NPBayes (wrap-up), Topic Modeling, and Simulation based Inference

CS772A: Probabilistic Machine Learning

Piyush Rai

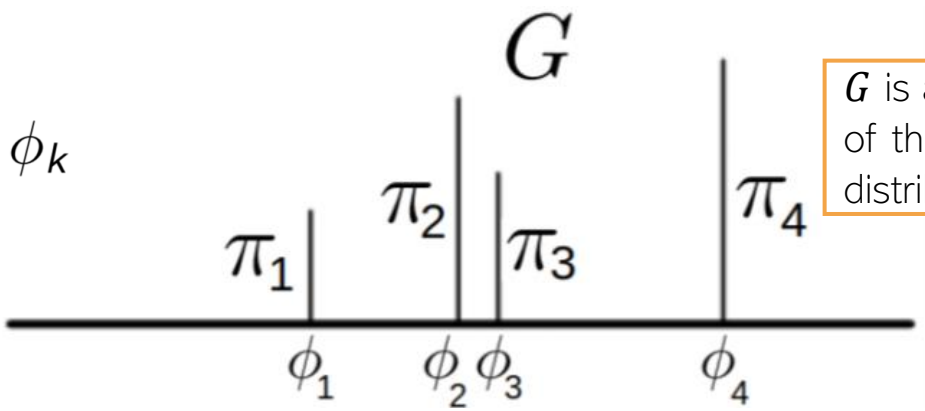
Plan today

- Wrap-up of nonparametric Bayesian models
  - Shrinkage-based construction
    - Stick-breaking process
    - Multiplicative gamma Process
- Simulation based inference
- A classic latent variable model: Topic Model (Latent Dirichlet Allocation)

Mixture Models

- Consider a finite mixture model with K components with params $(\mu_k, \Sigma_k)_{k=1}^K$

$$(\mu_k, \Sigma_k) = \phi_k$$



G is a representation of this mixture distribution

Defined by K locations or "atoms" with parameters $\{\phi_k\}_{k=1}^K$ with respective selection probabilities $\{\pi_k\}_{k=1}^K$

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

- In the finite case, we can assume $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$ and $\boldsymbol{\pi} \sim \mathbf{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$
- We can make it a nonparametric model by making $\boldsymbol{\pi}$ an infinite-dimensional vector

In practice, only a finite of these will have nonzero values, and others will shrink to very small (or zero), as we will see

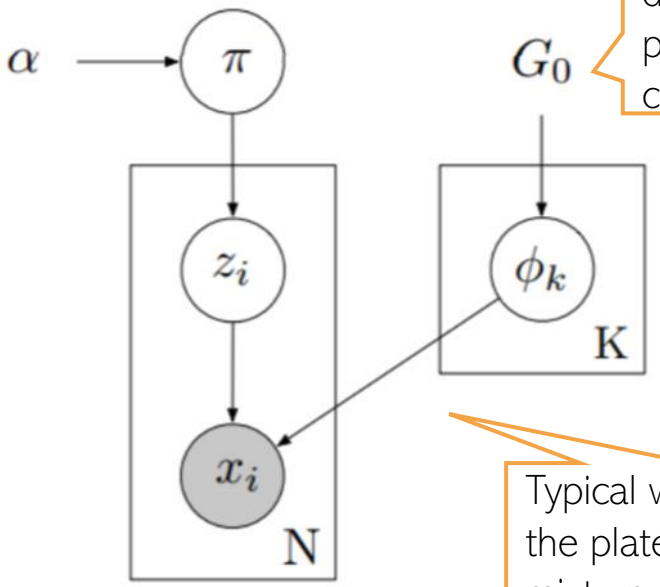
$$\pi_1, \pi_2, \pi_3, \ldots, \qquad \sum_{k=1}^{\infty} \pi_k = 1$$

Indeed. Called a "Dirichlet Process"

Related: "Stick-breaking Process"

- How to construct such a vector? Is there an infinite dimensional Dirichlet distribution?

