CSCI567 Machine Learning (Fall 2017)

Prof. Fei Sha

U of Southern California

Lecture on Nov. 9, 2017

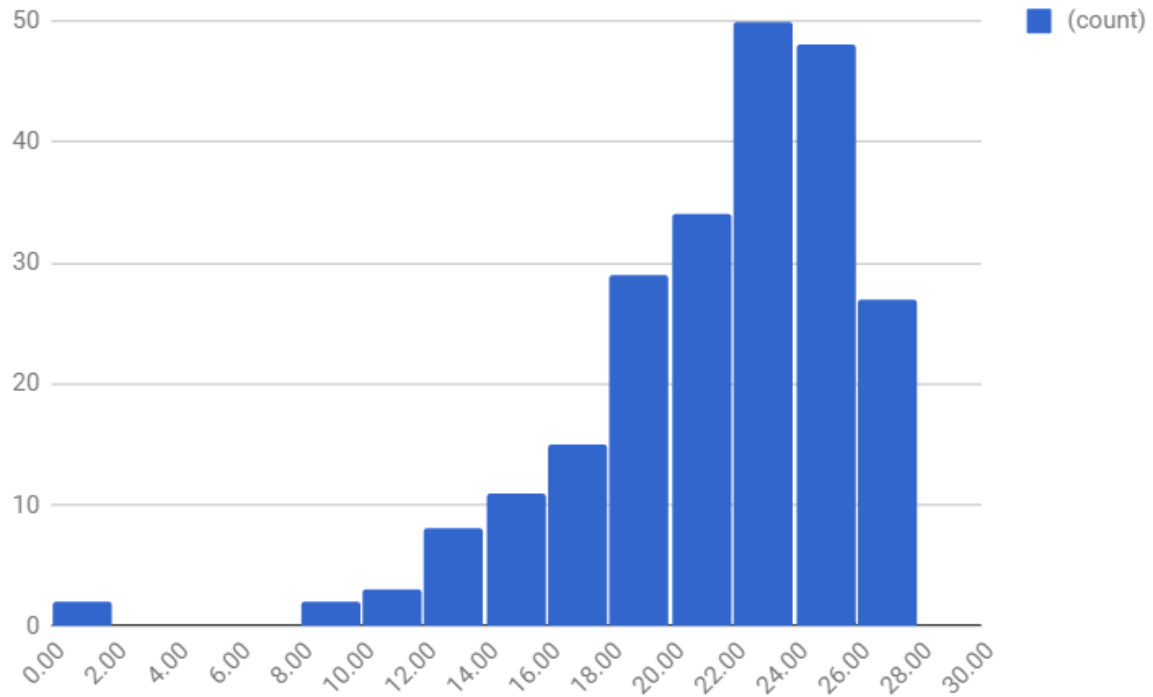
Outline

Outline

Administration

- Nov 14 (next Tuesday): Dr. Brian Milch from Google gives guest lecture.
- Today:
  - I will deliver lecture for LDA
  - Dr. Parisa Mansourifard will deliver lecture on dimensionality reduction

Quiz 2



- Mean: 21.4 (out of 27)
- Median: 22

General impression: the class is on track

Outline

Parameter estimation for HMM: what you need to know

- What are the parameters of HMM?
- How would you estimate them if complete data is given to you? For discrete HMMs, this amounts to counting and calculating the frequencies of event occurring (ie, the fraction/percentage of things happening)
- To estimate with incomplete data, what do you need to do?
  - High-level idea: EM
  - Concretely: compute the posteriors: *what are the posteriors?* How to compute them? What kind of procedures are involved?

Graphical models

Bayes nets

Probabilistic distribution represented with directed acyclic graphs (DAGs)

Markov networks

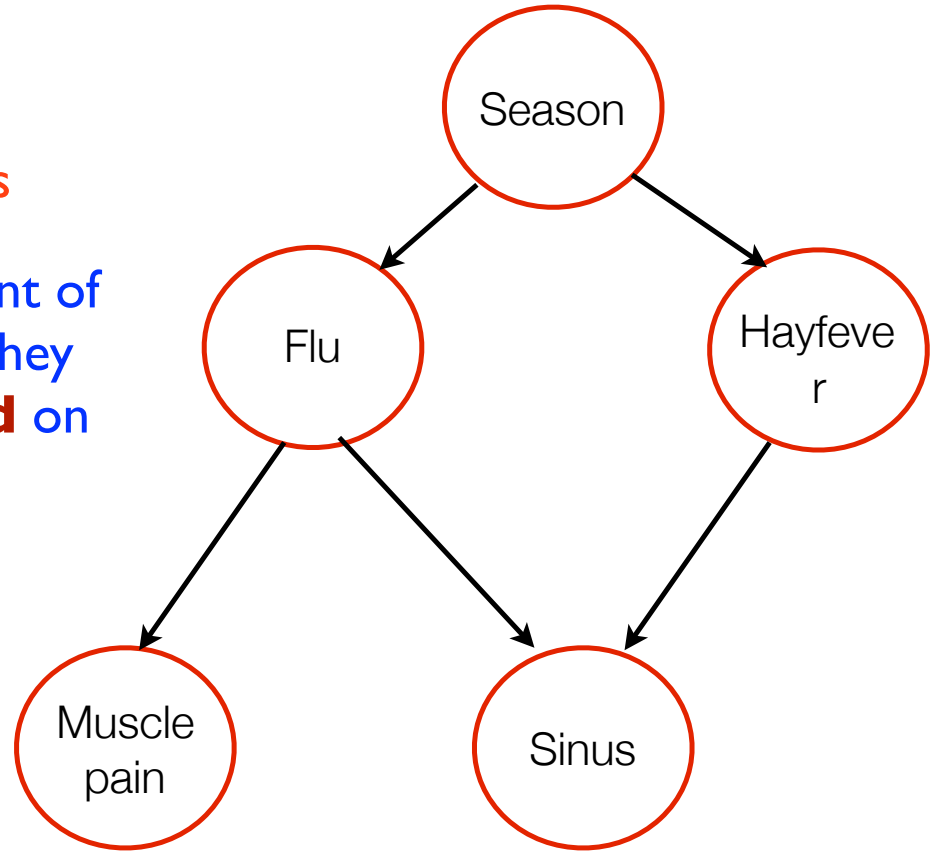
Probabilistic distribution represented with undirected graphs

Exploring structures

Draw links between variables

indicate dependencies, but more importantly, encode independencies

Ex: Flu and Hayfever are independent of each other in any given season; ie, they independently occur **conditioned** on season



This is an example of Bayes networks

Directed acyclic graphs

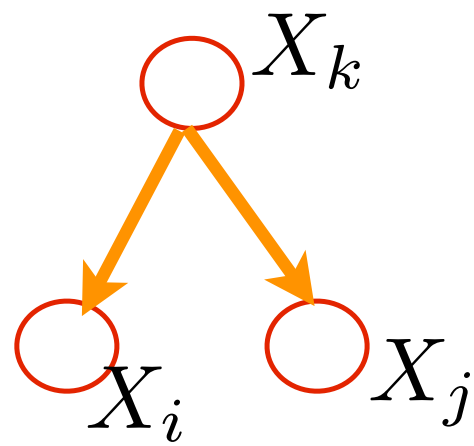
Compact representation of joint distribution

The key concept

conditional independence

$$X_i \perp\!\!\!\perp X_j \mid X_k$$

Representing it graphically



allows us to write

$$\begin{aligned} p(X_i, X_j, X_k) &= p(X_i|X_j, X_k)p(X_j, X_k) \\ &= p(X_i|X_k)p(X_j|X_k)p(X_k) \end{aligned}$$

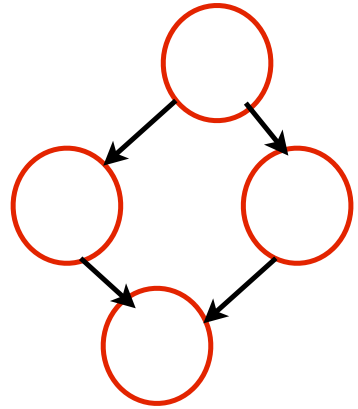
Formal definition of Bayesian networks

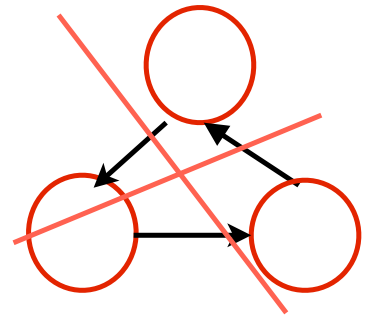
Structure (graph $\mathcal{G}$)

Vertex: random variable

Edge: directed, child vertex depends on parent

No "directed" loop: directed acyclic graph





Conditional probabilities distributions (CPD)

$$P(X_i|\mathbf{Pa}_{X_i})$$

for every vertex

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\mathbf{Pa}_{X_i})$$

also referred as CPT (cond. prob. table) with discrete variables

Models we have seen so far

- Naive Bayes
- Gaussian mixture models
- Hidden Markov models

Challenge: can you draw each above model as Bayesian network? (Note that I have never drawn GMM directly – please do this exercise.) Using the structure of the Bayesian network and the definition, can you write down the joint distribution? Are the same as we had been discussion about those models?

Required readings

- Eugene Charniak's famous "Bayesian Networks without Tears" `https://www.aaai.org/ojs/index.php/aimagazine/article/download/918/836`
- Kevin Murphy's tutorial: `https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html` Please read the "Representation" and "Inference".

Outline

Credits

The following slides are taken from Prof. David Blei's Machine Learning Summer School (2009) tutorial and the presentation by Deepak Santhanam from Prof. Erik Sudderth's Machine Learning Class at Brown University in 2010.

Topic Models

David M. Blei

Department of Computer Science
Princeton University

September 1, 2009

The problem with information

www.betaversion.org/~stefano/linotype/news/26/

As more information becomes available, it becomes more difficult to access what we are looking for.

We need new tools to help us organize, search, and understand these vast amounts of information.

Topic modeling

Candida Hofer

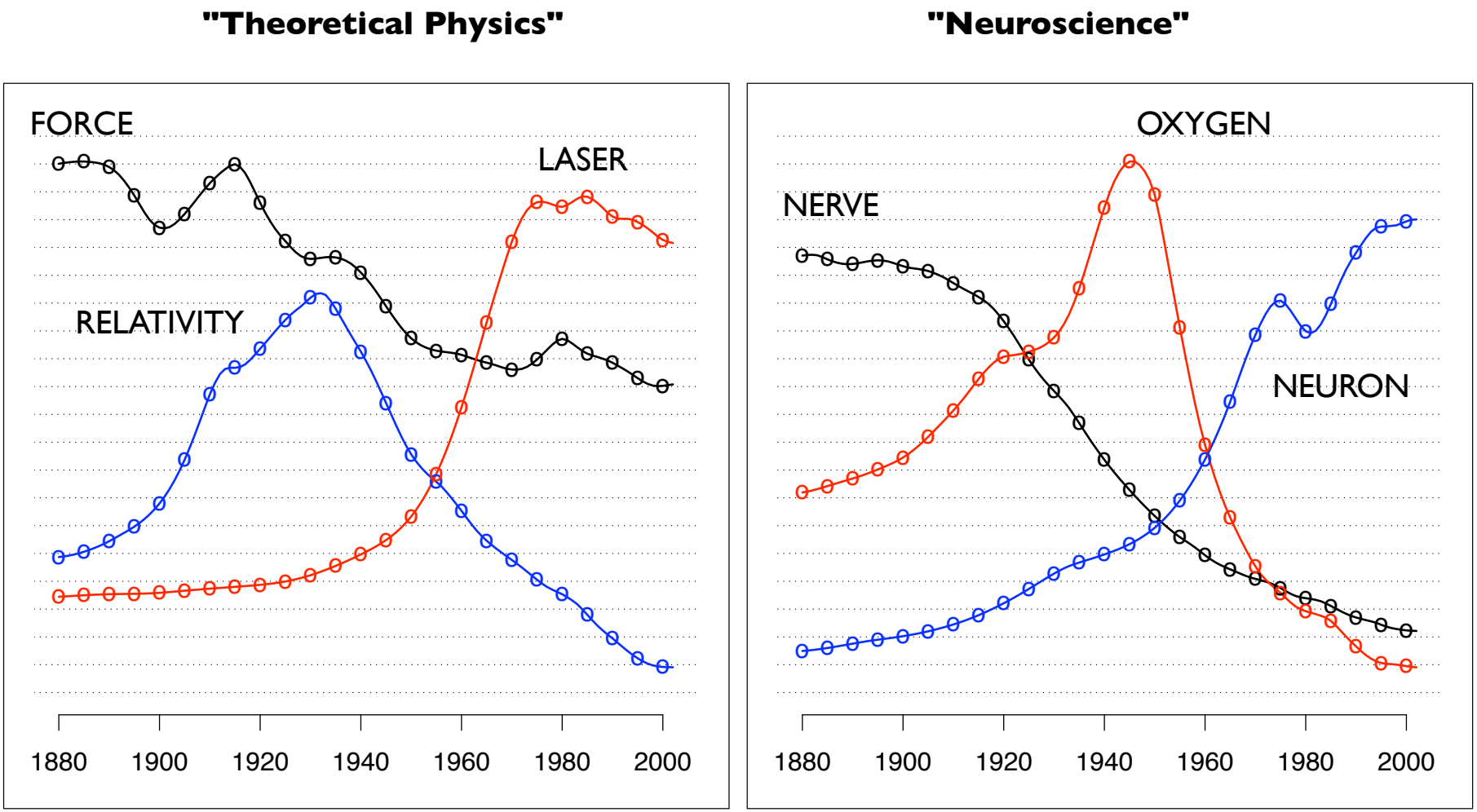
Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

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.

