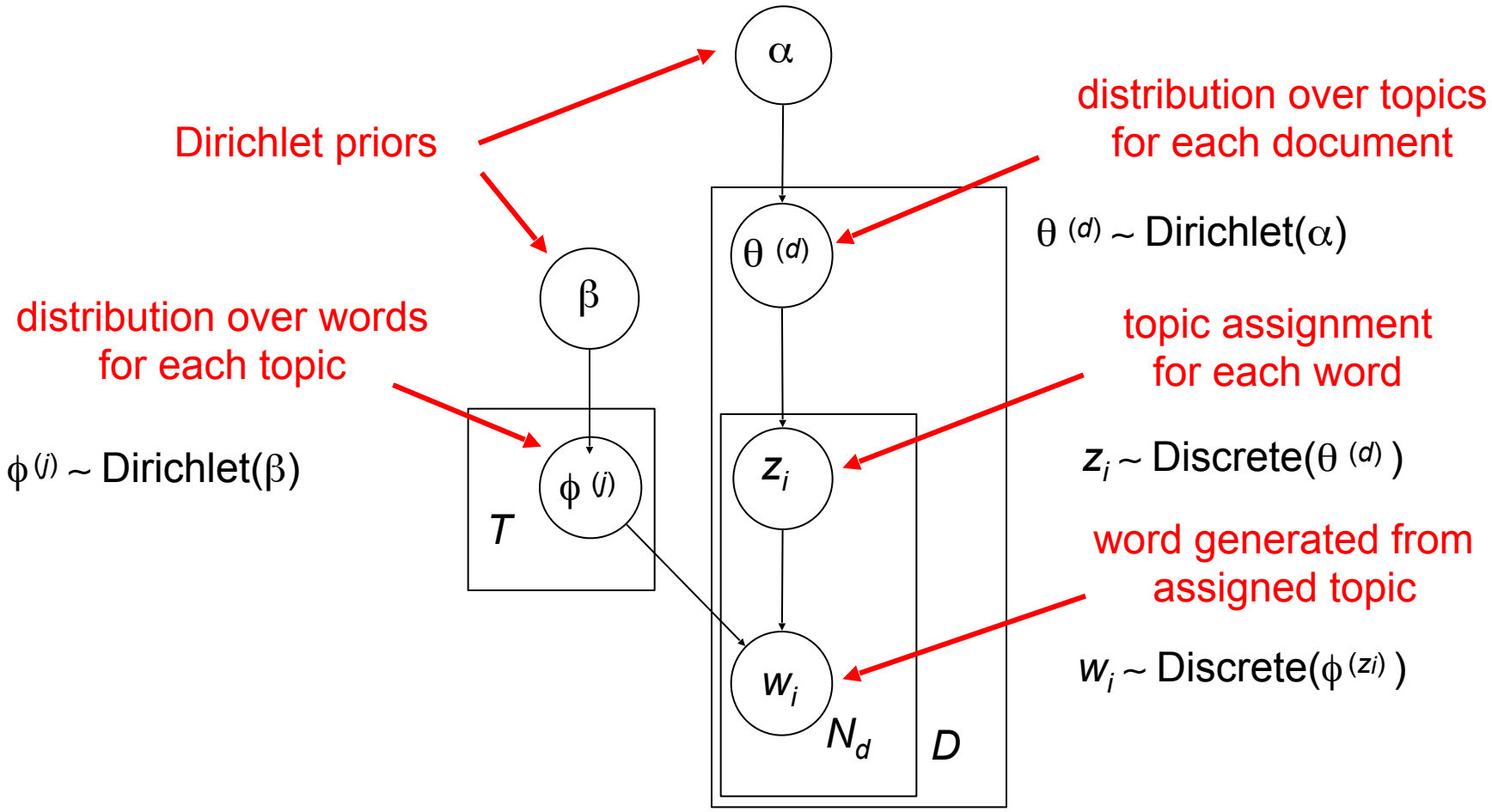
Bayesian document categorization



Latent Dirichlet allocation

(Blei, Ng, & Jordan, 2001; 2003)

Main difference: one topic per word



A generative model for documents

w	*P*(*w*|Cat = 1)
HEART	**0.2**
LOVE	**0.2**
SOUL	**0.2**
TEARS	**0.2**
JOY	**0.2**
SCIENTIFIC	0.0
KNOWLEDGE	0.0
WORK	0.0
RESEARCH	0.0
MATHEMATICS	0.0

topic 1

w	*P*(*w*|Cat = 2)
HEART	0.0
LOVE	0.0
SOUL	0.0
TEARS	0.0
JOY	0.0
SCIENTIFIC	**0.2**
KNOWLEDGE	**0.2**
WORK	**0.2**
RESEARCH	**0.2**
MATHEMATICS	**0.2**

topic 2

Choose mixture weights for each document, generate “bag of words”

