

(Deep) Generative Models for Unsupervised Learning (Part 2 – Classical Models - LDA)

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Latent Dirichlet Allocation (LDA) a.k.a. “Topic Model”



Motivation: Multinomial Mixture Model for Text

- Assume D documents, and document d has N_d words in it
- We can represent doc d by a word count vector \mathbf{w}_d
- Assuming a vocab of V unique words, \mathbf{w}_d is a $V \times 1$ vector of counts
 - w_{dv} = no of times word v appears in doc d
- Let's model the docs by a mixture of K multinomial distributions, each V -dim
 - The k^{th} multinomial modeled by a V -dim prob vector ϕ_k (sums to 1)
 - ϕ_k can be thought of as a "topic vector" (or just "topic"), ϕ_{kv} : prob of word v in topic k
- Generative model and plate diagram below

Each topic is a prob. distribution over word tokens

Each representing a "topic" (K topics)

Limitation: Each doc d belongs to a single cluster z_d and all words in a document assumed to be from the same topic. This is unrealistic/restrictive

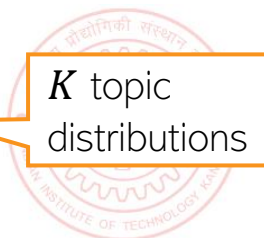
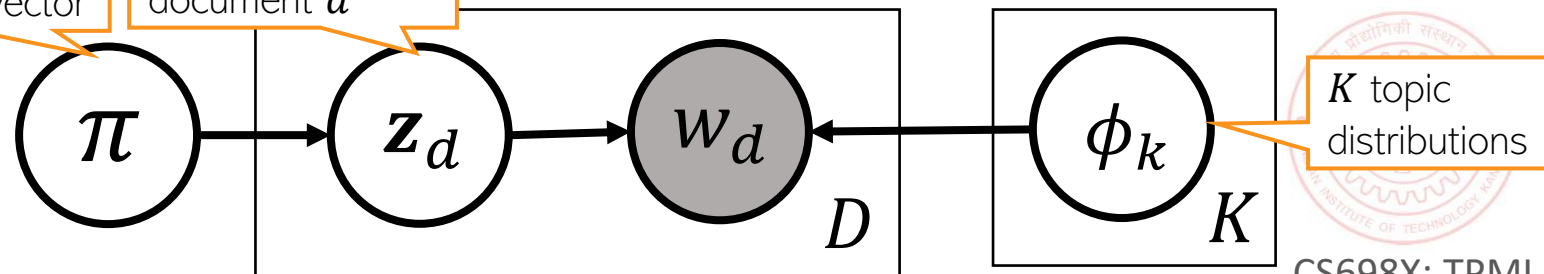
$$\mathbf{z}_d \sim \text{multinoulli}(\pi)$$

Topic Mixing proportion vector

Cluster/topic of document d

$$\mathbf{w}_d \sim \text{multinomial}(\phi_{z_d}, N_d)$$

Counts will sum to N_d



Documents can be about multiple topics

Seeking Life's Bare (Genetic) Necessities

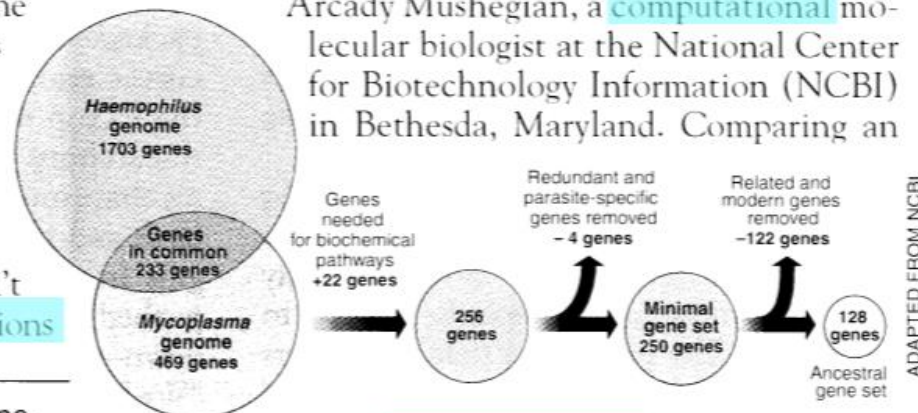
COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses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

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

SCIENCE • VOL. 272 • 24 MAY 1996

“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.

How do we find the word-topic associations in each document?

How do we use them to learn topics in the given text collection?

