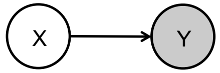
DD2434 Machine Learning, Advanced Course

Lecture 11: Topic Models

Hedvig Kjellström
hedvig@kth.se
https://www.kth.se/social/course/DD2434/

Latent Variable Models for Discrete Data

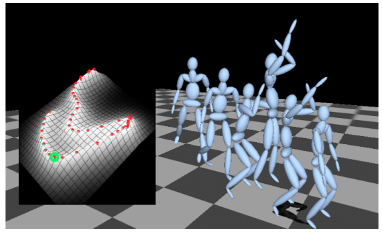
Previously: Latent Variable Models for continuous data



PPCA, HMM, GPLVM…

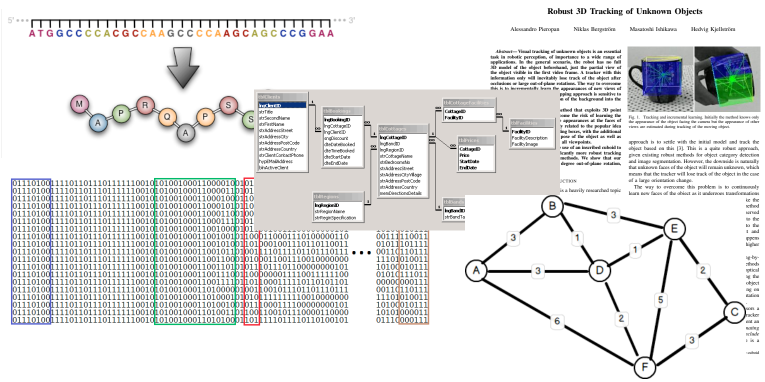
In general: Y noisy and high-dim observation, X structured and low-dim representation

Example from Lecture 6: GPLVM, X = latent low-dim motion space, Y = all joint angles of the human



Latent Variable Models for Discrete Data

Discrete data:



Today

Some slides from the tutorial on topic models at ICML 2012, given by David Blei

The idea of modeling text documents according to topics (David Blei ICML 2012 tutorial)

Text data and the bag of words model (Murphy 3.4, 27.1)

Plate notation (Murphy 10.4.1)

Latent Dirichlet Allocation (LDA) (Murphy 27.3, David Blei ICML 2012 tutorial)

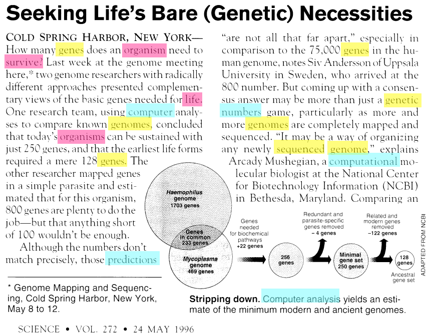
Text Documents and Topics

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

Discuss with your neighbor (2 min):
Can you see patterns in how words appear in the 4 columns?

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

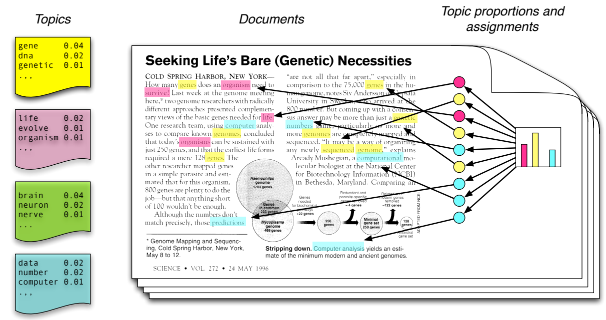
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

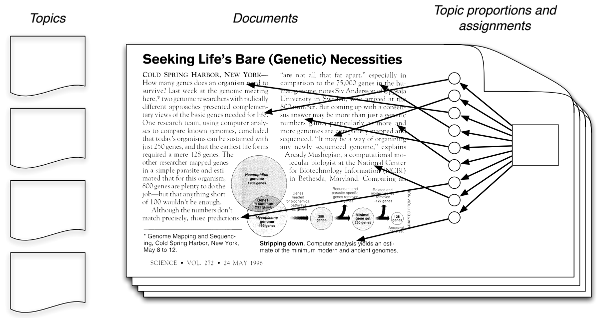
Simple intuition: Documents exhibit multiple topics.