$\{P(z = 1), P(z = 2)\}$

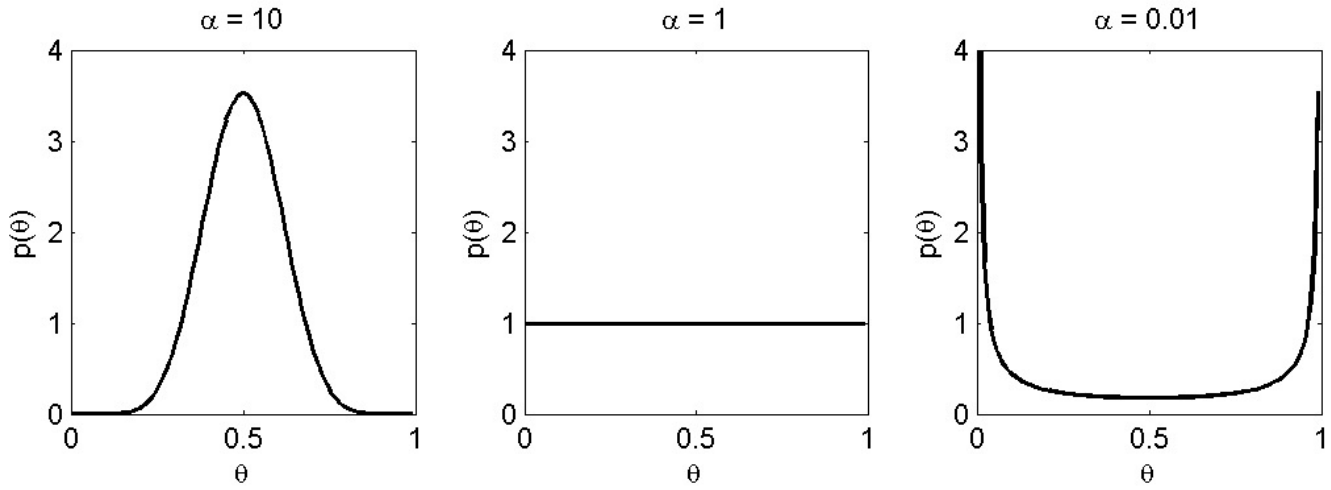
{0, 1}	MATHEMATICS KNOWLEDGE RESEARCH WORK MATHEMATICS RESEARCH WORK SCIENTIFIC MATHEMATICS WORK
{0.25, 0.75}	SCIENTIFIC KNOWLEDGE MATHEMATICS SCIENTIFIC HEART LOVE TEARS KNOWLEDGE HEART
{0.5, 0.5}	MATHEMATICS HEART RESEARCH LOVE MATHEMATICS WORK TEARS SOUL KNOWLEDGE HEART
{0.75, 0.25}	WORK JOY SOUL TEARS MATHEMATICS TEARS LOVE LOVE LOVE SOUL
{1, 0}	TEARS LOVE JOY SOUL LOVE TEARS SOUL SOUL TEARS JOY

Dirichlet priors

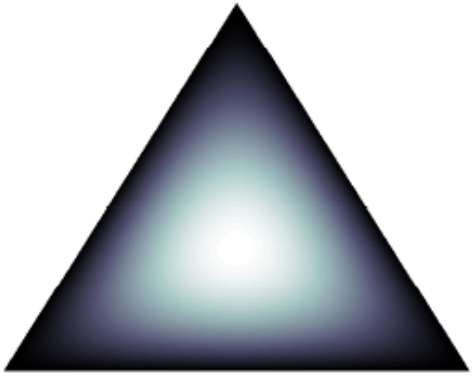
- Multivariate equivalent of Beta distribution

$$p(\theta \mid Cat) = \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \prod_{j=1}^{T} \theta_j^{\alpha-1}$$

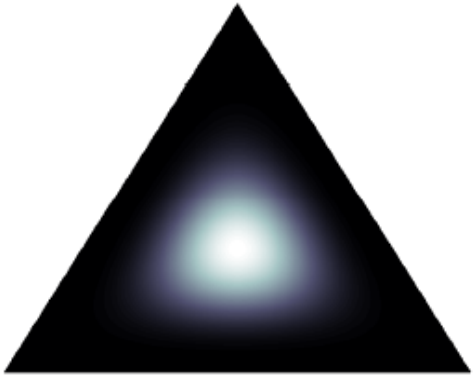
- Hyperparameters α determine form of the prior