How do we learn low-dim document representations in terms of the topics they represent?



A More Fine-Grained Mixture Model for Text

- Assume a corpus-level topic mixing proportions α ($K \times 1$ prob vector)
- Also assume doc-level topic mixing props θ_d ($K \times 1$ prob vector)
- Instead of assuming a single cluster \mathbf{z}_d for doc d , cluster each word in it
 - $\mathbf{z}_{d,n} \in \{1, 2, \dots, K\}$ denotes the cluster/topic of word $w_{d,n} \in \{1, 2, \dots, V\}$
- Can obtain the “average” clustering for doc d using θ_d or $\bar{\mathbf{z}}_d = \frac{1}{N_d} \sum_{n=1}^{N_d} \mathbf{z}_{d,n}$
- The generative model is as follows

Each assumed a one-hot $K \times 1$ vector

Locally-conjugate. Easy
Gibbs sampling, VI, etc

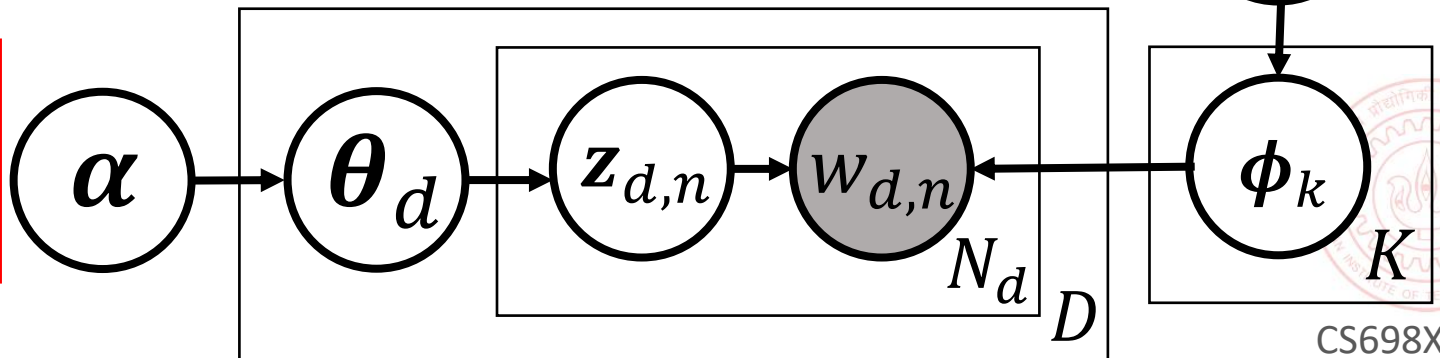
Latent Dirichlet
Allocation* (LDA)
Topic Model

Somewhat similar to
Dir-Mult PCA model

$$\phi_k \sim \text{Dirichlet}(\eta) \quad k = 1, 2, \dots, K \quad (V\text{-dim Dirichlet})$$

