

Latent Dirichlet Allocation

David Mareček

📅 October 22, 2024



EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

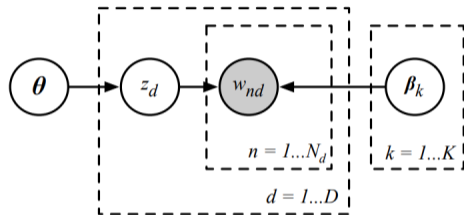
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated

Many of the slides in this presentation were taken from the presentations of Carl Edward Rasmussen (University of Cambridge)

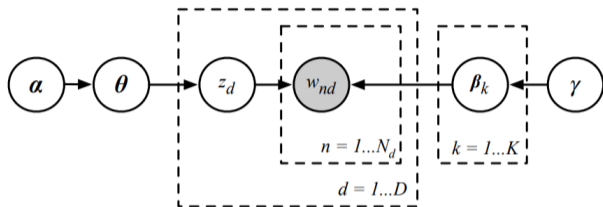
Mixture of Categoricals Model



$$z_d \sim \text{Cat}(\vec{\theta})$$
$$w_{nd} | z_d \sim \text{Cat}(\vec{\beta}_{z_d})$$

With the Expectation-Maximization algorithm we have essentially estimated $\vec{\theta}$ and $\vec{\beta}$ by maximum likelihood.

Bayesian Mixture of Categoricals Model



$$z_d \sim \text{Cat}(\vec{\theta})$$

$$\vec{\theta} \sim \text{Dir}(\vec{\alpha})$$

$$w_{nd} | z_d, \vec{\beta} \sim \text{Cat}(\vec{\beta}_{z_d})$$

$$\vec{\beta}_k \sim \text{Dir}(\vec{\gamma})$$

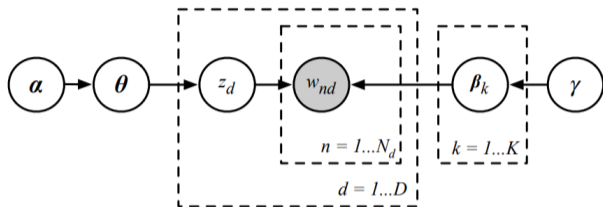
An alternative, Bayesian treatment infers these parameters starting from priors, e.g.:

- $\vec{\theta} \sim \text{Dir}(\vec{\alpha})$ is a symmetric Dirichlet over category probabilities,
- $\vec{\beta}_k \sim \text{Dir}(\vec{\gamma})$ are symmetric Dirichlets over vocabulary probabilities.

What is different?

- We no longer want to compute a point estimate of $\vec{\theta}$ and $\vec{\beta}$.
- We are now interested in computing posterior distributions.

Limitations of the mixture of categoricals model



$$\begin{aligned}z_d &\sim \text{Cat}(\vec{\theta}) \\ \vec{\theta} &\sim \text{Dir}(\vec{\alpha}) \\ w_{nd} | z_d, \vec{\beta} &\sim \text{Cat}(\vec{\beta}_{z_d}) \\ \vec{\beta}_k &\sim \text{Dir}(\vec{\gamma})\end{aligned}$$

A generative view of the mixture of categoricals model:

1. Draw a distribution $\vec{\theta}$ over K topics from a $\text{Dir}(\vec{\alpha})$.
2. For each topic k , draw a distribution $\vec{\beta}_k$ over words from a $\text{Dir}(\vec{\gamma})$.
3. For each document d , draw a topic z_d from a $\text{Cat}(\vec{\theta})$
4. For each document d , draw N_d words w_{nd} from a $\text{Cat}(\vec{\beta}_{z_d})$

Limitations:

- All words in each document are drawn from one specific topic distribution.
- This works if each document is exclusively about one topic, but if some documents span more than one topic, then “blurred” topics must be learnt.

Latent Dirichlet Allocation

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Seeking Life's Bare (Genetic) Necessities

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

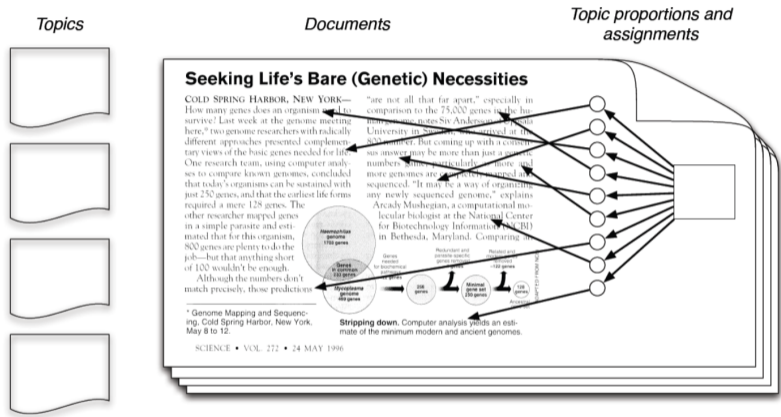
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