Mixture Models: Two Equivalent Views



Prior (a.k.a. "base distribution" for the parameters of each mixture component

Example: G_0 can be NIW if each component is a Gaussian and $\phi_k = (\mu_k, \Sigma_k)$

Typical way of showing the plate notation of a mixture model

$$\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

$$\phi_k \sim G_0 \qquad k = 1,2,\ldots,K$$

$$z_i \sim \text{multinoulli}(\boldsymbol{\pi}) \qquad i = 1,2,\ldots,N$$

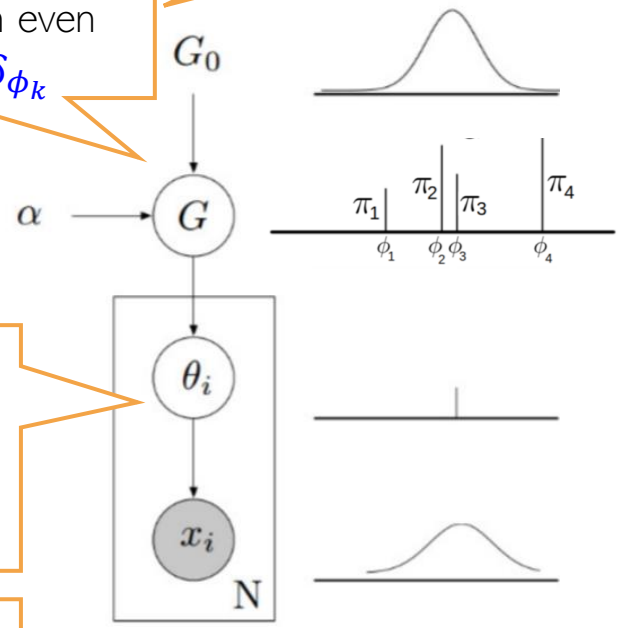
$$x_i \sim p(x|\phi_{z_i}) \qquad i = 1,2,\ldots,N$$

Similar representation even when $G = \sum_{k=1}^{\infty} \pi_k \, \delta_{\phi_k}$

But how to construct such a G distribution with potentially infinite components?

No explicit cluster ids; instead, θ_i denotes the param of the distribution which will generate x_i

Since G is discrete, there will at most be K distinct θ_i's, thereby achieving clustering



$$\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

$$\phi_k \sim G_0 \qquad k = 1,2,\ldots,K$$

$$G = \sum_{k=1}^{K} \pi_k \, \delta_{\phi_k}$$

$$\theta_i \sim G \qquad i = 1,2,\ldots,N$$

$$x_i \sim p(x|\theta_i) \qquad i = 1,2,\ldots,N$$

Stick-Breaking Process (Sethuraman'94)

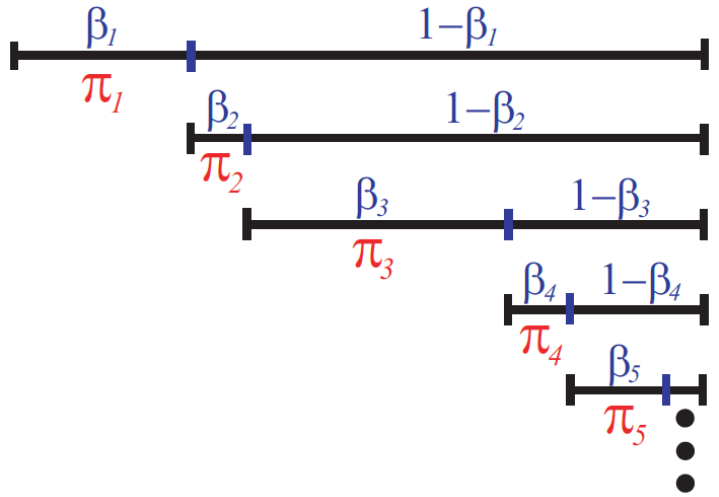
SBP gives us a way to construct infinite dimensional Dirichlet distribution known as the "Dirichlet Process"

- Recursively break a length 1 stick into two pieces
- Assume breaking point in each round is drawn from a Beta distribution

$$\beta_k \sim \text{Beta}(1, \alpha) \qquad k = 1, \ldots, \infty$$

$$\pi_1 = \beta_1$$

$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) \qquad k = 2, \ldots, \infty$$



- Can show that $\sum_{k=1}^{\infty} \pi_k - 1 \to 0$ which is what we want
- We can now have a "nonparametric/infinite" mixture distribution $G = \sum_{k=1}^{\infty} \pi_k \, \delta_{\phi_k}$
- "Location/atoms" $\boldsymbol{\phi_k}$ can be drawn from a "base" distr $\boldsymbol{G_0}$, say NIW if $\boldsymbol{\phi_k} = (\mu_k, \Sigma_k)$
- We basically replaced the Dirichlet prior on $\boldsymbol{\pi}$ by a Stick-Breaking Process (SBP) prior

Another NPBayes Prior: Multiplicative Gamma Process

- Consider the SVD-style probabilistic model with an *a priori* unbounded K

$$\mathbf{X} = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^{\mathsf{T}}$$