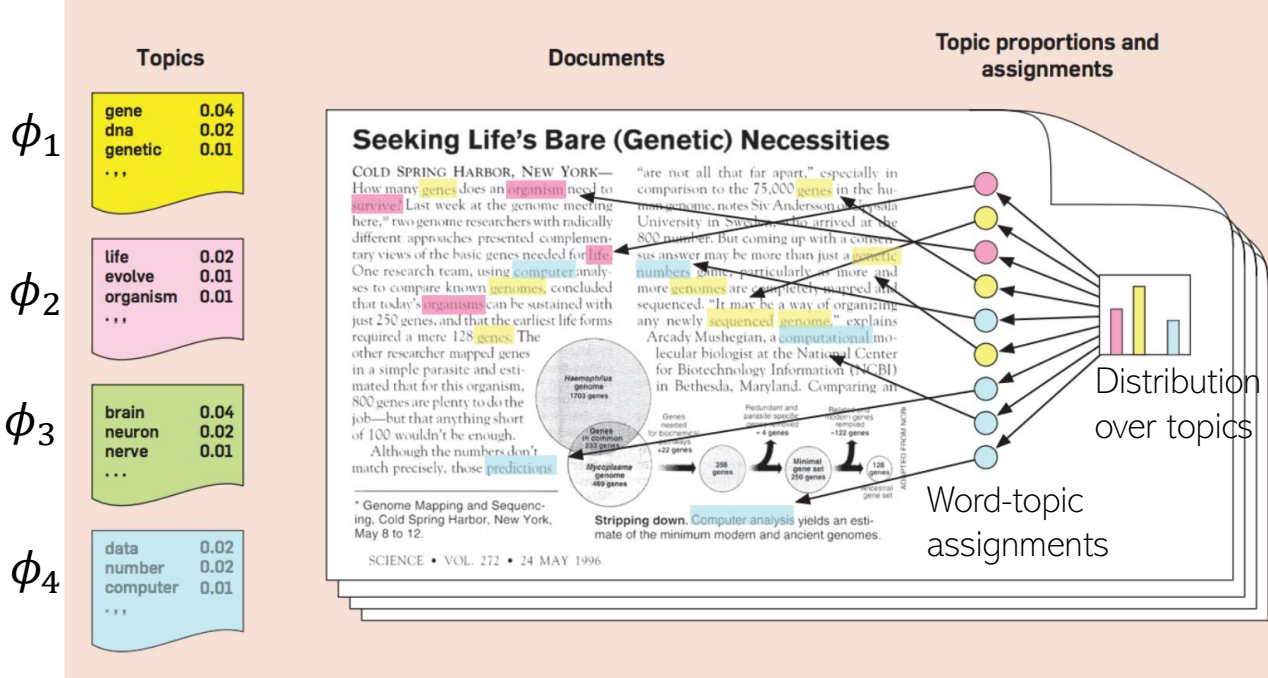
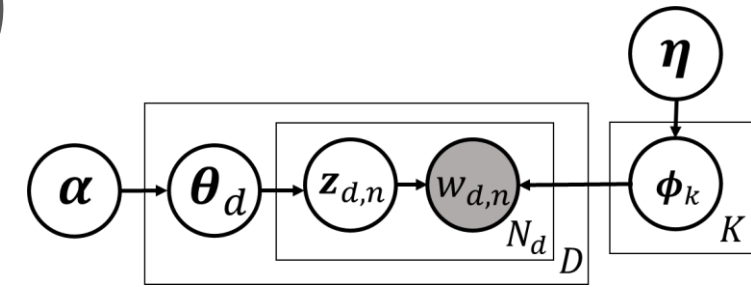
$$\theta_d \sim \text{Dirichlet}(\alpha) \quad d = 1, 2, \dots, D \quad (K\text{-dim Dirichlet})$$

$$\begin{aligned} \mathbf{z}_{d,n} &\sim \text{multinoulli}(\theta_d) \\ \mathbf{w}_{d,n} &\sim \text{multinoulli}(\phi_{\mathbf{z}_{d,n}}) \end{aligned}$$



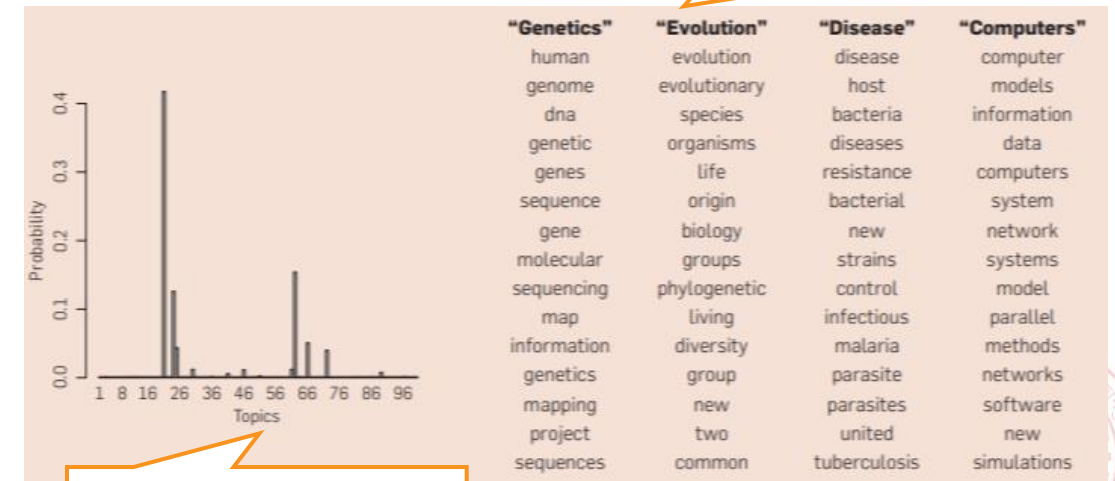
Latent Dirichlet Allocation (LDA)