Dirichlet Examples



$\alpha = (2, 2, 2)$



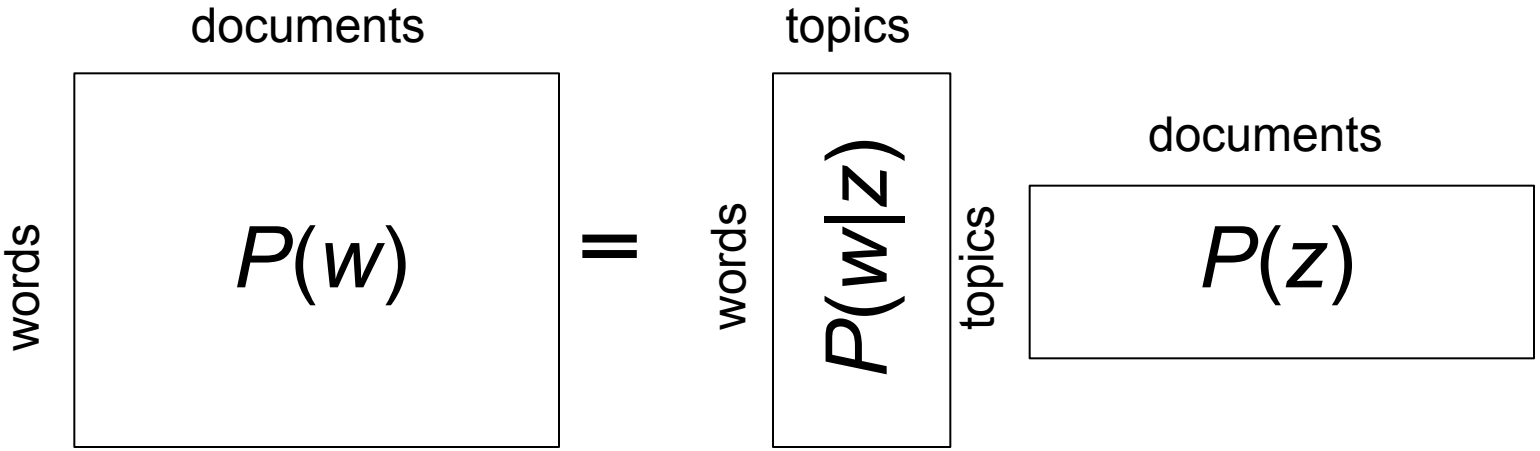
$\alpha = (5, 5, 5)$

$\alpha = (2, 2, 25)$

Darker implies lower magnitude

\alpha < 1 leads to sparser topics

Matrix factorization interpretation



Maximum-likelihood estimation is finding the factorization that minimizes KL divergence

(Hofmann, 1999)

Interpretable topics

DISEASE	WATER	MIND	STORY	FIELD	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	FIELD	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	FIELD	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	FIELD
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

each column shows words from a single topic, ordered by P($w|z$)

Handling multiple senses

DISEASE	WATER	MIND	STORY	**FIELD**	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	**FIELD**	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	**FIELD**	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	**FIELD**
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

each column shows words from a single topic, ordered by P($w|z$)

Explore and browse document collections

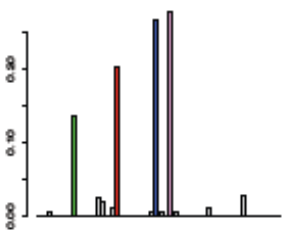
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



Abstract with the most likely topic assignments

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

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

Why does LDA “work” ?

Why does the LDA posterior put “topical” words together?

- Word probabilities are maximized by dividing the words among the topics. (More terms means more mass to be spread around.)
- In a mixture, this is enough to find clusters of co-occurring words.
- In LDA, the Dirichlet on the topic proportions can encourage sparsity, i.e., a document is penalized for using many topics.
- Loosely, this can be thought of as softening the strict definition of “co-occurrence” in a mixture model.
- This flexibility leads to sets of terms that more tightly co-occur.

Inverting the generative model

- Maximum a posteriori estimation (EM)
  - e.g., Hofmann (1999)
- Deterministic approximate algorithms
  - variational EM; Blei, Ng & Jordan (2001; 2003)
  - expectation propagation; Minka & Lafferty (2002)
- Markov chain Monte Carlo
  - full Gibbs sampler; Pritchard et al. (2000)
  - collapsed Gibbs sampler; Griffiths & Steyvers (2004)

The collapsed Gibbs sampler

- Using conjugacy of Dirichlet and multinomial distributions, integrate out continuous parameters

$$P(\mathbf{z}) = \int_{\Delta_T^D} P(\mathbf{z} \mid \Theta) p(\Theta) d\Theta = \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(\alpha)^T} \frac{\Gamma(T\alpha)}{\Gamma(\sum_j n_j^{(d)} + \alpha)}$$

$$P(\mathbf{w} \mid \mathbf{z}) = \int_{\Delta_W^T} P(\mathbf{w} \mid \mathbf{z}, \Phi) p(\Phi) d\Phi = \prod_{j=1}^{T} \frac{\prod_w \Gamma(n_w^{(j)} + \beta)}{\Gamma(\beta)^W} \frac{\Gamma(W\beta)}{\Gamma(\sum_w n_w^{(j)} + \beta)}$$