- Consider the following prior on each "singular values" $\boldsymbol{\lambda_k}$

$$\begin{aligned} \lambda_k &\sim \mathcal{N}(0, \tau_k^{-1}) \\ \tau_k &= \prod_{\ell=1}^{k} \delta_\ell \\ \delta_\ell &\sim \text{Gamma}(\alpha, 1) \quad \text{where } \alpha > 1 \end{aligned}$$

Precision keeps on getting larger and larger as k grows (thus variance keeps getting small and smaller)

Thus $\mathbb{E}[\delta_\ell] = \alpha$ (greater than 1 in expectation)

- In practice we can set K to be a sufficiently very large
  - Due to the shrinkage property, only a finite many $\boldsymbol{\lambda_k}$ will be nonzero
  - The nonzero $\boldsymbol{\lambda_k}$'s will dictate the effective K

Summary of NPBayes

- We saw some nonparametric Bayesian models (mainly used in unsup learning)
  - CRP/Dirichlet Process: For clustering problems
  - Multiplicative Gamma Process: For SVD-like matrix factorization
- Many applications of these models to solve a wide range of problems
- Also saw GP which is another example of a nonparametric Bayesian model
  - GPs are used for function approximation problems (both supervised and unsup. learning)
- These are only some of the examples of nonparametric Bayesian models
  - Many other such nonparametric Bayesian models for other problems in machine learning
  - "A tutorial on Bayesian nonparametric models" Gershman and Blei, 2011) is a nice survey
- Rich theory based on stochastic processes (beyond the scope of this course)
- Inspired other non-probabilistic algos, e.g., Using Dirichlet Process Mixture Model to get a K-means like clustering algorithm (DP-means) which doesn't require K

Simulation-based Inference

Simulation-based Inference

- Suppose we wish to compute the posterior $p(\theta|D)$
- However, suppose we can't compute the likelihood $p(D|\theta)$
  - Evaluation too expensive, or don't have explicit likelihood
- Simulation-based Inference (SBI) approximates $p(\theta|D)$ as follows
  - For $i = 1,2, \ldots S$
    - Draw a random $\theta^{(i)}$ the prior $p(\theta)$. Simulate a dataset $D^{(i)}$ from some simulator using $\theta^{(i)}$
    - Check how "similar" $D^{(i)}$ is to D. Define a suitable distance to measure this, e.g.,

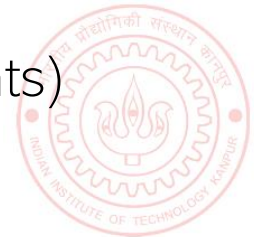
$$d_i = \left\| s\left(D^{(i)}\right) - s(D) \right\|$$

    - Define the weight of $\theta^{(i)}$ as inversely proportional to d_i, e.g., $w_i \propto \exp(-d_i)$
  - The approximate posterior is $\left\{w_i, \theta^{(i)}\right\}_{i=1}^{S}$
- The vanilla SBI/ABC can be inefficient in practice (most $\theta^{(i)}$'s may have low weights)
  - More efficient versions proposed in recent research, e.g, neural conditional density estimators
  - Check out this package for code and links to other methods: https://github.com/sbi-dev/sbi

SBI is also known as "Approximate Bayesian Computation" (ABC)

This simulator may be some domain-specific model of the data generation process (e.g., a physics engine, robotics/control simulator, etc)

Here $s(.)$ denotes a "summary statistics" which provides a summary of the dataset (e.g., its mean and covariance) which makes the comparison easier



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 $\boldsymbol{w}_d$
- Assuming a vocab of V unique words, $\boldsymbol{w}_d$ 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 $\boldsymbol{\phi}_k$ (sums to 1)
  - $\boldsymbol{\phi}_k$ 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

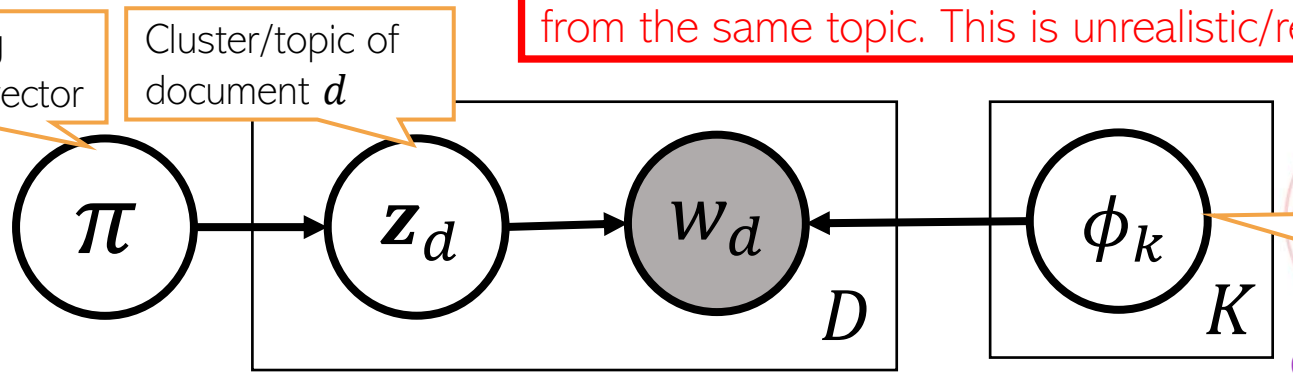
Each representing a "topic" (K topics)