- A very widely used probabilistic model for text data
- Nice and easy insights into the text collection
 - Each $\phi_k = [\phi_{k1}, \dots, \phi_{kV}]$ can be interpreted as topic (ϕ_{kv} = prob. of word v in topic k)
 - $\theta_d = [\theta_{d1}, \dots, \theta_{dK}]$: how much each topic is present in document d (topic distribution)
 - $\bar{z}_d = \frac{1}{N_d} \sum_{n=1}^{N_d} z_{d,n}$ also has a similar interpretation as θ_d



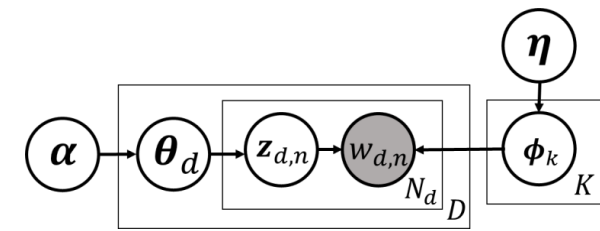
A topic is a set of words that tend to co-occur together

15 most frequent (most probable) words from four most prominent topics in this doc



LDA: Inference and Evaluation

7



- LDA is locally conjugate. Many inference methods (VI, variational EM, Gibbs samp, etc)

$$p(\mathbf{Z}, \Theta, \Phi | \mathbf{W}, \alpha, \eta) = \frac{p(\mathbf{W} | \Phi, \mathbf{Z}) p(\mathbf{Z} | \Theta) p(\Phi | \eta) p(\Theta | \alpha)}{p(\mathbf{W} | \alpha, \eta)} \quad (\text{assuming hyperparams } \alpha, \eta \text{ are fixed})$$

- Can even collapse some variables and do collapsed Gibbs or collapsed VB
 - E.g., collapse θ_d and ϕ_k (if needed, these can be approximated using \mathbf{Z})
- Many ways to evaluate how well LDA performs on some data
 - Extrinsic measures: Perform LDA and use its output for another task (e.g., classification)
 - Perplexity is another **intrinsic** measure to evaluate LDA-style models

Lower is better

Test set with M docs

Marginal likelihood of all words in the d^{th} test doc

$$perplexity(D_{\text{test}}) = \exp \left\{ -\frac{\sum_{d=1}^M \log p(\mathbf{w}_d)}{\sum_{d=1}^M N_d} \right\}$$



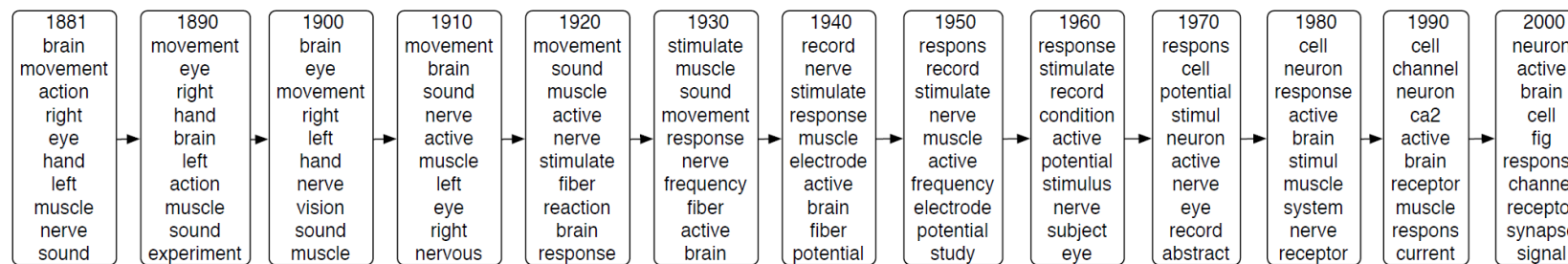
LDA: Limitations and Extensions

- LDA assumes topics remain static over time (improvement: Dynamic Topic Model)

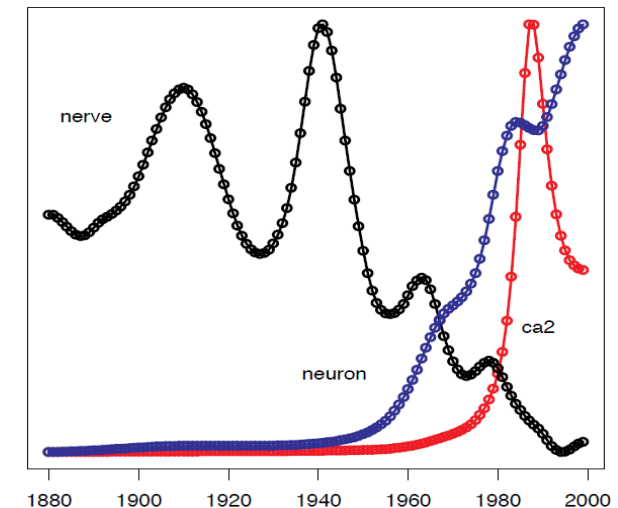
Assume a first-order Markov evolution for each topic w.r.t. time

$$w_k^t \sim \mathcal{N}(w_k^{t-1}, \sigma^2 I) \quad \phi_k^t = \mathcal{S}(w_k^t)$$