Latent Dirichlet allocation (LDA)



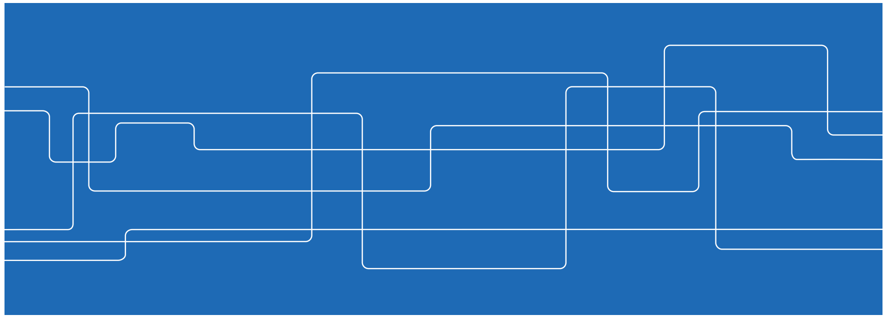
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Latent Dirichlet allocation (LDA)

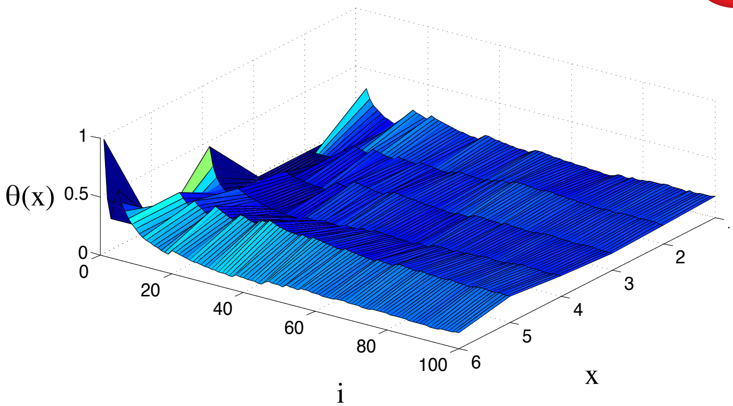


- In reality, we only observe the documents
- The other structure are **hidden variables**

Text Data and Bag of Words



From Lecture 10: Dice Roll



Dice Roll as an Example of Multinomial Distribution

Suppose that we observe $\mathcal{D} = \{x_1, ..., x_N\}$ where $x_i \in \{1, ..., K\}, K = 6$

The rolls are independent so the likelihood is

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{K} \theta_k^{N_k}$$

where N_k is the number of times the dice turned up *k*
This is a **Multinomial** distribution.

The prior and likelihood are both **Dirichlet**, the conjugate of multinomial – more later.

Multinomial Distribution of Text

Multinomial distribution – essentially normalized histogram over a finite set of outcomes

In dice case, set of outcomes $x_i \in \{1, ..., K\}, K = 6$

Discuss with your neighbor (5 min):

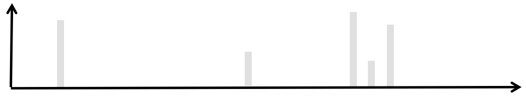
What is the set of possible outcomes if we think of a text document instead of a sequence of dice rolls?

Multinomial Distribution of Text

Statespace = set of unique words in the language in which the text document is written

High-dim   Sparse

Multinomial distribution (normalized histogram) of a text document is called a **bag of words**



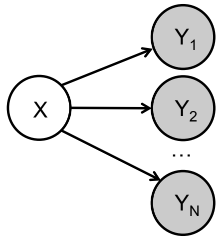
Discuss with your neighbor (5 min):

What information have you thrown away when you represent data as a bag of words?

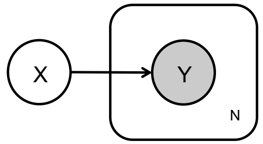
Plate Notation

Many Independent and Identically Distributed (i.i.d.) Variables – Variables That Repeat

Graphical model would look like this – not very convenient:

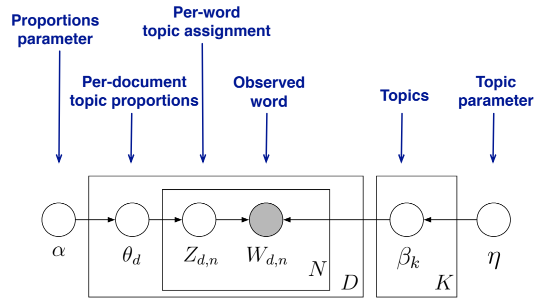


Therefore you use **plate notation**, which means the same thing:



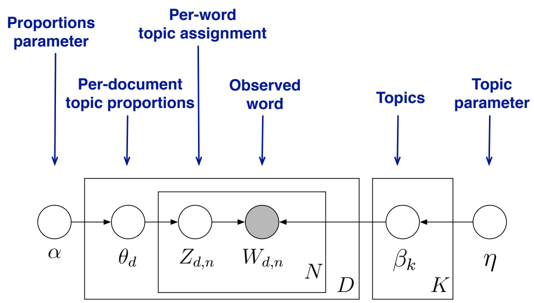
Latent Dirchlet Allocation (LDA)