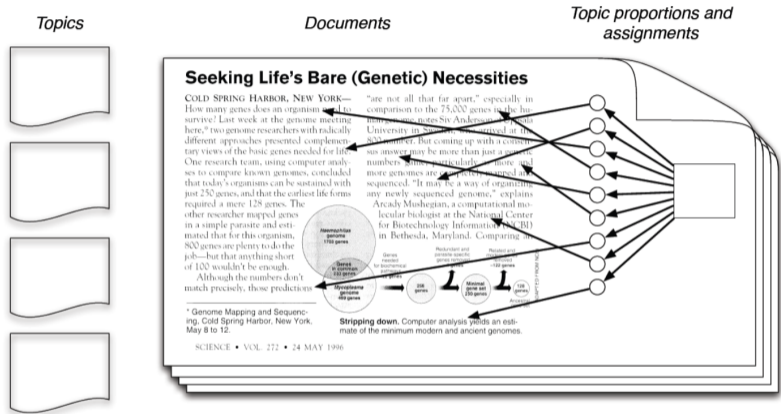
Latent Dirichlet Allocation: what we observe



In reality, we only observe the documents.

The other structure are hidden variables.

Latent Dirichlet Allocation: what we observe

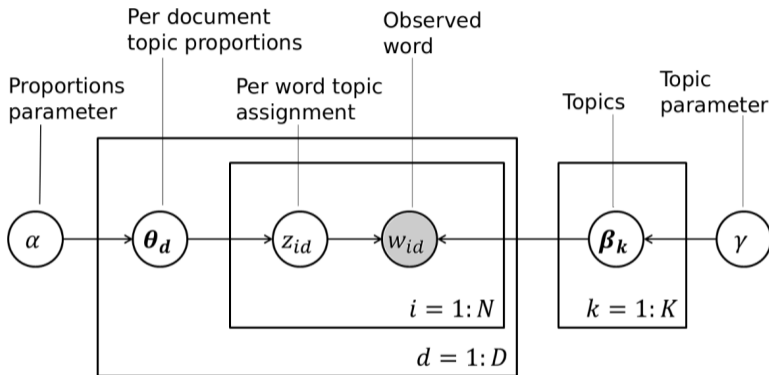


Our goal is to infer the hidden variables.

This means computing their distribution conditioned on the documents

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

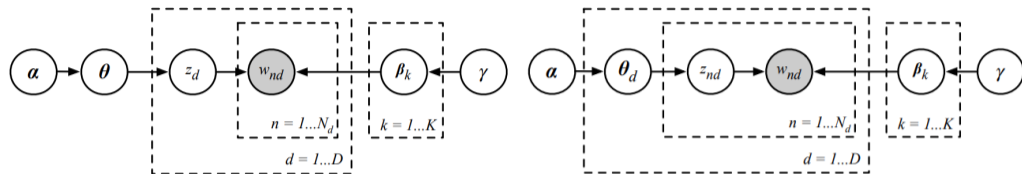
Latent Dirichlet Allocation: graphical model



Nodes are random variables; edges indicate dependence.

Shaded nodes indicate observed variables.

Mixture of Categoricals vs. LDA



A generative view of LDA:

1. For each document d draw a distribution $\vec{\theta}_d$ over topics from a $Dir(\vec{\alpha})$.
2. For each topic k draw a distribution $\vec{\beta}_k$ over words from a $Dir(\vec{\gamma})$.
3. Draw a topic z_{nd} for the n -th word in document d from a $Cat(\vec{\theta}_d)$.
4. Draw word w_{nd} from a $Cat(\vec{\beta}_{z_{nd}})$.

Differences with the mixture of categoricals model:

- In LDA, every word in a document can be drawn from a different topic,
- and every document has its own distribution over topics $\vec{\theta}_d$.

Gibbs sampling algorithm

- Initialize z_{nd} randomly for all words in all documents
- Choose random word and sample a new category based on all other words in all other documents.
- The distribution over categories is the predictive distribution of the posterior Dirichlet distribution (integration across all possible $\vec{\theta}$ and $\vec{\beta}$).
- Perform these small changes in many iterations over the data. The algorithm will converge to good solutions.