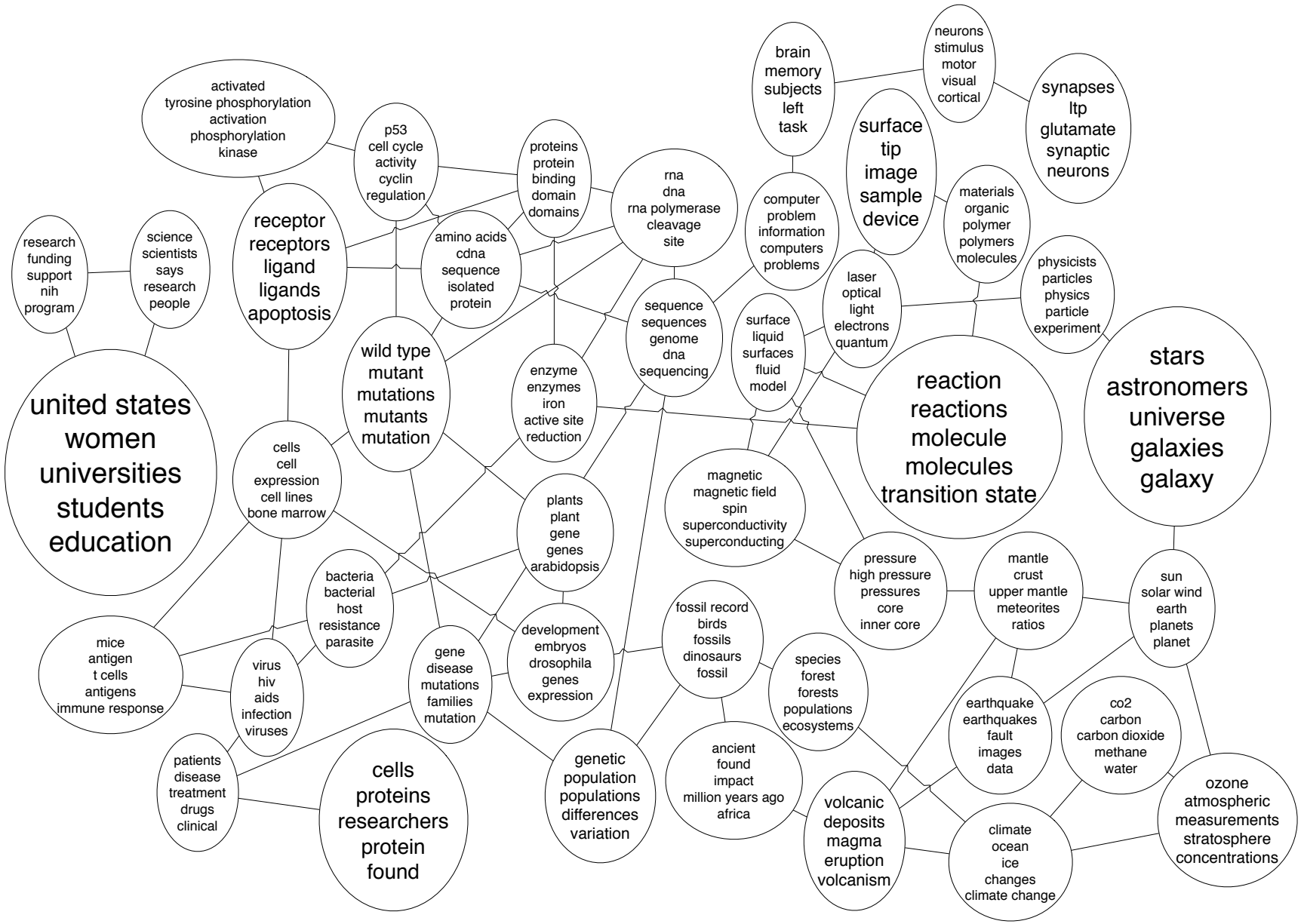
Discover topics from a corpus

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 the evolution of topics over time



Model connections between topics



Annotate images

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

Topic modeling topics

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

Latent Dirichlet allocation (LDA)

1. Introduction to LDA
2. The posterior distribution for LDA

Approximate posterior inference

1. Gibbs sampling
2. Variational inference
3. Comparison/Theory/Advice

Other topic models

1. Topic models for prediction: Relational and supervised topic models
2. The logistic normal: Dynamic and correlated topic models
3. "Infinite" topic models, i.e., the hierarchical Dirichlet process

Interpreting and evaluating topic models

Latent Dirichlet Allocation

Probabilistic modeling

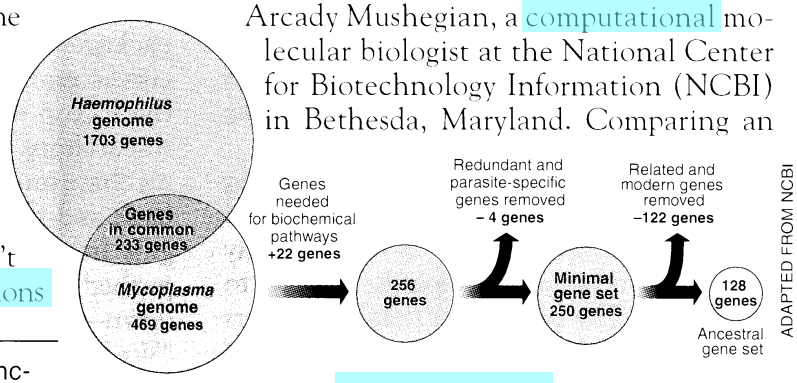
1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables
   - For documents, the hidden variables reflect the thematic structure of the collection.
2. Infer the hidden structure using *posterior inference*
   - What are the topics that describe this collection?
3. Situate new data into the estimated model.
   - How does this query or new document fit into the estimated topic structure?

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



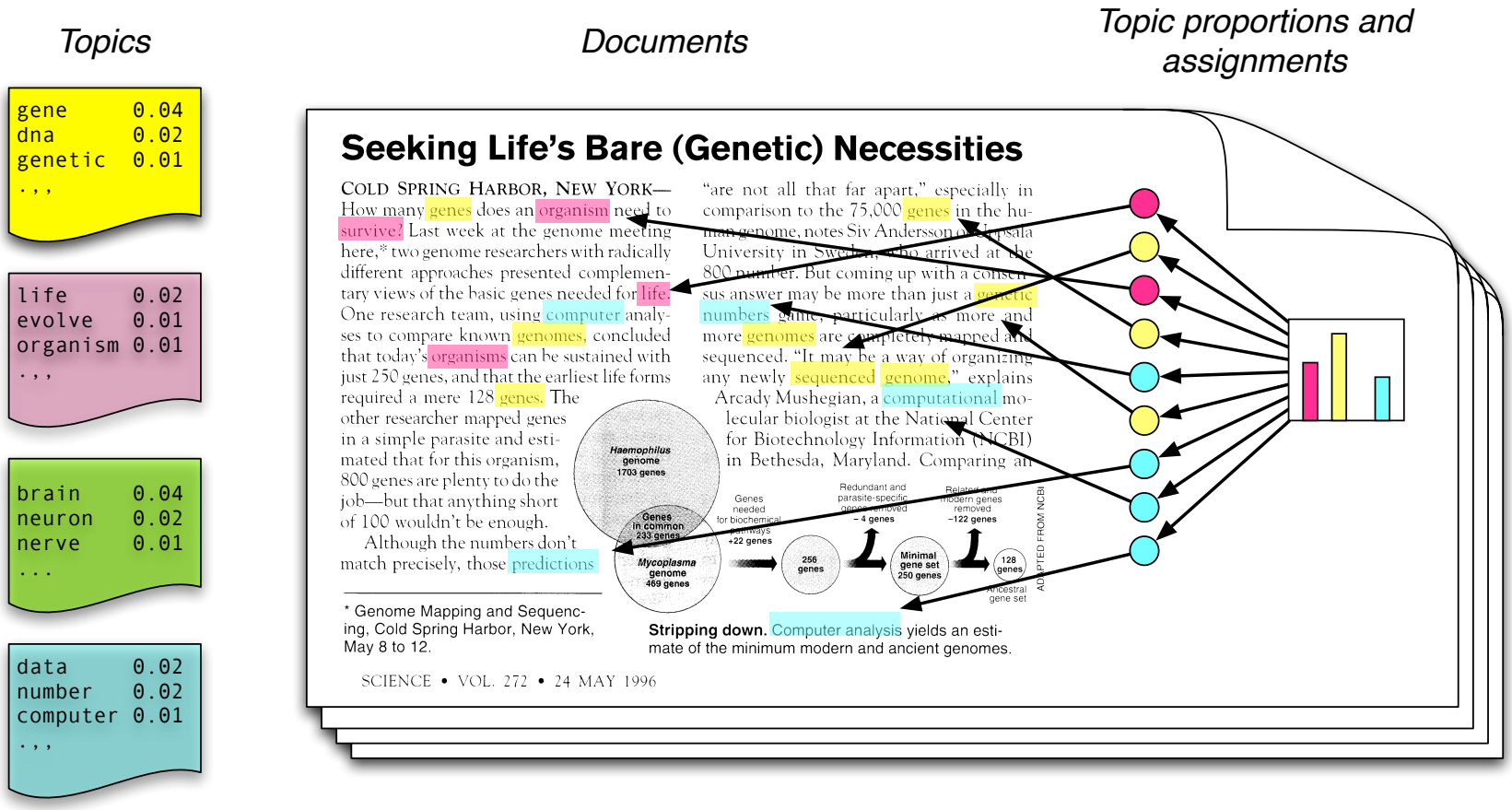
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

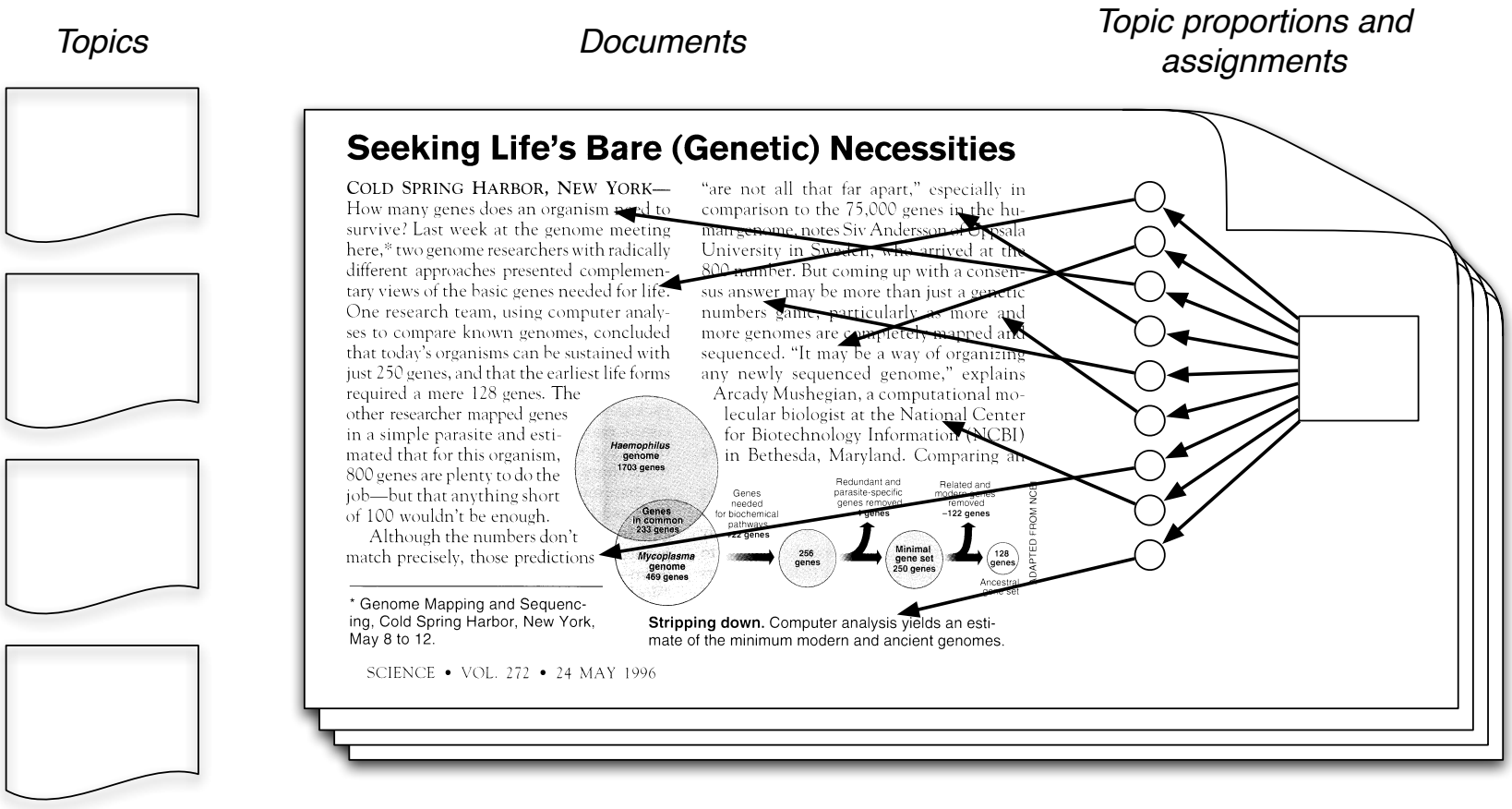
Simple intuition: Documents exhibit multiple topics.