Simplex transformation (convert w_k^t into a probability vector)



Evolution of topic “Neuroscience”
(learned from the journal Science)



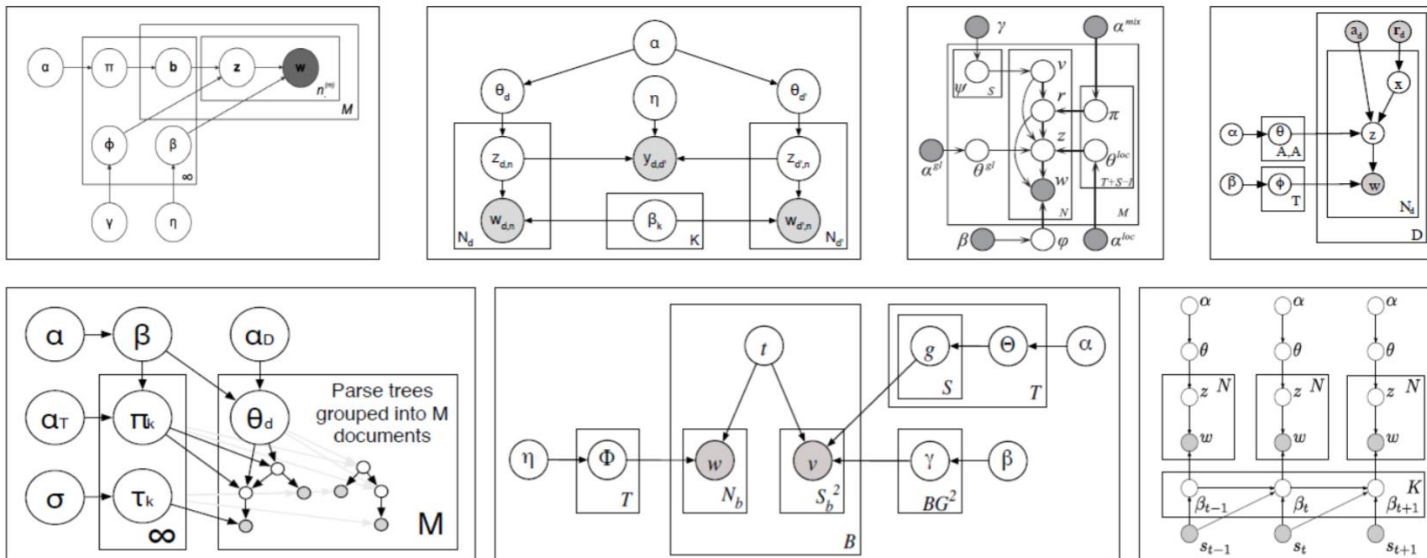
- LDA assumes topics are uncorrelated (improvement: Corr-LDA)
 - Use a **logistic normal** distribution on θ_d (cov matrix of log-normal makes component correlated)
- LDA ignores the sequential structure in the text (improvement: HMM-LDA)



LDA Extensions (Contd)

- LDA for non-text data, e.g., images
 - Each image can be represented as a bag of “visual words” and LDA can be applied
- Supervised/Labeled LDA (when we have a label for each document)
- LDA for paired/multimodality data (e.g., images and text caption)
- LDA for graph-structured data instead of documents

Plate diagrams for some LDA extensions



LDA is also equivalent to doing a non-negative matrix fact. of the $V \times D$ word-document matrix \mathbf{X} using a Poisson likelihood model*

$$\mathbf{X} \sim \text{Poisson}(\Phi\Theta)$$

Φ ($V \times K$) and Θ ($K \times D$) can be given any non-negative priors (Dirichlet/gamma)

This can be extended to “deep” matrix factorization** (modeling Θ using many layers)

*Sec 4 and 5 of “Beta-Negative Binomial Process and Poisson Factor Analysis” (Zhou et al, 2012)

** Poisson-gamma belief networks” (Zhou et al, 2015)