- Defines a distribution on discrete ensembles **z**

$$P(\mathbf{z} \mid \mathbf{w}) = \frac{P(\mathbf{w} \mid \mathbf{z}) P(\mathbf{z})}{\sum_{\mathbf{z}} P(\mathbf{w} \mid \mathbf{z}) P(\mathbf{z})}$$

The collapsed Gibbs sampler

- Sample each z_i conditioned on $\mathbf{z}_{-i}$

$$P(z_i \mid \mathbf{w}, \mathbf{z}_{-i}) \propto \frac{n_{w_i}^{(z_i)} + \beta}{n_{\bullet}^{(z_i)} + W\beta} \frac{n_j^{(d_i)} + \alpha}{n_{\bullet}^{(d_i)} + T\alpha}$$

- This is nicer than your average Gibbs sampler:
  - memory: counts can be cached in two sparse matrices
  - optimization: no special functions, simple arithmetic
  - the distributions on Φ and Θ are analytic given **z** and **w**, and can later be found for each sample

Gibbs sampling in LDA

			iteration 1
i	w_i	d_i	z_i
1	MATHEMATICS	1	2
2	KNOWLEDGE	1	2
3	RESEARCH	1	1
4	WORK	1	2
5	MATHEMATICS	1	1
6	RESEARCH	1	2
7	WORK	1	2
8	SCIENTIFIC	1	1
9	MATHEMATICS	1	2
10	WORK	1	1
11	SCIENTIFIC	2	1
12	KNOWLEDGE	2	1
.	.	.	.
.	.	.	.
.	.	.	.
50	JOY	5	2

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	?
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	?
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	?
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

Gibbs sampling in LDA

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	2
5	MATHEMATICS	1	1	?
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