Generative model



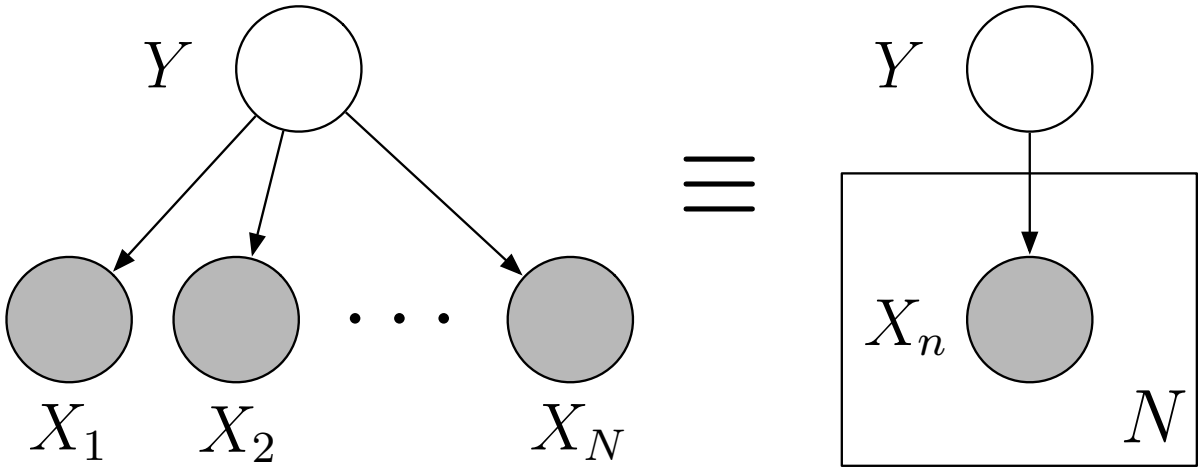
- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

The posterior distribution



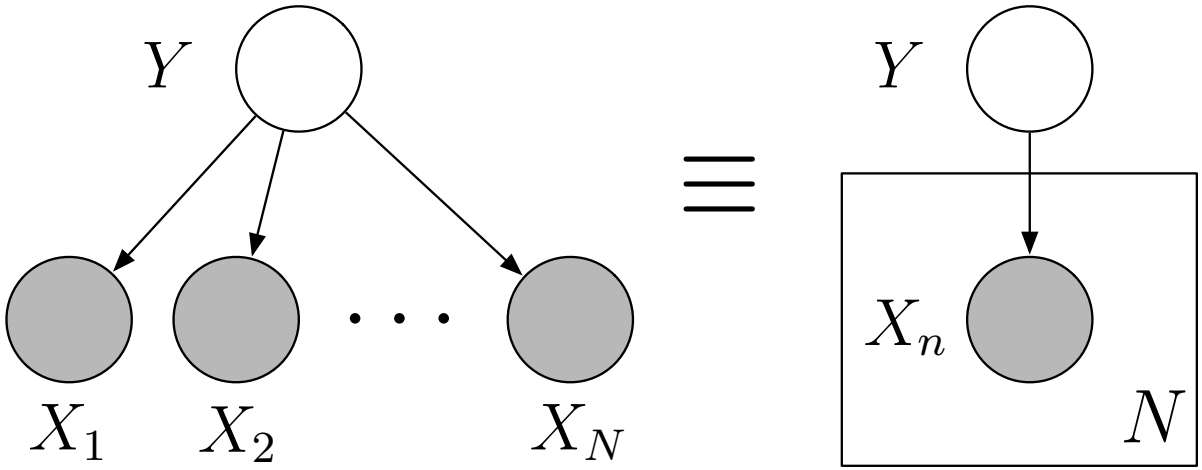
- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

Graphical models (Aside)



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

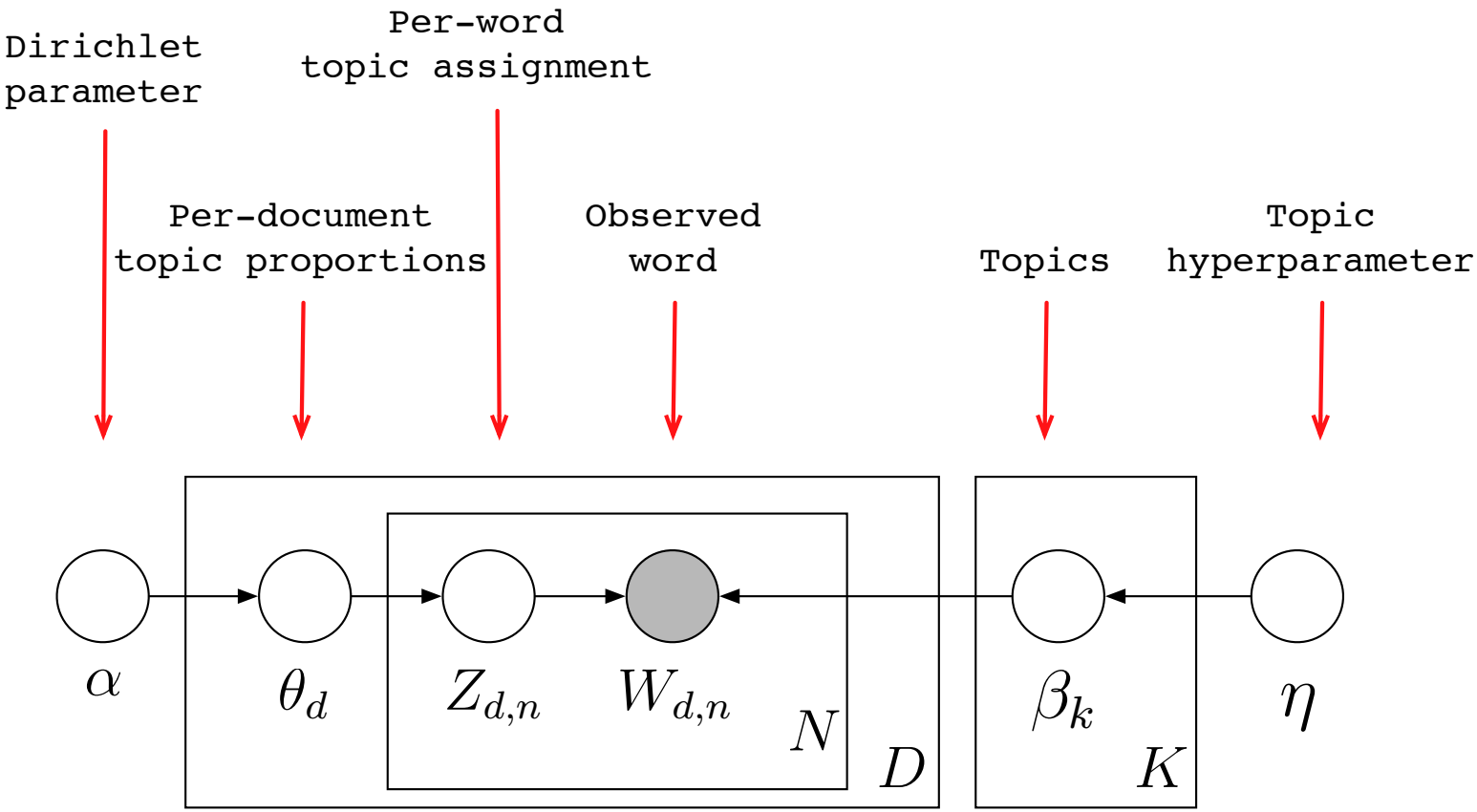
Graphical models (Aside)



- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

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

Latent Dirichlet allocation



Each piece of the structure is a random variable.

The Dirichlet distribution

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

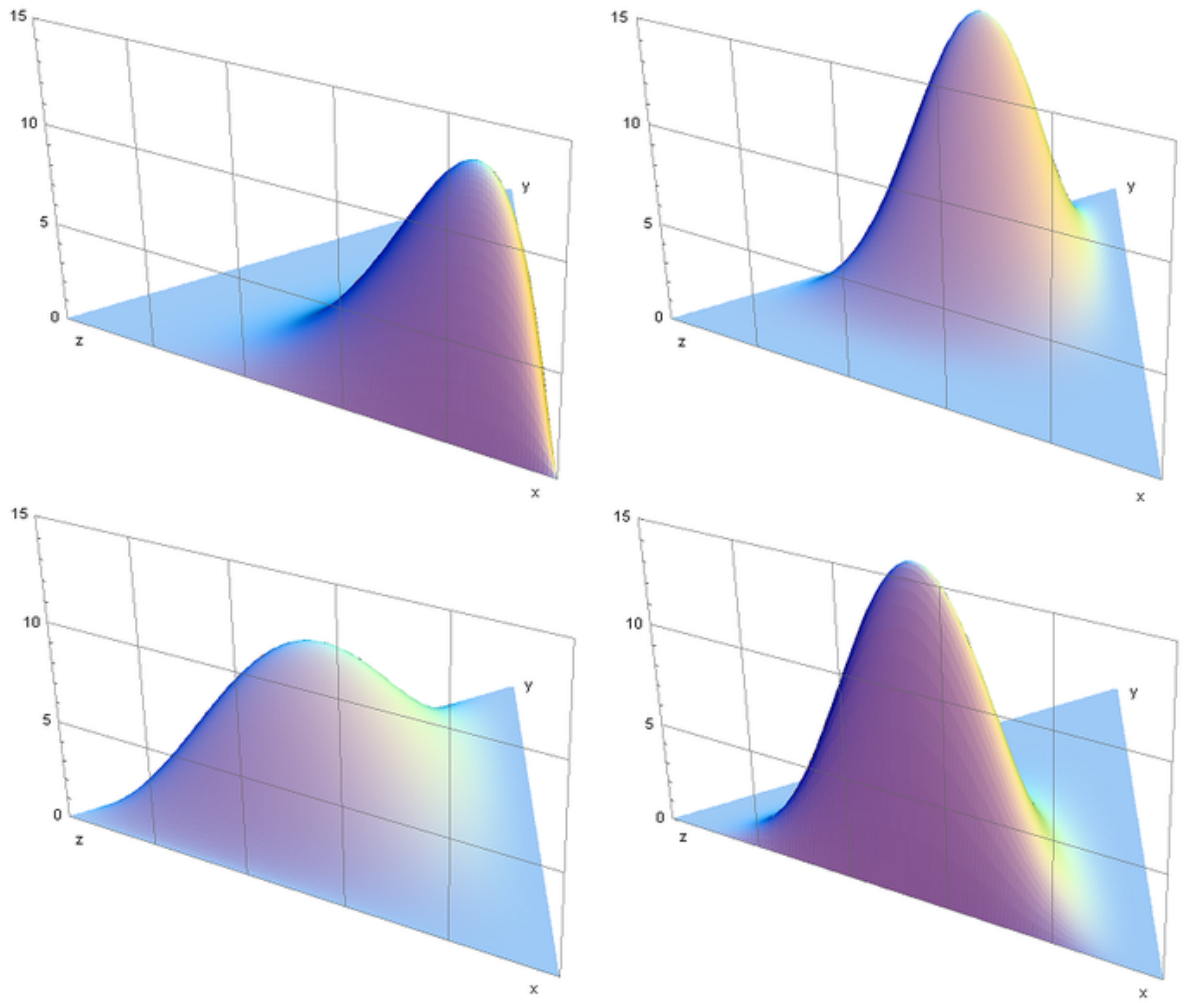
$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

- The Dirichlet 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.

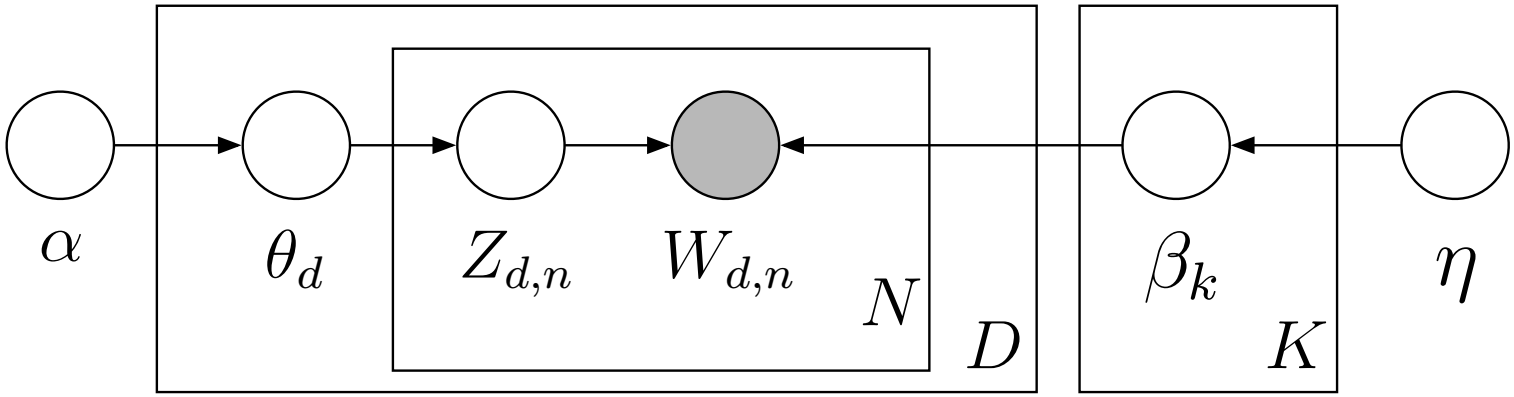
- The topic proportions are a K dimensional Dirichlet.
  The topics are a V dimensional Dirichlet.

The Dirichlet distribution



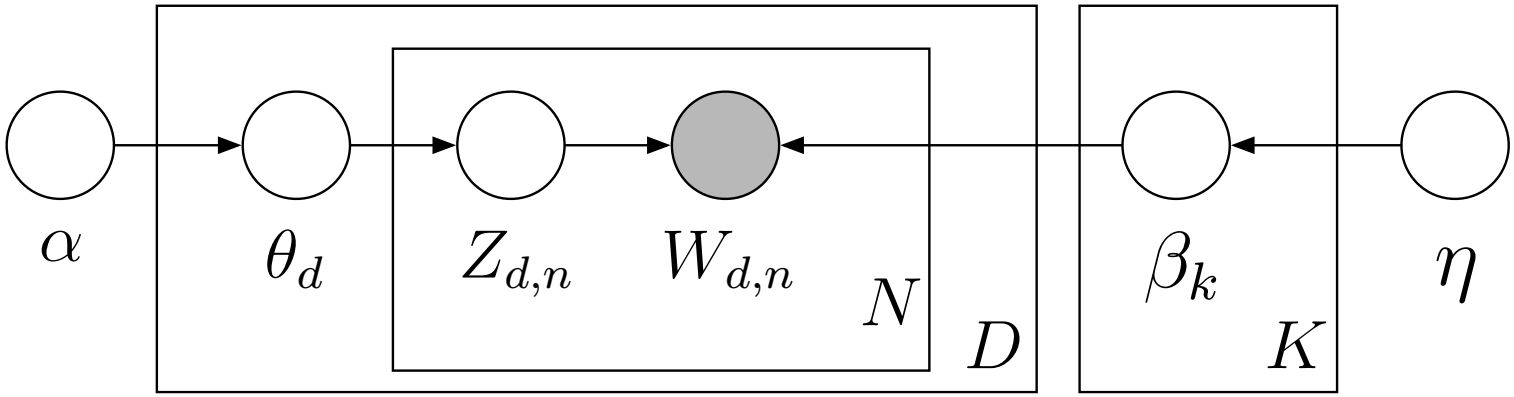
(From Wikipedia)

Latent Dirichlet allocation



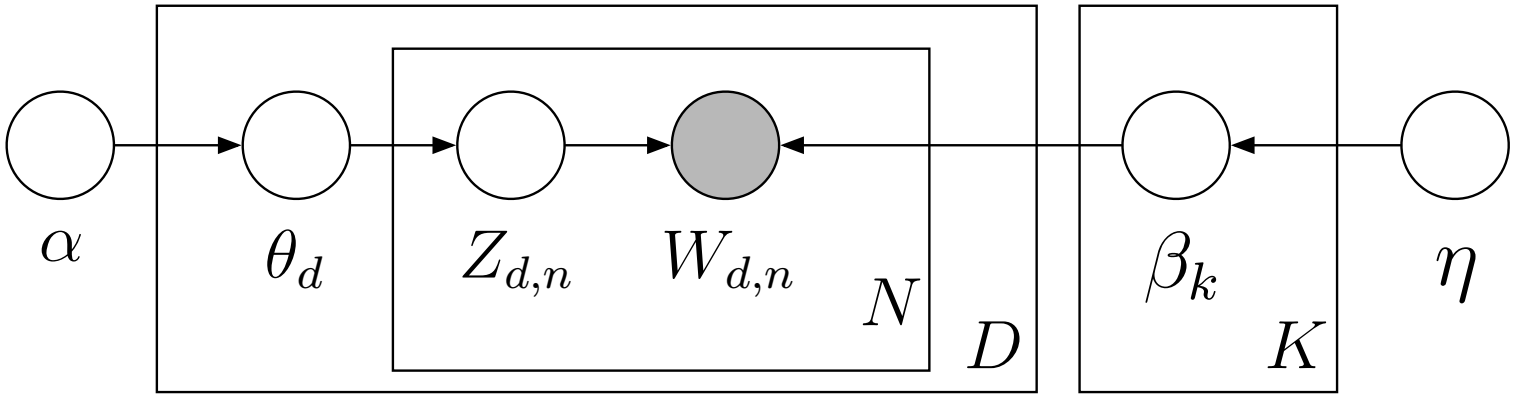
- LDA is a mixed membership model (Erosheva, 2004) that builds on the work of Deerwester et al. (1990) and Hofmann (1999).
- For document collections and other grouped data, this might be more appropriate than a simple finite mixture.
- The same model was independently invented for population genetics analysis (Pritchard et al., 2000).