LDA as a graphical model



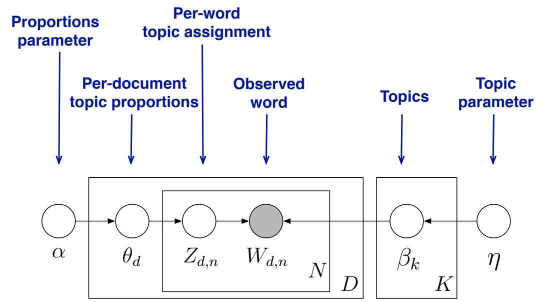
- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

LDA as a graphical model



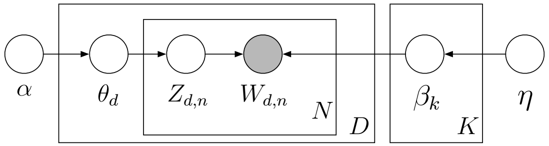
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed.
- Plates indicate replicated variables.

LDA as a graphical model



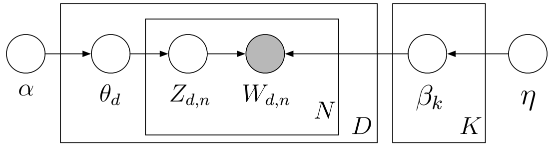
$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

LDA as a graphical model



- This joint defines a posterior.
- 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, e.g., information retrieval, document similarity, exploration, ...

LDA as a graphical model



Approximate posterior inference algorithms

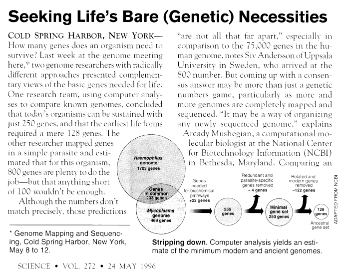
- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)
- Online variational inference (Hoffman et al., 2010)

Also see Mukherjee and Blei (2009) and Asuncion et al. (2009).

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

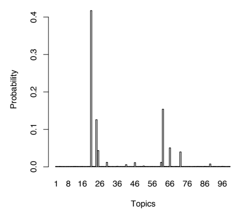
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

* 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

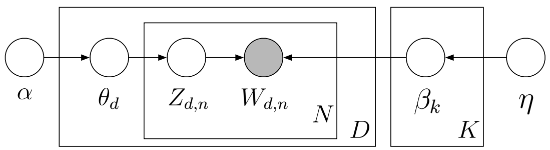
Aside: The Dirichlet distribution

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

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

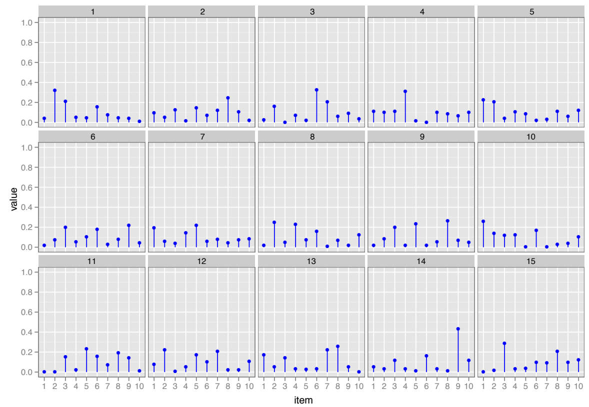
- It is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- The parameter α controls the mean shape and sparsity of θ.
- The topic proportions are a K dimensional Dirichlet. The topics are a V dimensional Dirichlet.

LDA as a graphical model

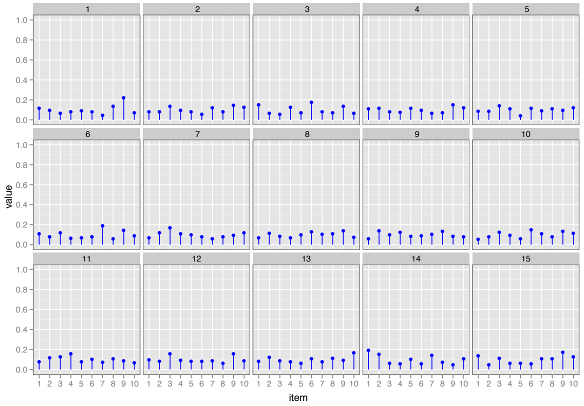


Discuss with your neighbor (5 min):
What would happen to the topic distribution if we removed α?