Limitation: Each doc d belongs to a single cluster $\boldsymbol{z}_d$ and all words in a document assumed to be from the same topic. This is unrealistic/restrictive

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

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

Counts will sum to N_d

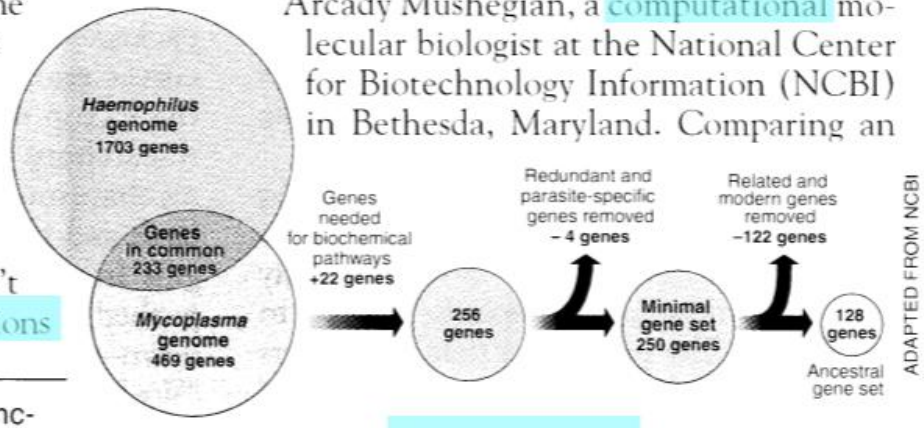


Documents can be about multiple topics

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 "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

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 $\boldsymbol{\alpha}$ ($K \times 1$ prob vector)
- Also assume doc-level topic mixing props $\boldsymbol{\theta}_d$ ($K \times 1$ prob vector)
- Instead of assuming a single cluster $\boldsymbol{z}_d$ for doc $\boldsymbol{d}$, cluster each word in it
  - $\boldsymbol{z}_{d,n} \in \{1,2,\ldots,K\}$ denotes the cluster/topic of word $\boldsymbol{w}_{d,n} \in \{1,2,\ldots,V\}$
- Can obtain the "average" clustering for doc $\boldsymbol{d}$ using $\boldsymbol{\theta}_d$ or $\bar{\boldsymbol{z}}_d = \frac{1}{N_d}\sum_{n=1}^{N_d} \boldsymbol{z}_{d,n}$