Latent Dirichlet allocation



- 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
- Use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, etc.

Latent Dirichlet allocation

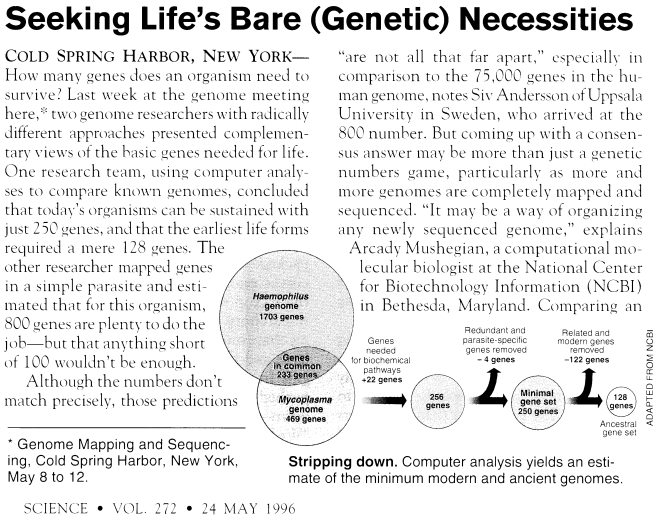


Approximate posterior inference algorithms

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)

For comparison, see Mukherjee and Blei (2009) and Asuncion et al. (2009).

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

* 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

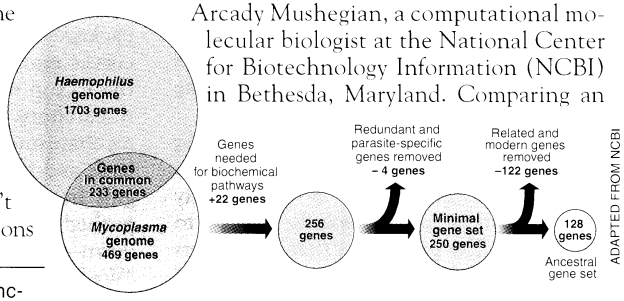
- **Data**: The OCR'ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.

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.

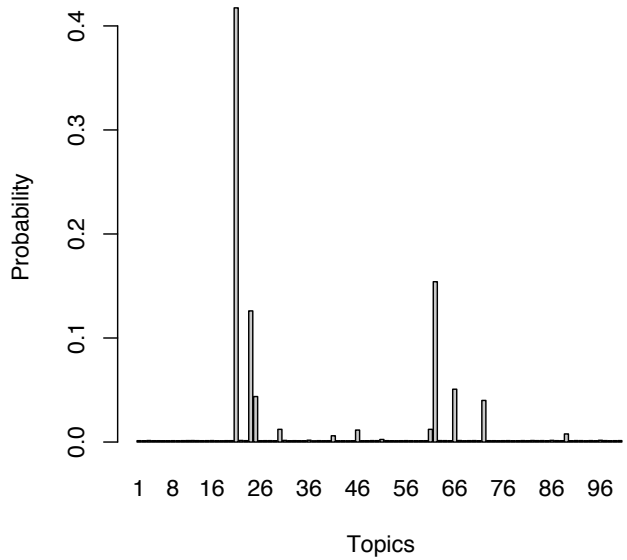
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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



* 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



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

Example inference (II)

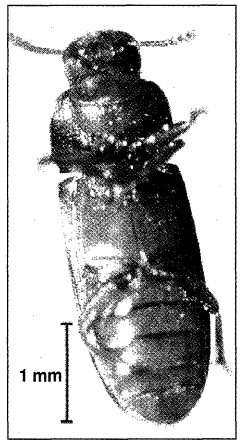
Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al.* (*2*) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (*3*).



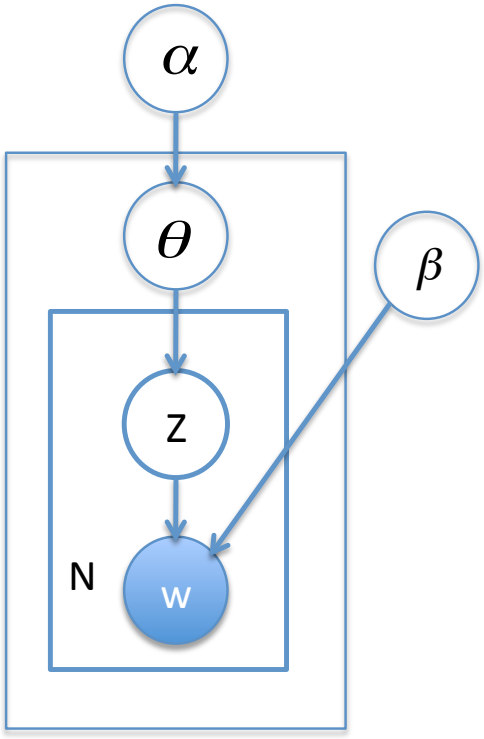
Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

Example inference (II)

problem	model	selection	species
problems	rate	male	forest
mathematical	constant	males	ecology
number	distribution	females	fish
new	time	sex	ecological
mathematics	number	species	conservation
university	size	female	diversity
two	values	evolution	population
first	value	populations	natural
numbers	average	population	ecosystems
work	rates	sexual	populations
time	data	behavior	endangered
mathematicians	density	evolutionary	tropical
chaos	measured	genetic	forests
chaotic	models	reproductive	ecosystem

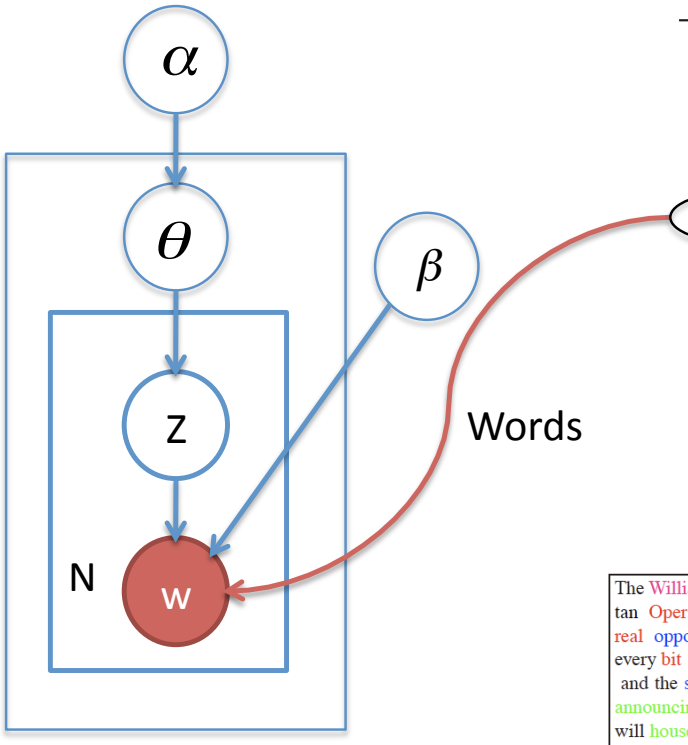
An Early Example..



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

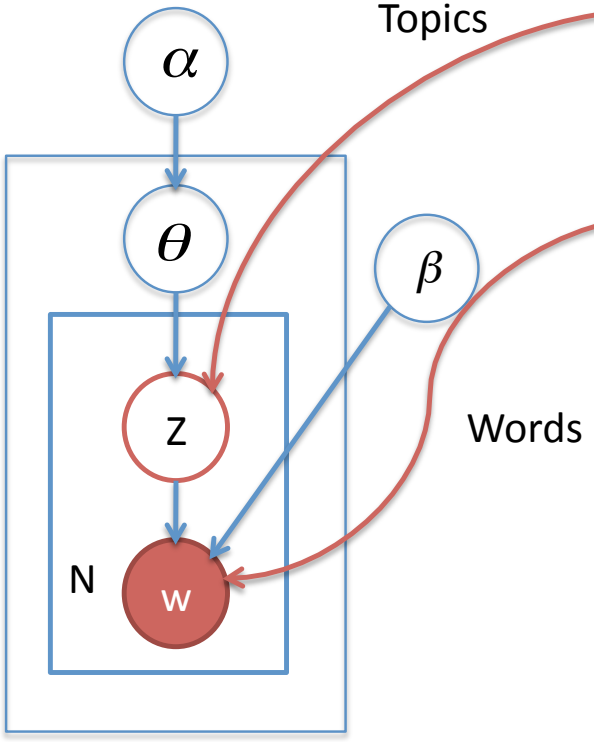
An Early Example..



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

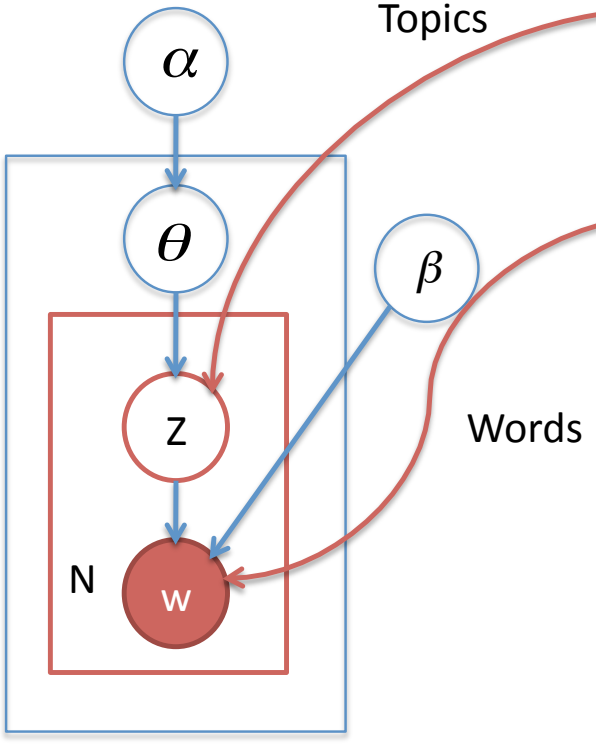
An Early Example..



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

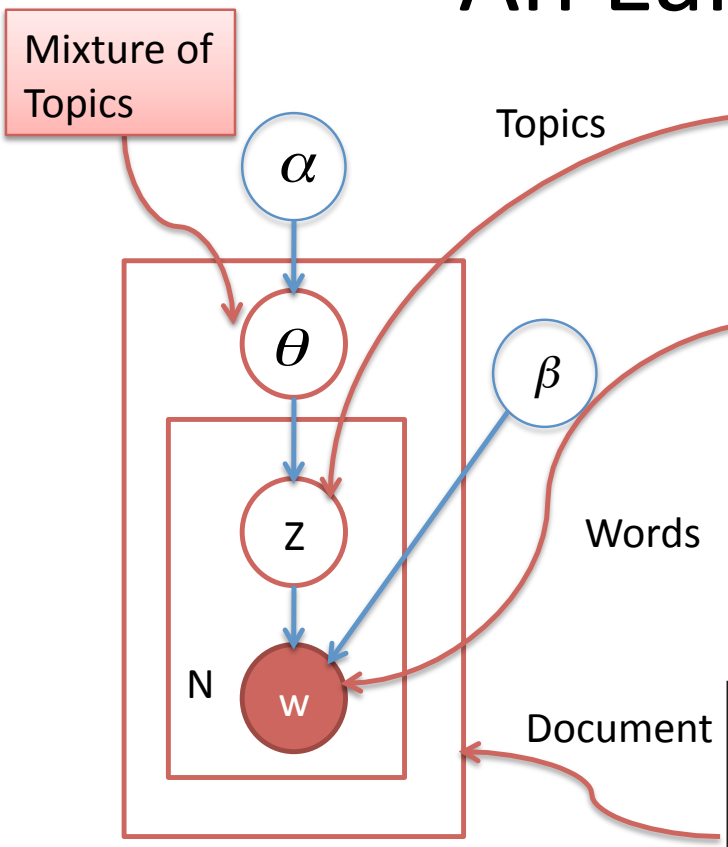
An Early Example..



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

An Early Example..



Topics

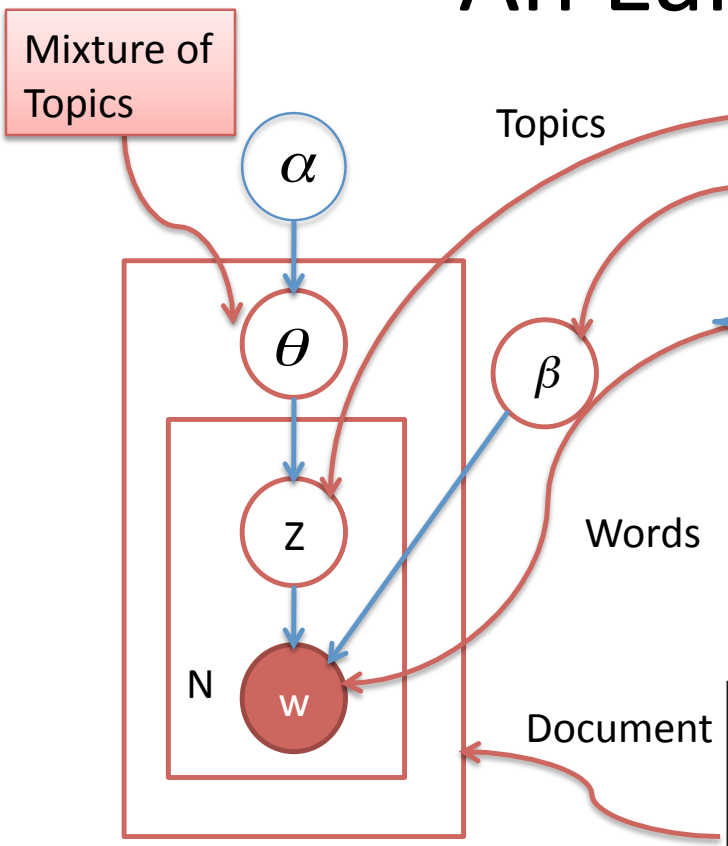
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Words

Document

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

An Early Example..



Topics

Words

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Document

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Learning and Inference in LDA

The most important part is to figure out posterior distribution.

Graphical Model of LDA

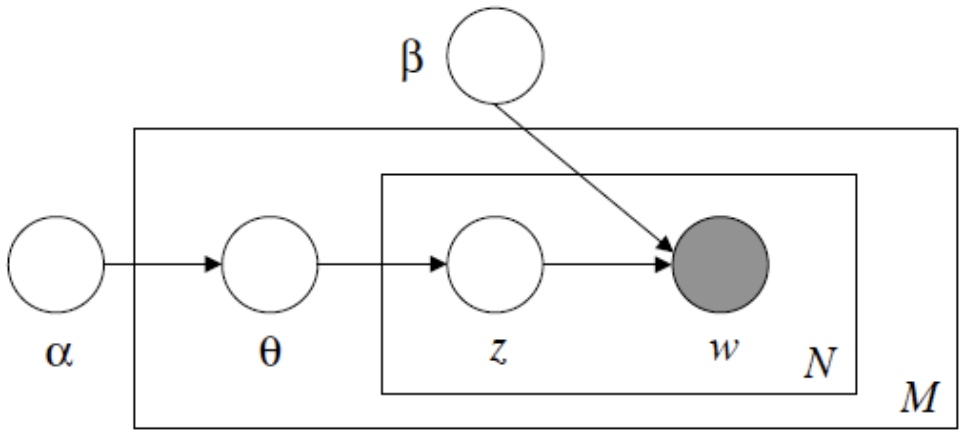
The joint over the topics and words is given by,

Sampled once per corpus

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta),$$

Sampled once every document

Sampled once every word



The Marginal of a Document and The Probability of the Corpus.

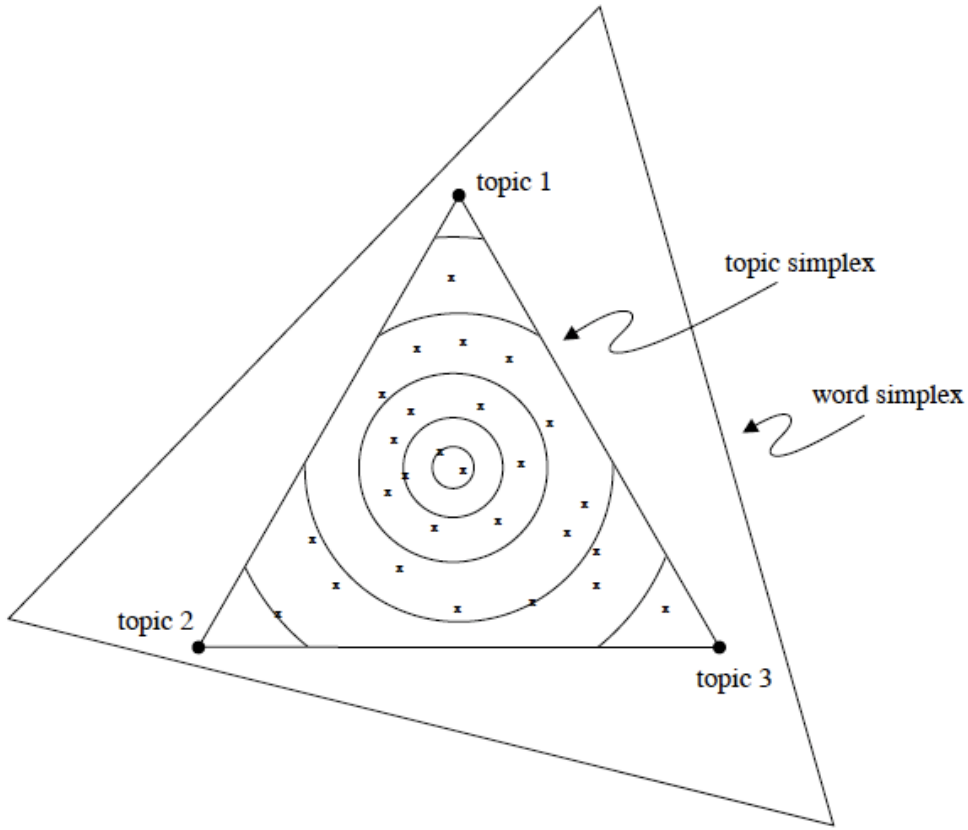
- Integrating over the topic mixtures and summing over the words gives the Marginal of a document.

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta.$$

- Product of the Marginals of all documents gives the probability of the corpus

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d.$$

Geometric Representation



Inference Problem

We have to find the Posterior of the latent variables of a document.

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

Intractable cause we need to marginalize over hidden variables.

$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta,$$

Tight Coupling between two parameters

Use approximate inference like ***MCMC*** or ***variational*** methods.

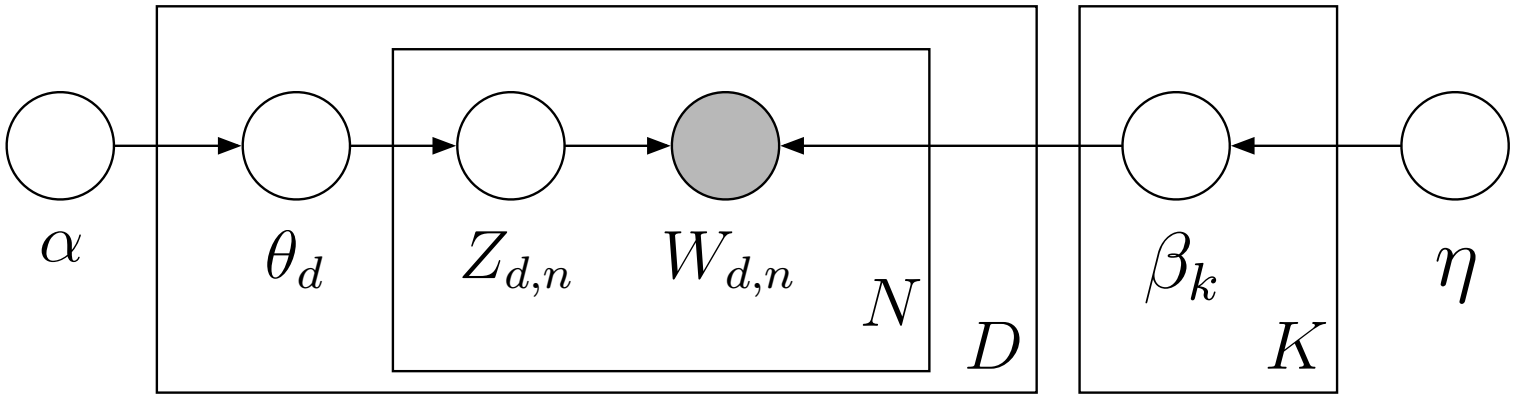
Posterior distribution for LDA

- For now, assume the topics $\beta_{1:K}$ are fixed.
  The per-document posterior is

$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

- This is intractable to compute
- It is a "multiple hypergeometric function" (see Dickey, 1983)
- Can be seen as sum of N^K (tractable) Dirichlet integral terms

Posterior distribution for LDA

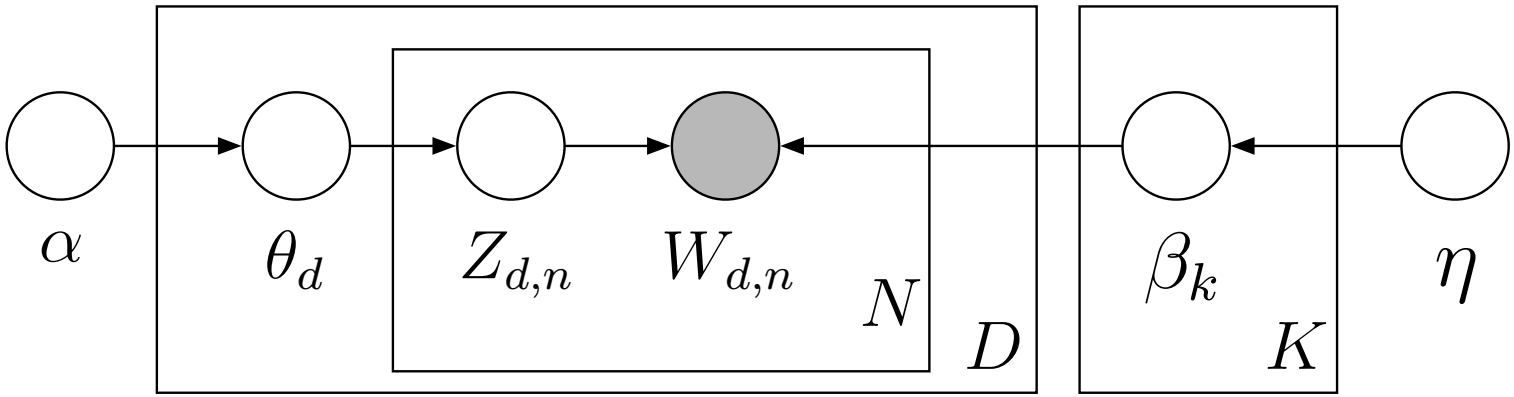


We appeal to approximate posterior inference of the posterior,

$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

- Gibbs sampling
- Variational methods
- Particle filtering

Gibbs sampling for LDA



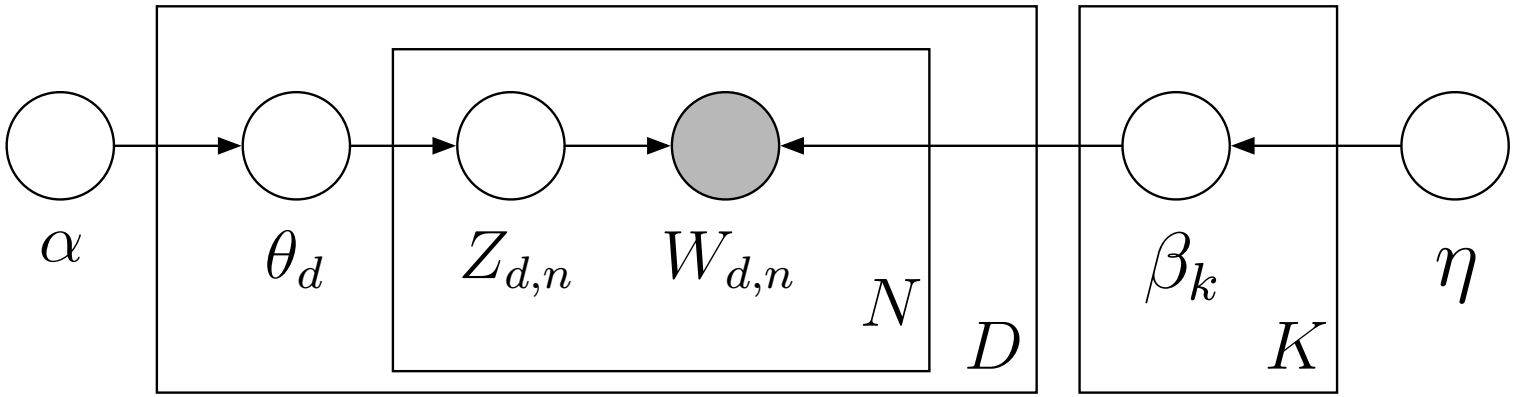
Define $n(z_{1:N})$ to be the counts vector. A simple Gibbs sampler is

$$\begin{aligned} \theta \mid w_{1:N}, z_{1:N} &\sim \mathrm{Dir}(\alpha + n(z_{1:N})) \\ z_i \mid z_{-i}, w_{1:N} &\sim \mathrm{Mult}(\pi(z_{-i}, w_i)) \end{aligned}$$

where

$$\pi(z_{-i}, w_i) \propto (\alpha + n(z_{1:N}))p(w_i \mid \beta_{1:K})$$

Gibbs sampling for LDA



- The topic proportions θ can be integrated out.
- A **collapsed Gibbs sampler** draws from

$$p(z_i \mid z_{-i}, w_{1:N}) \propto p(w_i \mid \beta_{1:K}) \prod_{k=1}^{K} \Gamma(n_k(z_{-i})),$$

  where $n_k(z_{-i})$ is the number of times we've seen topic k in the collection of topic assignments z_{-i}.
- Integrating out variables leads to a faster mixing chain.

Variational inference (in general)

- Variational methods are a deterministic alternative to MCMC.
- Let $x_{1:N}$ be observations and $z_{1:M}$ be latent variables
- Our goal is to compute the posterior distribution

$$p(z_{1:M} \mid x_{1:N}) = \frac{p(z_{1:M}, x_{1:N})}{\int p(z_{1:M}, x_{1:N}) dz_{1:M}}$$

- For many interesting distributions, the marginal likelihood of the observations is difficult to efficiently compute

Variational inference

- Use Jensen's inequality to bound the log prob of the observations:

$$\begin{aligned}
\log p(x_{1:N}) &= \log \int p(z_{1:M}, x_{1:N}) dz_{1:M} \\
&= \log \int p(z_{1:M}, x_{1:N}) \frac{q_\nu(z_{1:M})}{q_\nu(z_{1:M})} dz_{1:M} \\
&\geq \mathrm{E}_{q_\nu}[\log p(z_{1:M}, x_{1:N})] - \mathrm{E}_{q_\nu}[\log q_\nu(z_{1:M})]
\end{aligned}$$

- We have introduced a distribution of the latent variables with free *variational parameters* ν.
- We optimize those parameters to tighten this bound.
- This is the same as finding the member of the family q_ν that is closest in KL divergence to $p(z_{1:M} \mid x_{1:N})$.