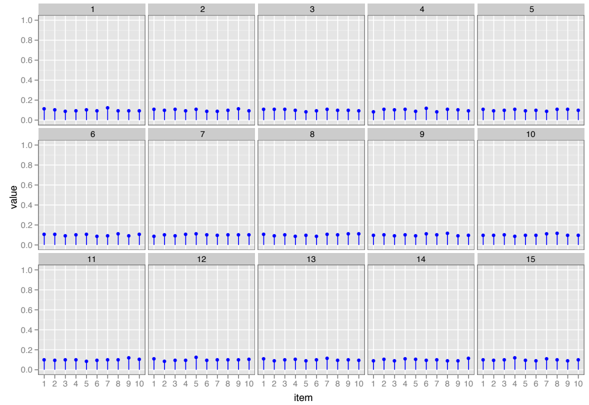
$\alpha = 1$



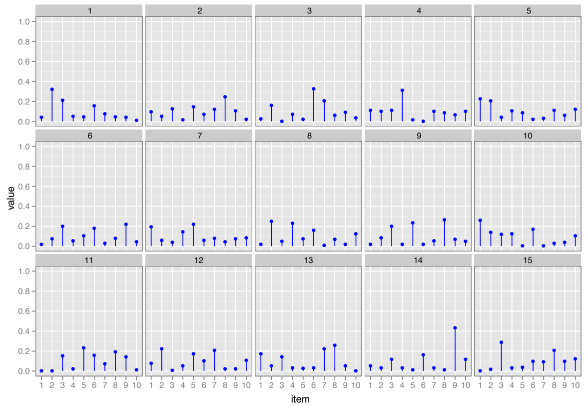
$\alpha = 10$



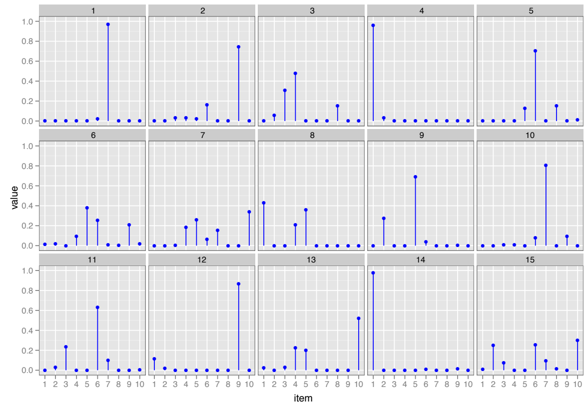
$\alpha = 100$



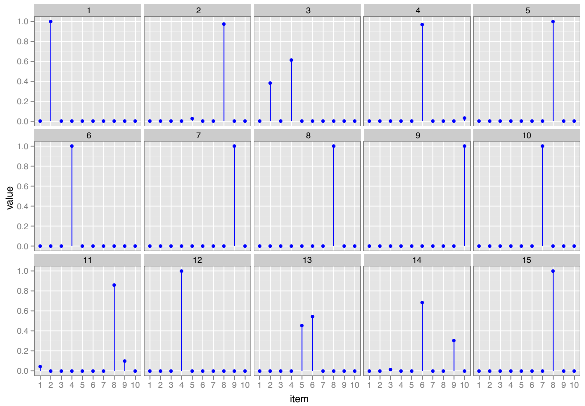
$\alpha = 1$



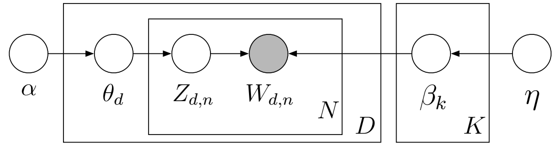
$\alpha = 0.1$



$\alpha = 0.01$

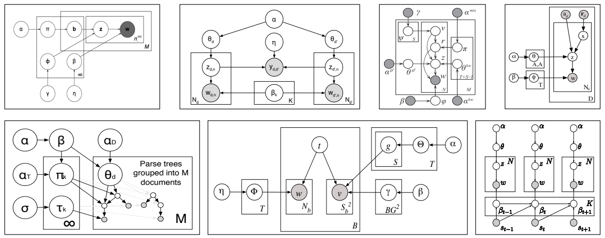


LDA summary



- 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 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



- Organizing and finding patterns in data has become important in the sciences, humanities, industry, and culture.
- LDA can be embedded in more complicated models that capture richer assumptions about the data.
- Algorithmic improvements let us fit models to massive data.

What is next?

Assignment 3 – report in tomorrow 17 Dec by NOON

Project – talk to your group and send your supervisor an email about what you plan to do

Wed 17 Dec 15:15-17:00 Q31
Lecture 12: Method and Model Selection
Hedvig Kjellström
Readings: Murphy Chapter Murphy Chapter 1, 5.3, 8.6
Presentation of "early group" project (students who are leaving KTH and therefore finish before New Year)