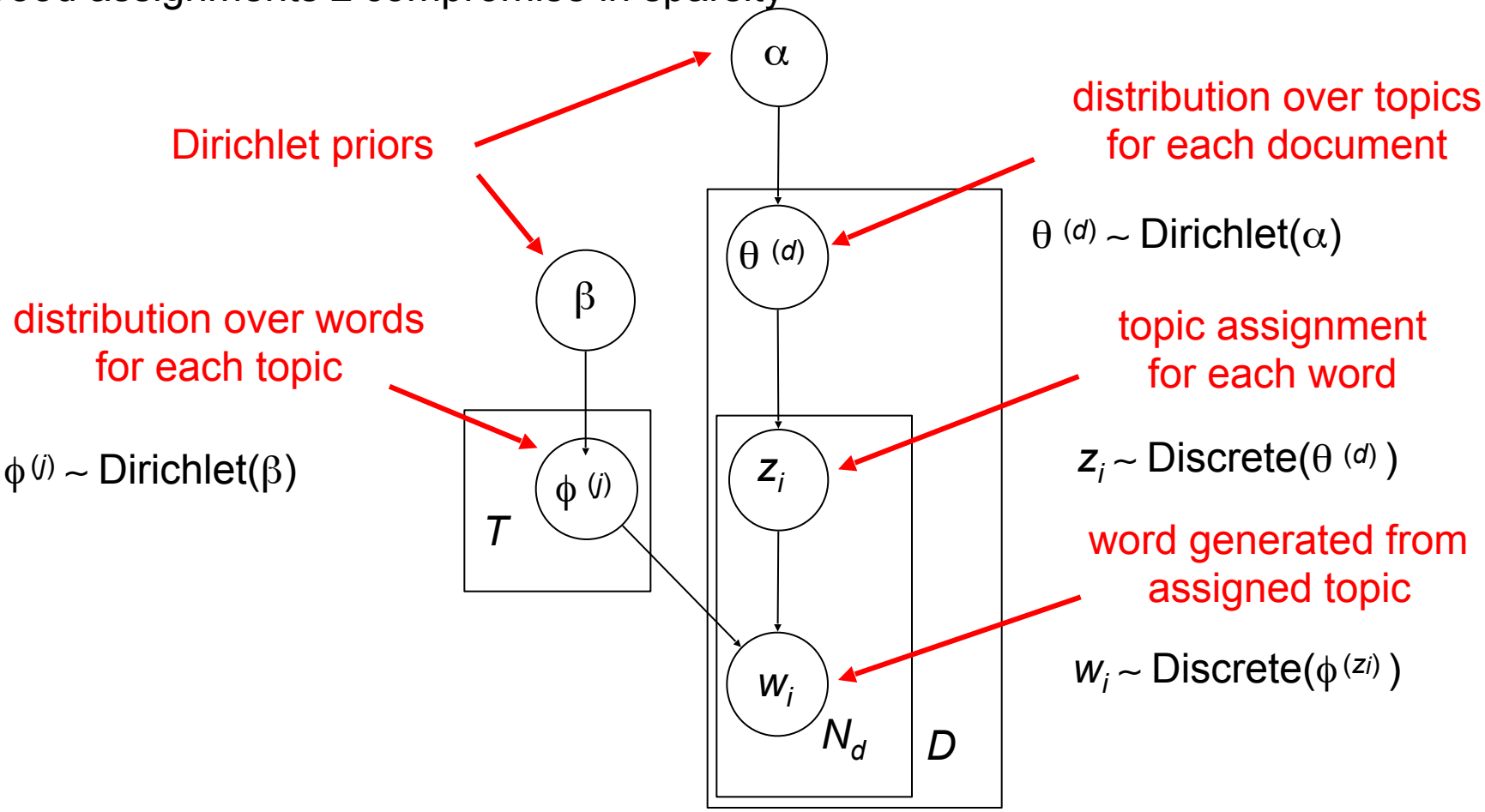
Gibbs sampling in LDA

			iteration 1	2	…	1000
i	w_i	d_i	z_i	z_i		z_i
1	MATHEMATICS	1	2	2		2
2	KNOWLEDGE	1	2	1		2
3	RESEARCH	1	1	1		2
4	WORK	1	2	2		1
5	MATHEMATICS	1	1	2		2
6	RESEARCH	1	2	2		2
7	WORK	1	2	2		2
8	SCIENTIFIC	1	1	1	…	1
9	MATHEMATICS	1	2	2		2
10	WORK	1	1	2		2
11	SCIENTIFIC	2	1	1		2
12	KNOWLEDGE	2	1	2		2
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
50	JOY	5	2	1		1

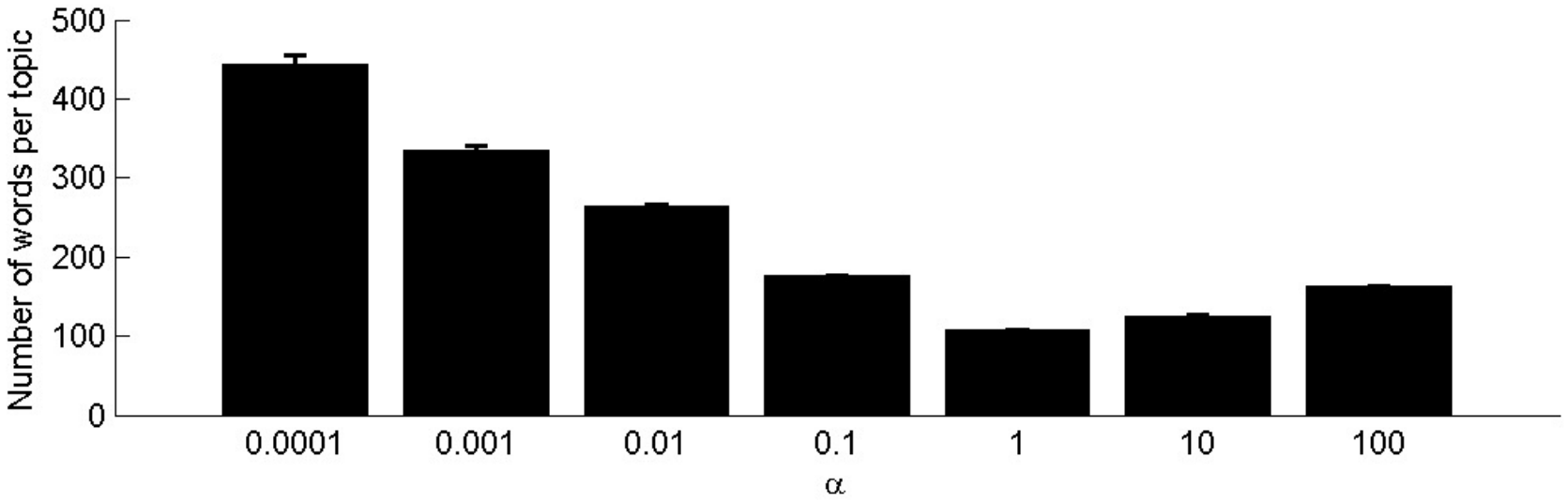
$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