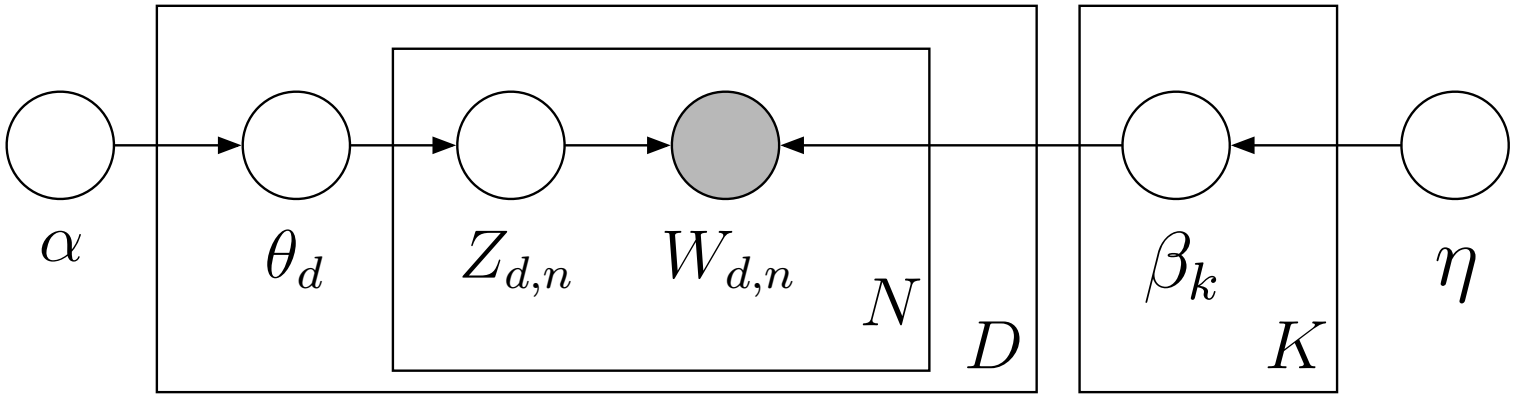
Mean-field variational inference

- Complexity of optimization is determined by the factorization of q_ν
- In *mean field variational inference* we choose q_ν to be fully factored

$$q_\nu(z_{1:M}) = \prod_{m=1}^{M} q_{\nu_m}(z_m).$$

- The latent variables are independent.
  - Each is governed by its own variational parameter ν_m.
- In the true posterior they can exhibit dependence (often, this is what makes exact inference difficult).

Variational Inference for LDA

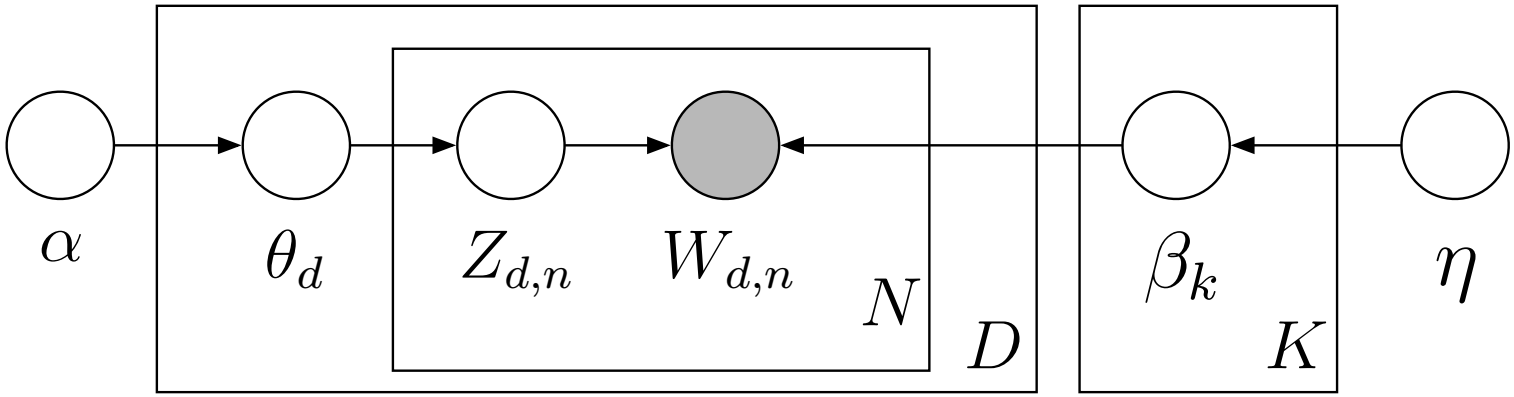


- The mean field variational distribution is

$$q(\theta, z_{1:N} \mid \gamma, \phi_{1:N}) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi)$$

- This is a family of distributions over the latent variables, where all variables are independent and governed by their own parameters.
- In the true posterior, the latent variables are **not** independent.

Variational Inference for LDA

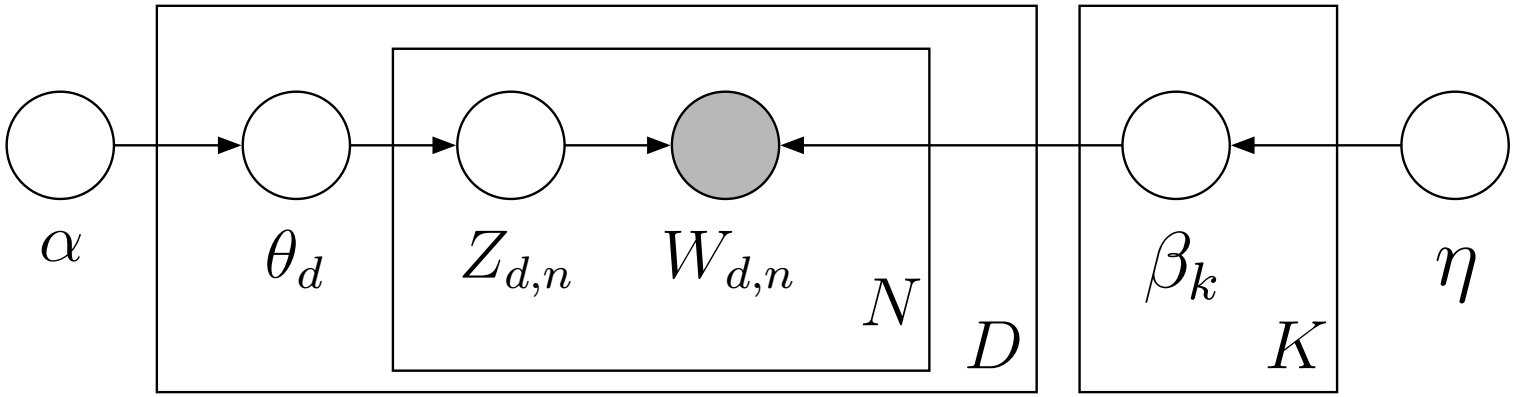


The variational paramters are:

γ	Dirichlet parameters
$\phi_{1:N}$	Multinomial parameters for K-dim variables

There is a separate variational Dirichlet distribution for each document; there is a separate multinomial distribution for each word in each document. (Contrast this to the model.)

Variational Inference for LDA



Coordinate ascent on the variational objective,

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

where

$$\mathrm{E}[\log \theta_i] = \Psi(\gamma_i) - \Psi(\textstyle\sum_j \gamma_j).$$

Estimating the topics

Maximum likelihood: Expectation-Maximization

- E-step: Use variational or MCMC to approximate the per-document posterior
- M-step: Find MLE of $\beta_{1:K}$ from expected counts

Bayesian topics

- Put a Dirichlet prior on the topics (usually exchangeable)
  Note/Warning: This controls the sparsity of the topics
- Collapsed Gibbs sampling is still possible—we only need to keep track of the topic assignments.
- Variational: Use a variational Dirichlet for each topic

Inference comparison

- Conventional wisdom says that:
  - Gibbs is easiest to implement
  - Variational can be faster, especially when dealing with nonconjugate priors (more on that later)
- There are other options:
  - Collapsed variational inference
  - Parallelized inference for large corpora
  - Particle filters for on-line inference
- An ICML paper examining these issues is Asuncion et al. (2009).

Outline

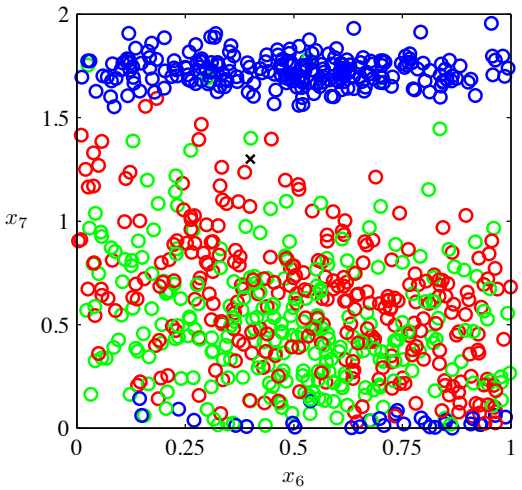
Dimensionality reduction

Motivation Given data that are high-dimensional $\boldsymbol{x} \in \mathbb{R}^D$, we want to find a low-dimensional representation $\boldsymbol{y} \in \mathbb{R}^M$ such that $M < D$:

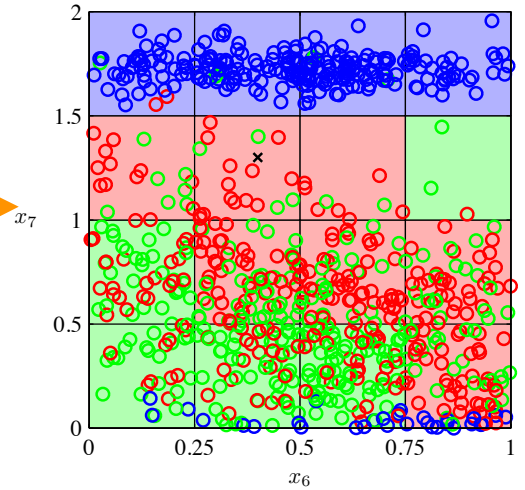
- Visualize data and discover intrinsic structures
- Save computational and storage cost
- Robust statistical modeling: curse of dimensionality

What is curse of dimensionality?

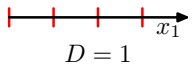
Ex: a simple classification scheme (related to decision tree)

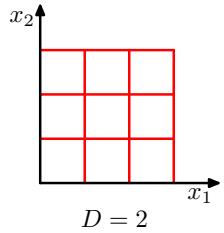


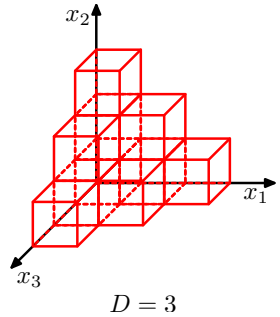
divide up
into small cells



of cells grows exponentially







# of cells r^D

r: number of divisions in each dimension

That is a lot even if D is just moderately large!

So to cover the whole space reasonably well,
you need exponentially number of training data points!

Another example

Nearest neighbor classification

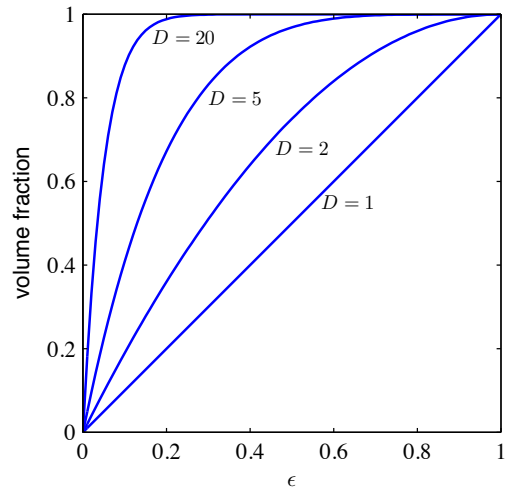
Say we want points of different classes are at least ε away from each other

To put things in scale (otherwise, you can scale ε arbitrarily), assume all points are at most 1 away from origin.

How many points between 1 and 1-ε from the origin?

$$1 - (1 - \epsilon)^D$$

Namely, when D is high, you cannot figure out data points that are ε away even ε is pretty big



Curse of dimensionality: summary

Intuition The higher the dimensionality, the more data points we need to train a model.

- To fill a unit-cube in $\mathbb{R}^D$ uniformly with data points, we need r^D where r is the edge length of small cells (i.e., dividing each axis in equal size of r.)

  Thus, if data is distributed that way, models such as decision trees need r^D training samples in order to make sure every cell is covered – in case a test sample falls into one of those cells.

- For a unit-ball $\boldsymbol{x} \leq 1$, a large percentage of data live in the shell — between the surface$\boldsymbol{x} = 1$ and the surface $\boldsymbol{x} = 1 - \epsilon$. The percentage is

$$1 - (1 - \epsilon)^D$$

  approaches 1. Thus, most data points in the high-dimensional space are crowded in the shell and are about the same distant from each other. To have data points that "look" substantially different from others, we need a exponentially large (in D) number of samples to fill the void $\boldsymbol{x} < 1 - \epsilon$.

An overview of dimensionality reduction

Parametric approaches

Linear: PCA, Fisher LDA, NMF, random projection, CCA, etc

Nonlinear: Neural networks, generative topological mapping, ICA

Nonparametric approaches

kernelized version of parametric methods: k-PCA, kICA, kCCA

manifold learning

Gaussian processes

Dimensionality reduction

- Linear: the low-dimensional coordinates $\boldsymbol{y}$ is parameterized as

$$\boldsymbol{y} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{x},$$

  where $\boldsymbol{U} \in \mathbb{R}^{M \times D}$
- Nonlinear: the relationship is through a nonlinear mapping

$$\boldsymbol{y} = f(\boldsymbol{x})$$

In both categories, there are many methods. We will focus on Principal Component Analysis (PCA) — a linear method for dimensionality reduction.

Linear dimensionality reductions

Many examples

Principal component analysis (PCA)

Fisher discriminant analysis (FDA)

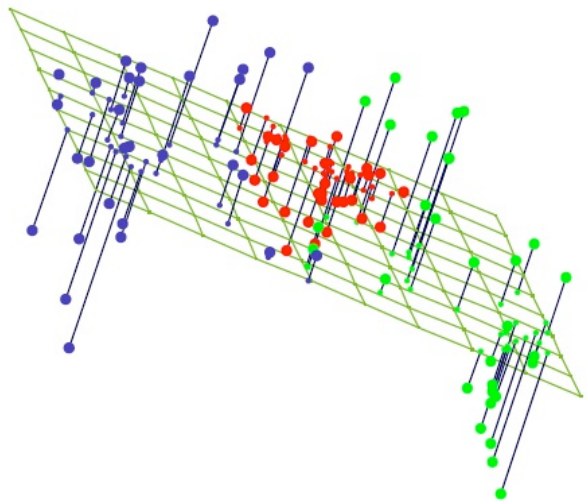
Nonnegative matrix factorization (NMF)

Framework

$$\boldsymbol{x} \in \Re^D \rightarrow \boldsymbol{y} \in \Re^M$$

$$D \gg M$$

$$\boldsymbol{y} = \boldsymbol{U}^\top \boldsymbol{x}$$



linear transformation of original space

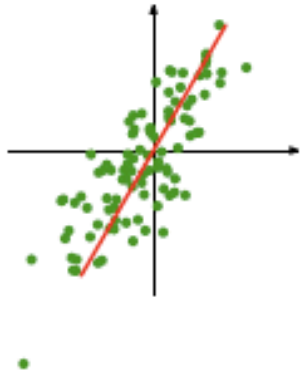
Linear dimensionality reduction: PCA

Intuition Consider the special case $M = 1$, namely, we are transforming $\boldsymbol{x}$ into a scalar via

$$y = \boldsymbol{u}^{\mathrm{T}} \boldsymbol{x}$$

which $\boldsymbol{u}$ is sensible?

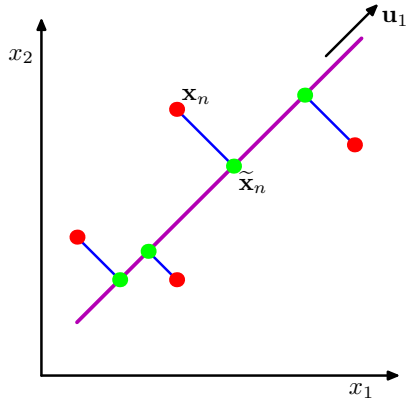
PCA: first principal component



Objective:

maximize variance of points projected on the line

Derivation on whiteboard



The framework of PCA

Assumption:

Centered inputs (if not, subtract mean)

$$\boldsymbol{x} \in \mathbb{R}^D \quad \sum_i \boldsymbol{x}_i = \mathbf{0}$$

Projection into subspace

$$\boldsymbol{U} \in \mathbb{R}^{D \times M} \quad \boldsymbol{y}_i = \boldsymbol{U}^{\mathrm{T}} \boldsymbol{x}_i$$

$$\boldsymbol{U}^{\mathrm{T}} \boldsymbol{U} = \boldsymbol{I}$$

Solution

each row is a data sample

Computer covariance matrix

$$\boldsymbol{S} = \frac{1}{N} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}$$

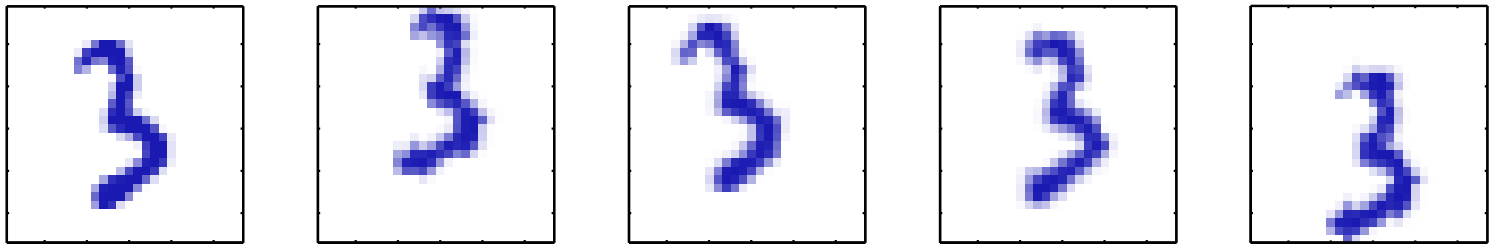
Diagonalize S: $(\boldsymbol{u}_d, \lambda_d)$ $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$

Use top D eigenvectors to form U

$$\boldsymbol{U} = (\boldsymbol{u}_1 \ \boldsymbol{u}_2 \ \cdots \ \boldsymbol{u}_M)$$

Examples of running PCA

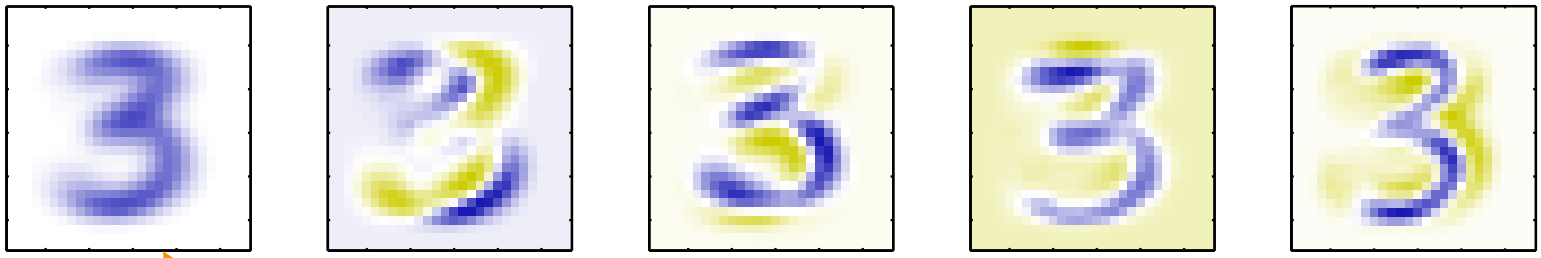
Original Images



Eigenvectors

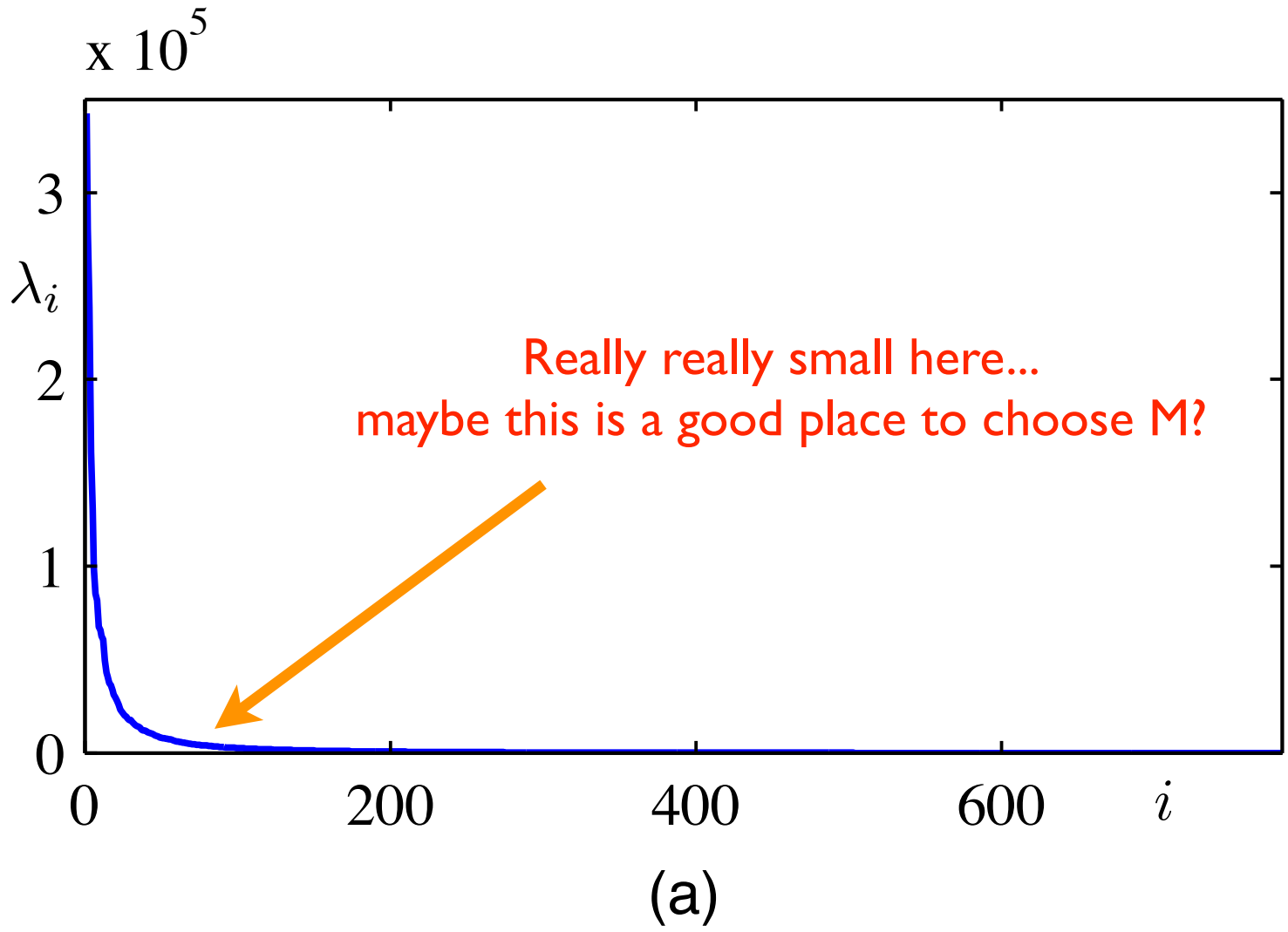
they look like blurred original images

Mean $\lambda_1 = 3.4 \cdot 10^5$ $\lambda_2 = 2.8 \cdot 10^5$ $\lambda_3 = 2.4 \cdot 10^5$ $\lambda_4 = 1.6 \cdot 10^5$



Used to centralize inputs

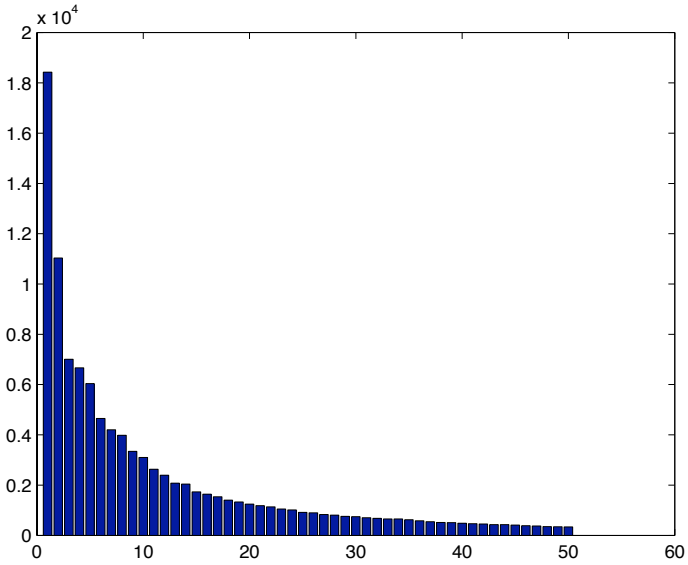
Eigenspectrum of the covariance matrix

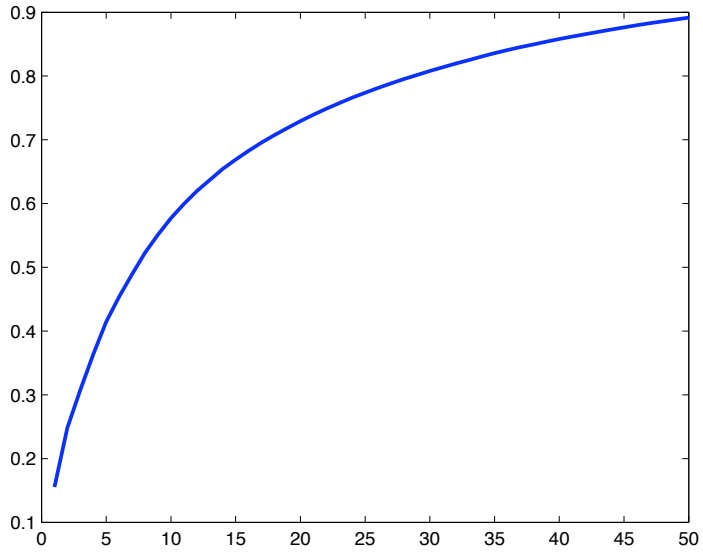


A slightly sensible approach

common choice is 95% or 90%

$$\frac{\sum_{d=1}^{M} \lambda_d}{\sum_{d=1}^{D} \lambda_d} \geq \text{Threshold}$$





Application of PCA

Preprocessing

Diagonalize data

$$\boldsymbol{y}_i = \boldsymbol{U}^{\mathrm{T}} \boldsymbol{x}_i$$

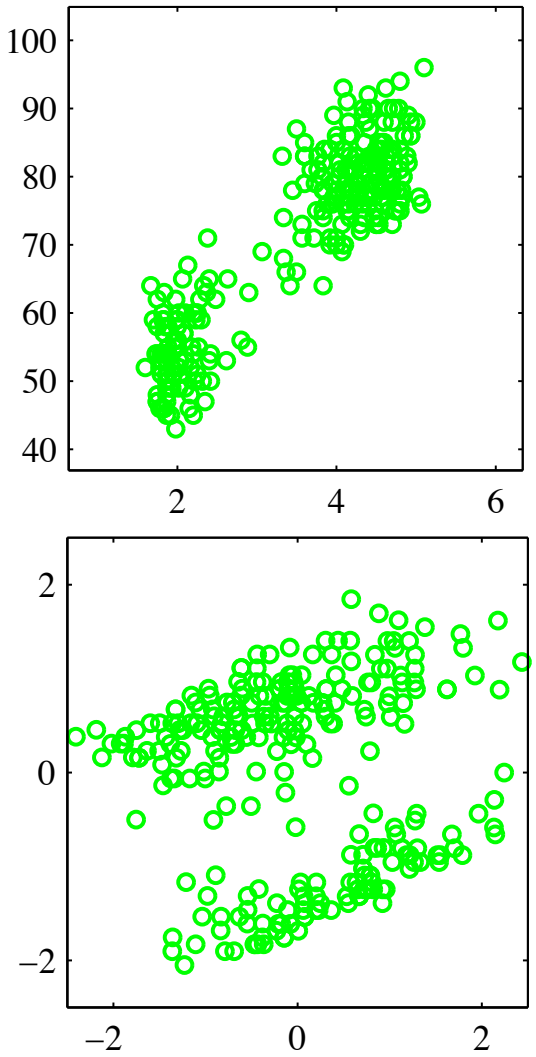
Normalize data (whitening)

$$\boldsymbol{y}_i = \boldsymbol{\lambda}^{-1/2} \boldsymbol{U}^{\mathrm{T}} \boldsymbol{x}_i$$