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

- The generative model is as follows

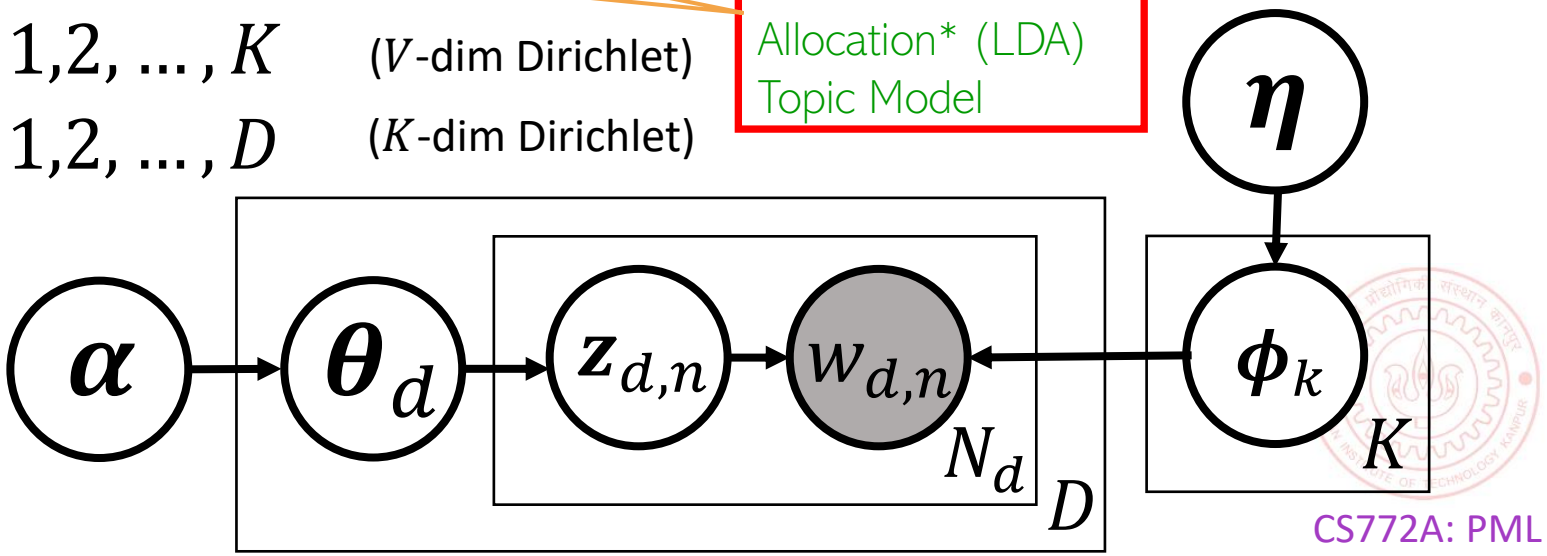
$$\phi_k \sim \text{Dirichlet}(\boldsymbol{\eta}) \quad k = 1,2,\ldots,K \quad (V\text{-dim Dirichlet})$$
$$\theta_d \sim \text{Dirichlet}(\boldsymbol{\alpha}) \quad d = 1,2,\ldots,D \quad (K\text{-dim Dirichlet})$$

Locally-conjugate. Easy Gibbs sampling, VI, etc

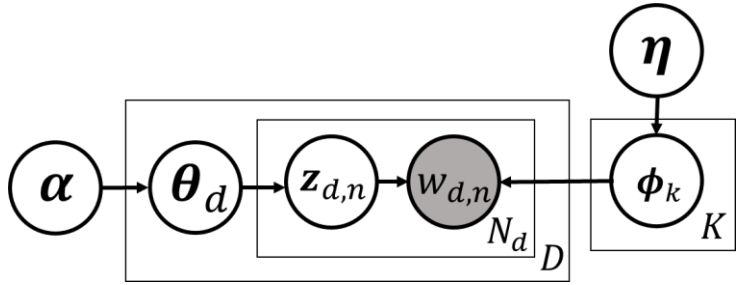
Somewhat similar to Dir-Mult PCA model

Latent Dirichlet Allocation* (LDA) Topic Model

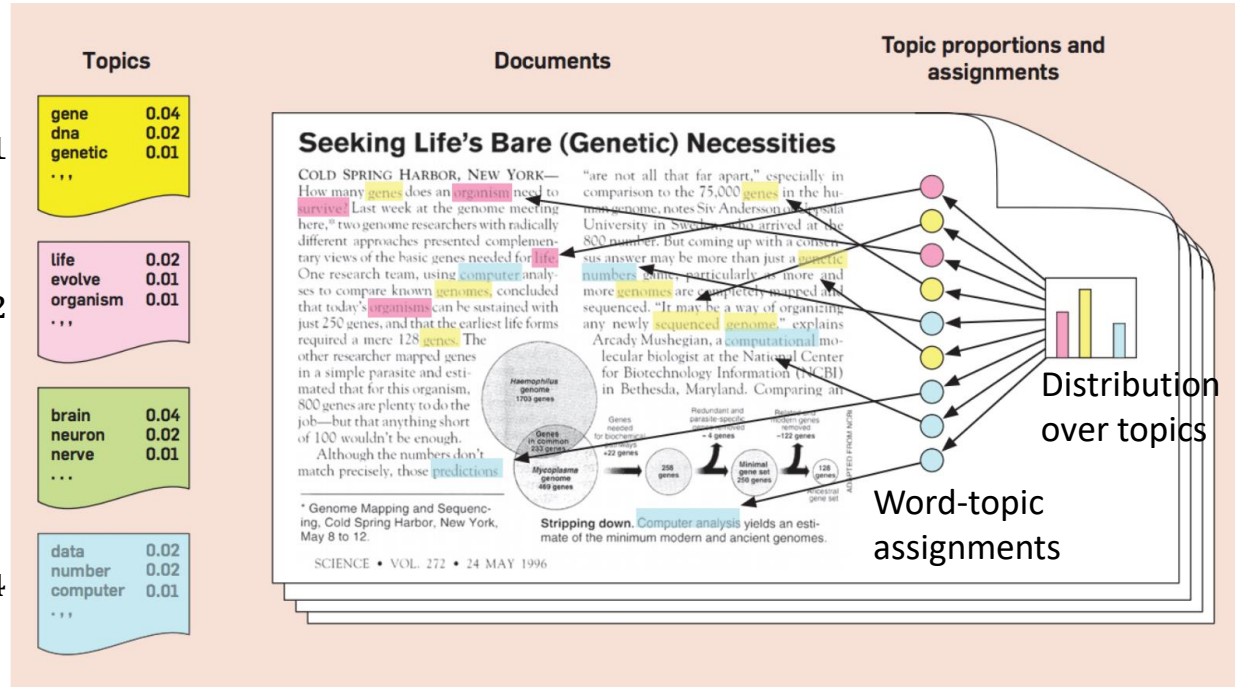
$$\boldsymbol{z}_{d,n} \sim \text{multinoulli}(\theta_d)$$
$$\boldsymbol{w}_{d,n} \sim \text{multinoulli}(\phi_{z_{d,n}})$$