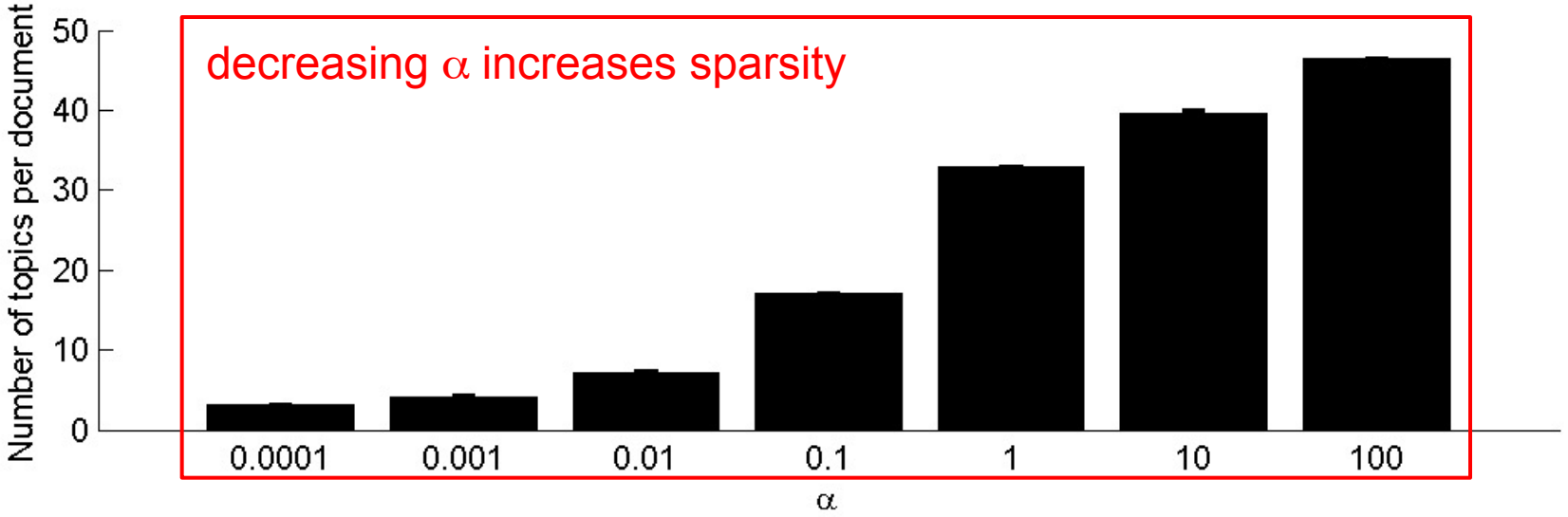
Effects of hyperparameters

- α and β control the relative sparsity of Φ and Θ
  - smaller α, fewer topics per document
  - smaller β, fewer words per topic
- Good assignments **z** compromise in sparsity