Benefits:

1) depress noisy features

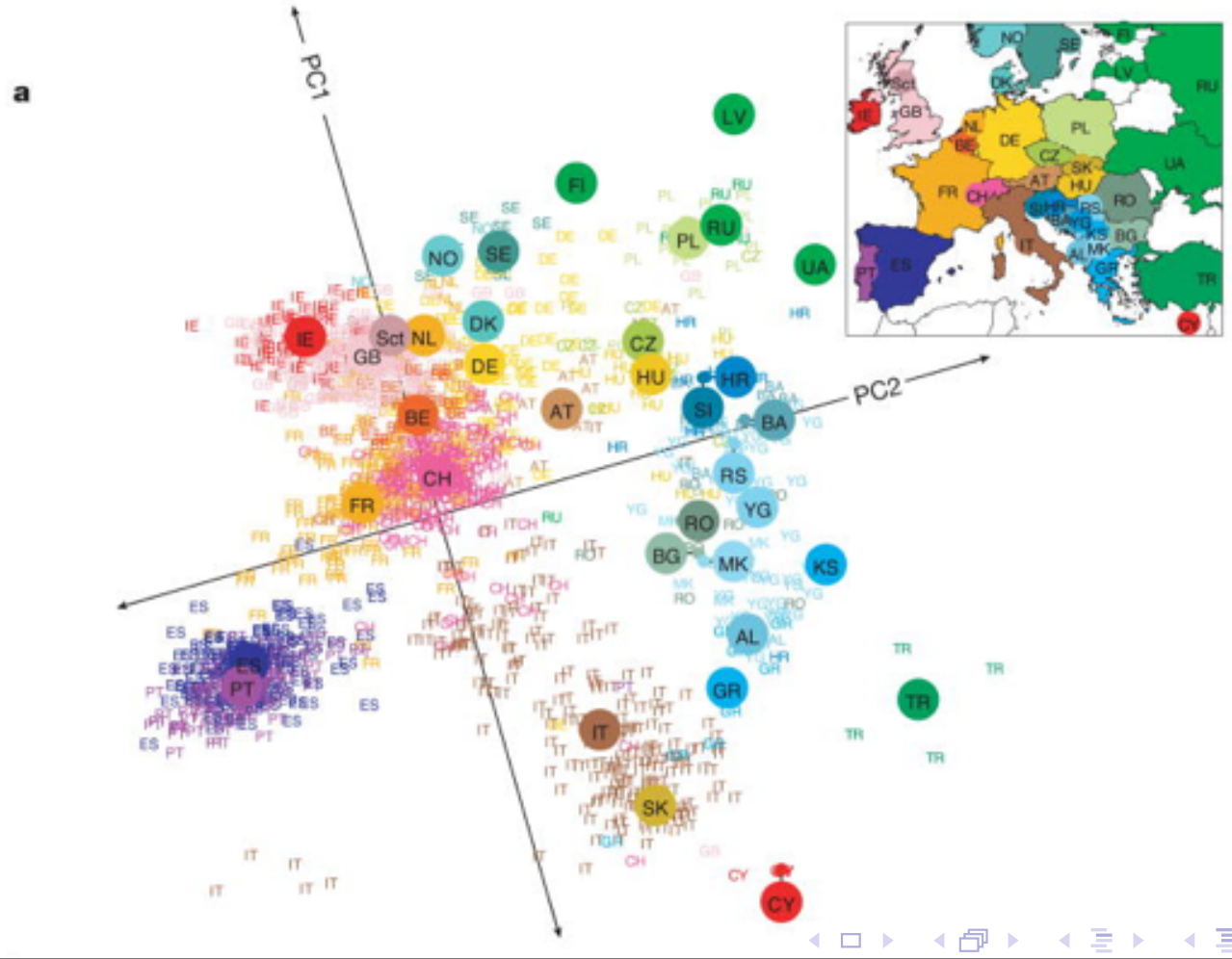
2) couple with other models



A very cool application

From genetic data to discover the origin of your ancestors:
`https://www.ncbi.nlm.nih.gov/pubmed/18758442`

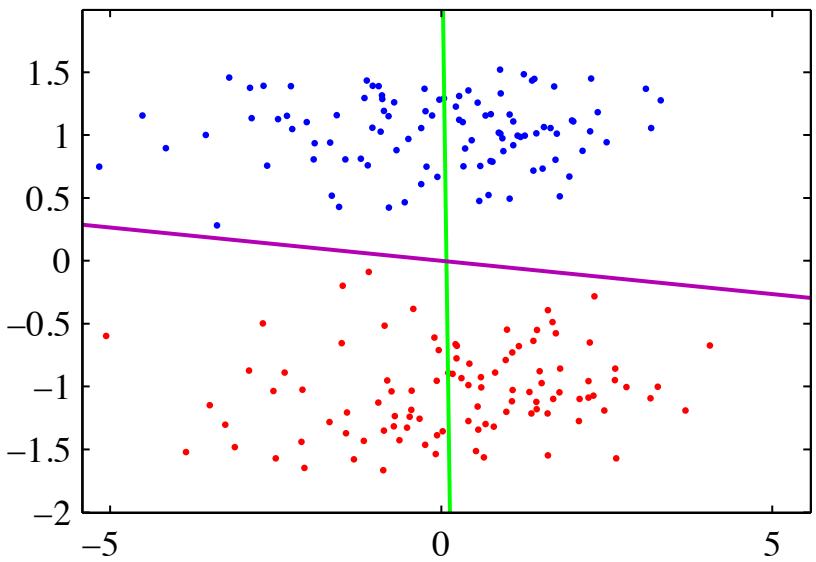
Visualizing data with PCA



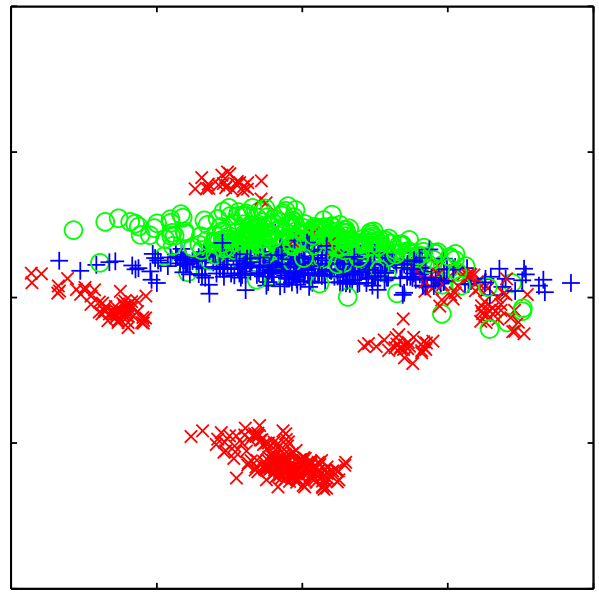
Does PCA help for classification?

Not always!

no help



help



Visualization

The second interpretation of PCA

minimum reconstruction error

Consider a basis of a subspace with M-dimension where M < D

$$\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_M \in \mathbb{R}^D$$

Consider the approximation error by linearly combining the bassis

$$\sum \|\boldsymbol{x}_n - \sum_m \alpha_{nm} \boldsymbol{u}_m\|_2^2$$

If we minimize the error by optimizing both the basis and the coefficients, we get

$$\alpha_{nm} = \boldsymbol{x}_n^{\mathrm{T}} \boldsymbol{u}_m$$

$\boldsymbol{u}_m$ is the m-th largest eigenvector of S(ample) covariance matrix

Unexplained variance and residual error

Unexplained variance

Each projection direction u_i contributes λ_i variance

With D-dimensional data, M projection directions, the "leftover" is

$$\sum_{d=1}^{D} \lambda_d - \sum_{m=1}^{M} \lambda_m = \sum_{m=M+1}^{D} \lambda_m$$

Reconstruction error

$$\sum \| \boldsymbol{x}_n - \sum_m \alpha_{nm} \boldsymbol{u}_m \|_2^2$$

The two are exactly the same!!!

Issues with PCA

Choose dimensionality

ad hoc approach: we have seen that

more principled approach: bayesian PCA, etc

Missing values

what if data is missing?

Probabilistic PCA: a little bit later

How to get nonlinear dimensionality reduction

- Kernel PCA
- Autoencoder with nonlinear neural networks
- Manifold learning

We need nonlinear projection

Intuition:

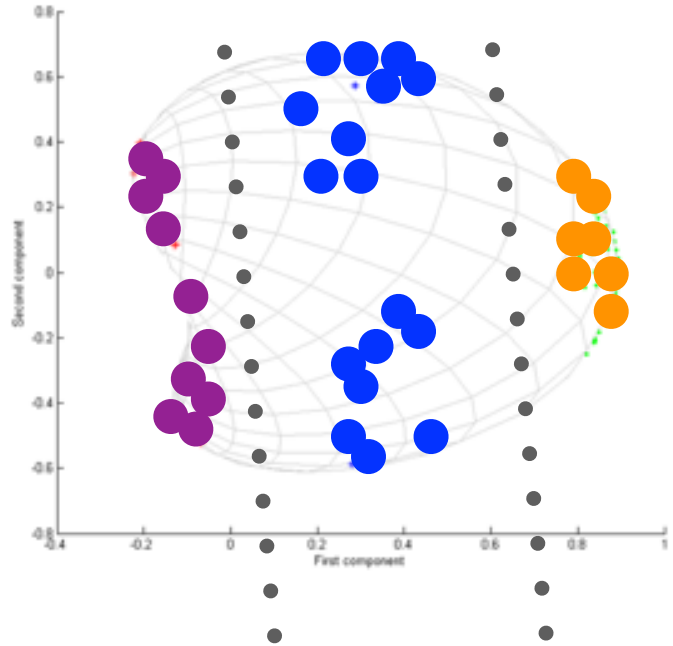
Find a good space through

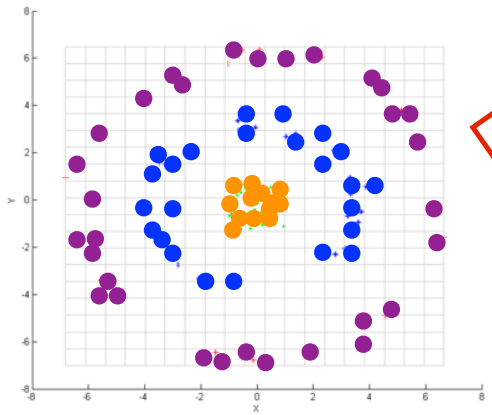
nonlinear mapping;

Then do linear projection

(as in PCA)

✓ Data separated!

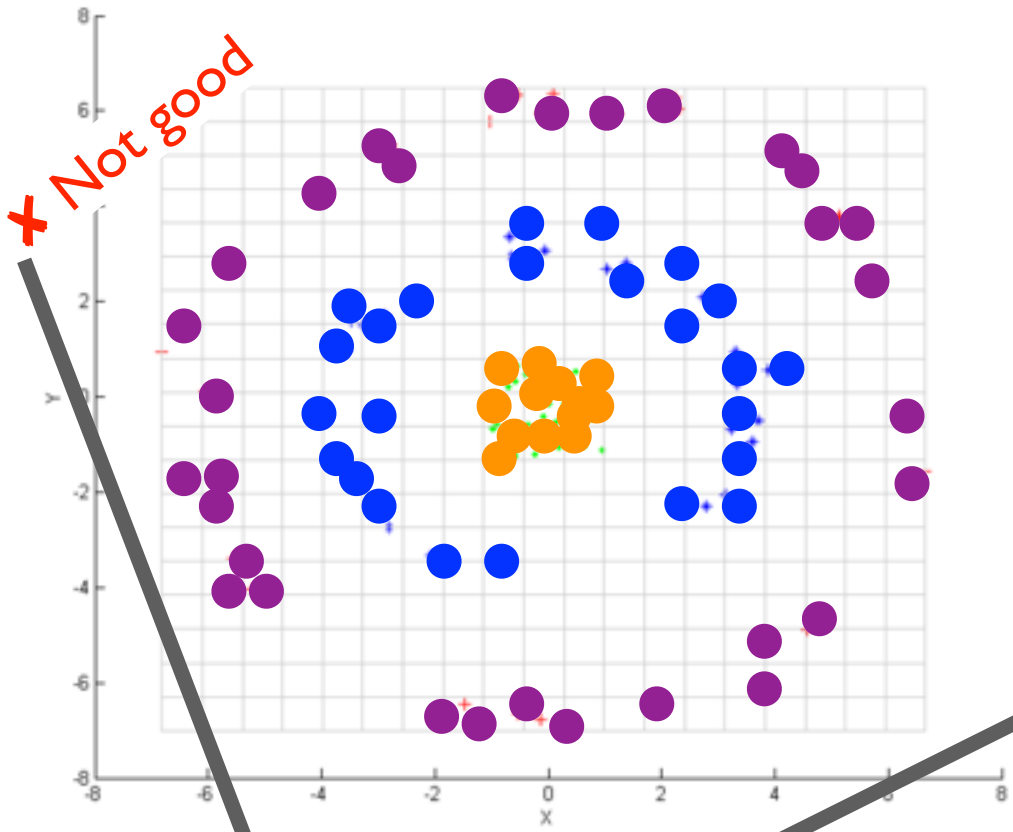




How to find this
nonlinear mapping?

it is called kernel PCA!

Why? ---- Linear is not enough



Issue:
projection on any 1D line will cause mixing of data from different classes.

✘ Not good

Details of kernel PCA

Derivation

see note

Demo