Latent Dirichlet Allocation (LDA)

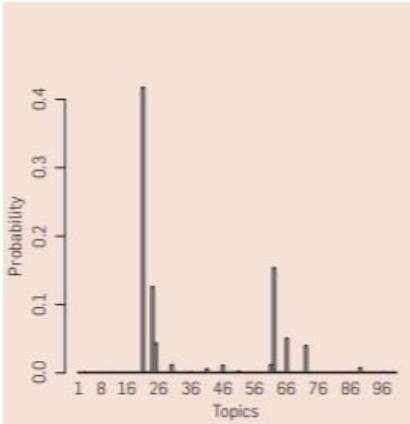


- A very widely used probabilistic model for text data
- Nice and easy insights into the text collection
  - Each $\phi_k = [\phi_{k1}, \ldots, \phi_{kV}]$ can be interpreted as topic (ϕ_{kv} = prob. of word v in topic k)
  - $\theta_d = [\theta_{d1}, \ldots, \theta_{dK}]$: how much each topic is present 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

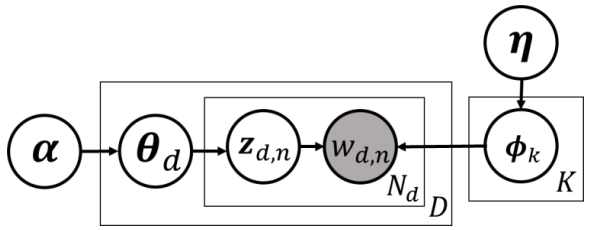
15 most frequent (most probable) words from four most prominent topics in this doc



"Genetics"	"Evolution"	"Disease"	"Computers"
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

Topic distribution for the document on left

LDA: Inference and Evaluation



- 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 $\boldsymbol{\theta_d}$ and $\boldsymbol{\phi_k}$ (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 $\boldsymbol{M}$ docs

Marginal likelihood of all words in the $\boldsymbol{d^{th}}$ 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)

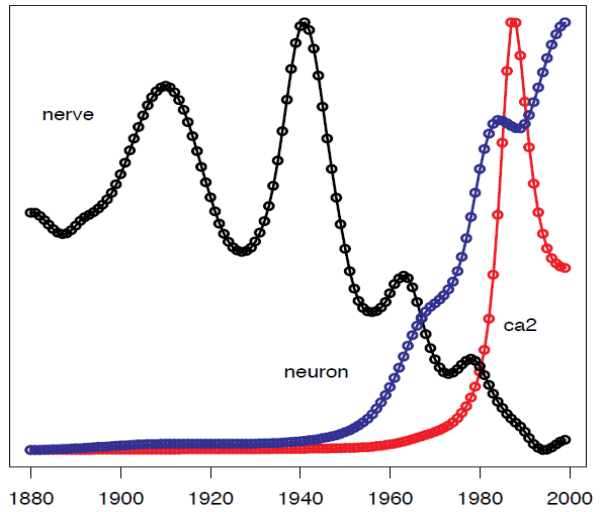
Assume a first-order Markov evolution for each topic w.r.t. time