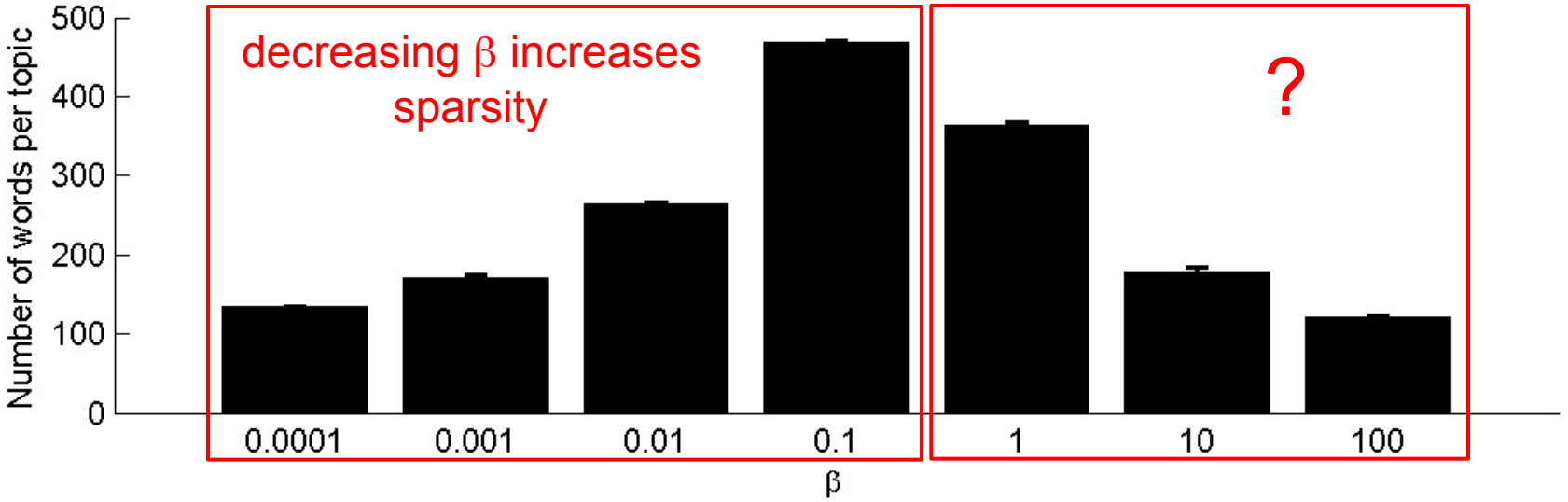


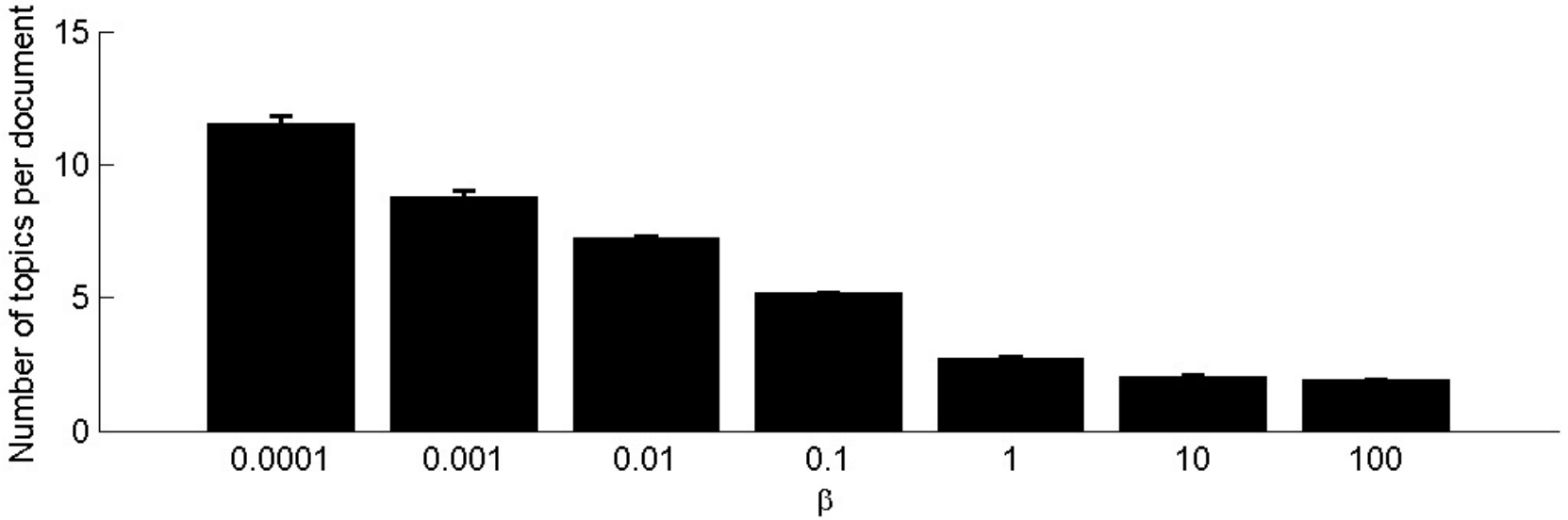
Varying α

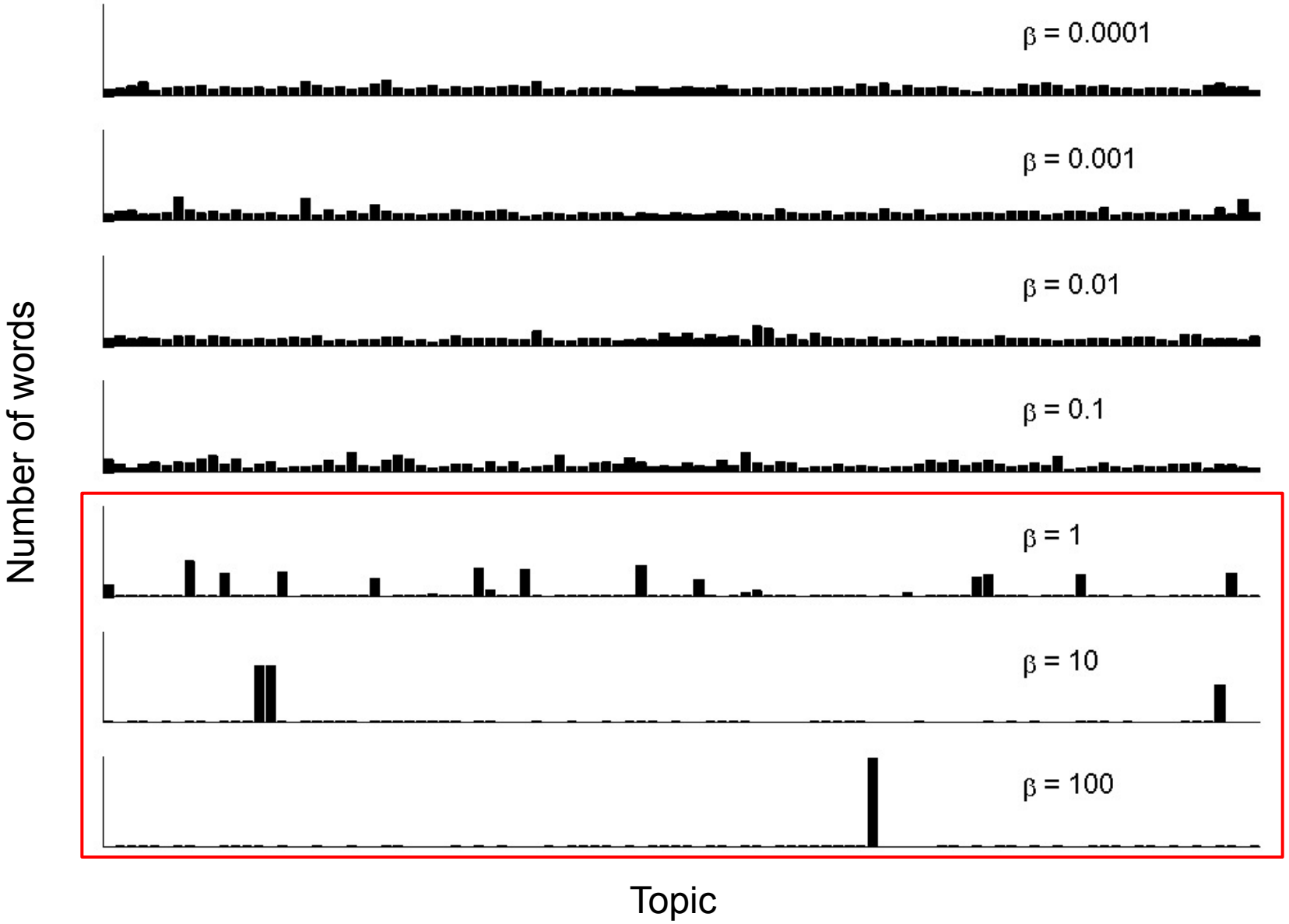




Varying β


Number of words per topic
500
400
300
200
100
0
decreasing β increases sparsity
?
0.0001
0.001
0.01
0.1
1
10
100
β


Number of topics per document
15
10
5
0
0.0001
0.001
0.01
0.1
1
10
100
β


β = 0.0001
β = 0.001
β = 0.01
β = 0.1
β = 1
β = 10
β = 100
Number of words
Topic

Learning the number of topics

- Can use standard Bayes factor methods to evaluate models of different dimensionality
- Alternative: nonparametric Bayes
  - fixed number of topics per document, unbounded number of topics per corpus
    (Blei, Griffiths, Jordan, & Tenenbaum, 2004)
  - unbounded number of topics for both (the hierarchical Dirichlet process)
    (Teh, Jordan, Beal, & Blei, 2004)

Analysis of PNAS abstracts

- Test topic models with a real database of scientific papers from PNAS

- All 28,154 abstracts from 1991-2001
- All words occurring in at least five abstracts, not on “stop” list (20,551)
- Total of 3,026,970 tokens in corpus

(Griffiths & Steyvers, 2004)

A selection of topics

FORCE	HIV	MUSCLE	STRUCTURE	NEURONS	TUMOR
SURFACE	VIRUS	CARDIAC	ANGSTROM	BRAIN	CANCER
MOLECULES	INFECTED	HEART	CRYSTAL	CORTEX	TUMORS
SOLUTION	IMMUNODEFICIENCY	SKELETAL	RESIDUES	CORTICAL	HUMAN
SURFACES	CD4	MYOCYTES	STRUCTURES	OLFACTORY	CELLS
MICROSCOPY	INFECTION	VENTRICULAR	STRUCTURAL	NUCLEUS	BREAST
WATER	HUMAN	MUSCLES	RESOLUTION	NEURONAL	MELANOMA
FORCES	VIRAL	SMOOTH	HELIX	LAYER	GROWTH
PARTICLES	TAT	HYPERTROPHY	THREE	RAT	CARCINOMA
STRENGTH	GP120	DYSTROPHIN	HELICES	NUCLEI	PROSTATE
POLYMER	REPLICATION	HEARTS	DETERMINED	CEREBELLUM	NORMAL
IONIC	TYPE	CONTRACTION	RAY	CEREBELLAR	CELL
ATOMIC	ENVELOPE	FIBERS	CONFORMATION	LATERAL	METASTATIC
AQUEOUS	AIDS	FUNCTION	HELICAL	CEREBRAL	MALIGNANT
MOLECULAR	REV	TISSUE	HYDROPHOBIC	LAYERS	LUNG
PROPERTIES	BLOOD	RAT	SIDE	GRANULE	CANCERS
LIQUID	CCR5	MYOCARDIAL	DIMENSIONAL	LABELED	MICE
SOLUTIONS	INDIVIDUALS	ISOLATED	INTERACTIONS	HIPPOCAMPUS	NUDE
BEADS	ENV	MYOD	MOLECULE	AREAS	PRIMARY
MECHANICAL	PERIPHERAL	FAILURE	SURFACE	THALAMIC	OVARIAN

Software

- MALLET (java)
- in R: *topicmodels* and *lda* packages
- lda (python)
- LDAvis (R)
- ... (lots more!)

Web demo

http://cpsievert.github.io/LDAvis/reviews/reviews.html