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

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

1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
brain	movement	brain	movement	movement	stimulate	record	respons	response	respons	cell	cell	neuron
movement	eye	eye	brain	sound	muscle	nerve	record	stimulate	cell	neuron	channel	active
action	right	movement	sound	muscle	sound	stimulate	stimulate	condition	potential	response	neuron	brain
right	hand	right	nerve	active	movement	response	nerve	active	stimul	active	ca2	cell
eye	brain	left	active	nerve	response	muscle	muscle	potential	neuron	brain	active	fig
hand	left	hand	muscle	stimulate	nerve	electrode	active	stimulus	active	stimul	brain	response
left	action	nerve	left	fiber	frequency	active	frequency	nerve	nerve	muscle	receptor	channel
muscle	muscle	vision	eye	reaction	fiber	brain	electrode	subject	eye	system	muscle	receptor
nerve	sound	sound	right	brain	active	fiber	potential	eye	record	nerve	respons	synapse
sound	experiment	muscle	nervous	response	brain	potential	study		abstract	receptor	current	signal

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)

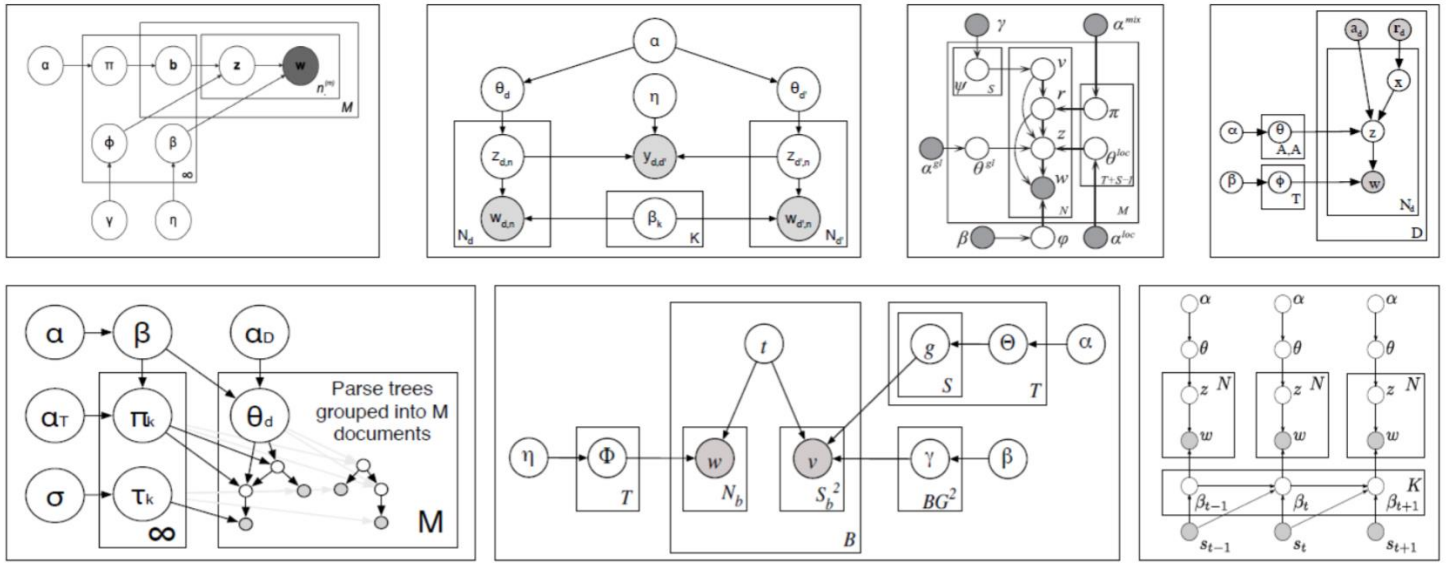
Fig courtesy: Dynamic Topic Models (Blei and Lafferty, 2006)

LDA Extensions (Contd)

Also: "Neural" Topic Models are popular nowadays ($\boldsymbol{z}$ to $\boldsymbol{x}$ mapping and vice-versa modeled via deep nets). Also, some topic models use pre-computed word-embeddings rather than one-hot Representation of each word

- 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 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*

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

$\mathbf{\Phi}$ $(V \times K)$ and $\mathbf{\Theta}$ $(K \times D)$ can be given any non-negative priors (Dirichlet/gamma)

This can be extended to "deep" matrix factorization** (modeling $\mathbf{\Theta}$ using many layers)

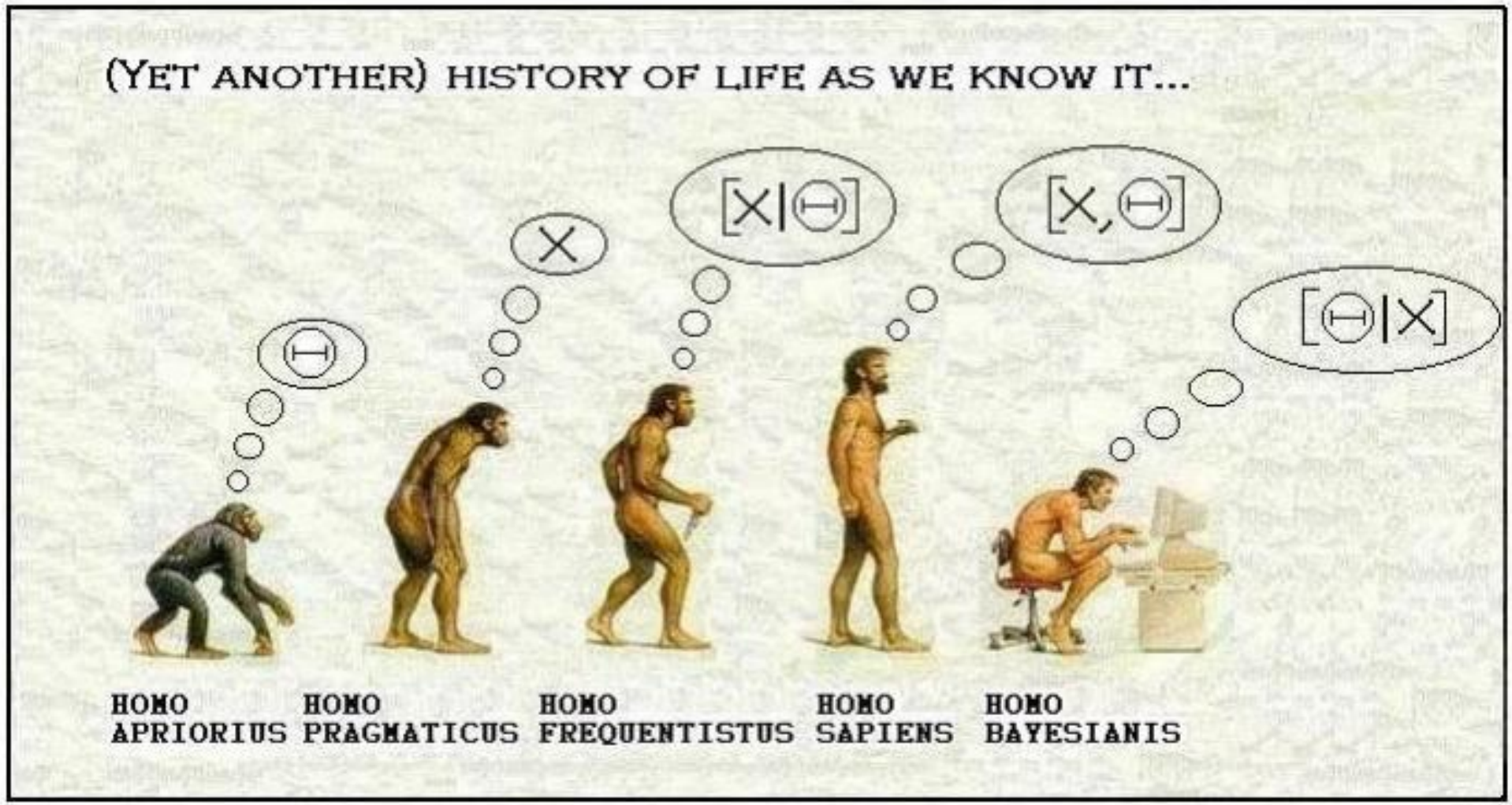
*Sec 4 and 5 of "Beta-Negative Binomial Process and Poisson Factor Analysis" (Zhou et al, 2012)

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

Conclusion

- Probabilistic modeling provides a natural way to think about models of data
- Many benefits as compared to non-probabilistic approaches
  - Easier to model and leverage uncertainty in data/parameters
  - Principle of marginalization while making prediction
  - Easier to encode prior knowledge about the problem (via prior/likelihood distributions)
  - Easier to handle missing data (by marginalizing it out if possible, or by treating as latent variable)
  - Easier to build complex models can be neatly combining/extending simpler probabilistic models
  - Easier to learn the "right model" (hyperparameter estimation, nonparametric Bayesian models)
- Bayesian approaches as well as single model based uncertainty
- Uncertainty is important but proper calibration of uncertainty is also important
- Fast-moving field, lots of recent advances on new models and inference methods

Thank You!


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS