Case Study 5: Mixed Membership Modeling

Clustering Documents Revisited, Latent Dirichlet Allocation

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
May 26th, 2015

Task 2: Cluster Documents

- **Then examined:**
  - Cluster documents based on topic

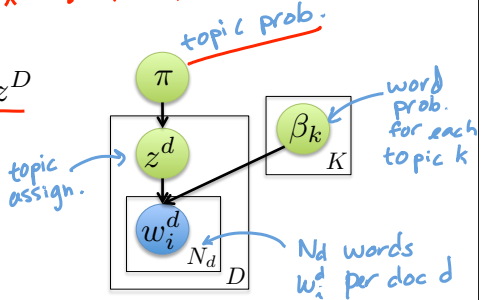
sports

world news

A Generative Model

- Documents: $x^1, \ldots, x^D$
- Associated topics: $z^1, \ldots, z^D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:

$x^d = \{w_1^d, \ldots, w_{N_d}^d\}$



Sample topic:
$z^d \sim \pi$

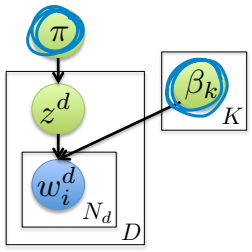
Sample words:
$w_i^d \mid z^d \sim \beta_{z^d}$ $i = 1, \ldots, N_d$
topic of doc d

Given topic $z^d = k$ for doc d, draw each word from β_k

Bayesian Document Model

- Model parameters $\pi, \{\beta_k\}$ unknown
- Bayesian approach

place priors on parameters



- Need distribution on pmf's

$\sum_{k=1}^{K} \pi_k = 1$ $\sum_{v=1}^{V} \beta_{kv} = 1$

What is a distribution on the simplex?
First, what is the simplex?

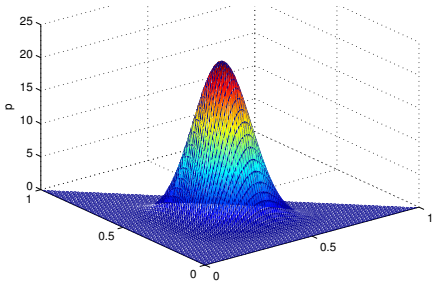
Dirichlet Distributions

- The Dirichlet distribution is defined on the simplex

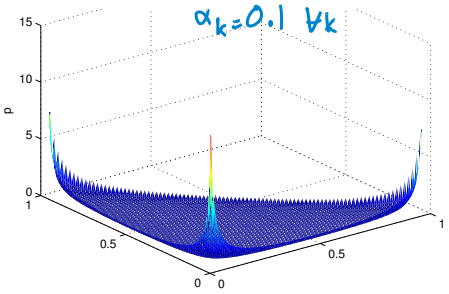
$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \Rightarrow \sum_{k=1}^{K} \pi_k = 1$ (K-dim pmf)

$$p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

$\alpha_k = 10 \ \forall k$



$\alpha_k = 0.1 \ \forall k$



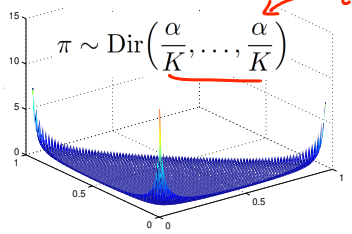
Moments:

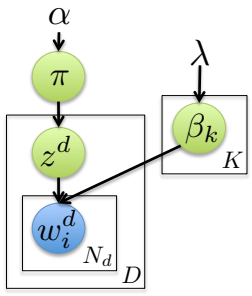
$$\mathbb{E}_\alpha[\pi_k] = \frac{\alpha_k}{\alpha_0}$$

$$\mathrm{Var}_\alpha[\pi_k] = \frac{K-1}{K^2(\alpha_0+1)}$$

Model Summary

- Prior on model parameters
  - E.g., symmetric Dirichlet for π (chose sym. Dir.)





  - Dirichlet prior for topic parameters $\beta_k \sim \mathrm{Dir}\left(\frac{\lambda}{V}, \ldots, \frac{\lambda}{V}\right)$ $k = 1, \ldots, K$

- Sample observations as

$$z^d \sim \pi \qquad d = 1, \ldots, D$$

$$w_i^d \mid z^d \sim \beta_{z^d} \qquad i = 1, \ldots, N_d$$

Posterior Inference via Sampling

- Iterate between sampling

actual obs.

"full conditionals"

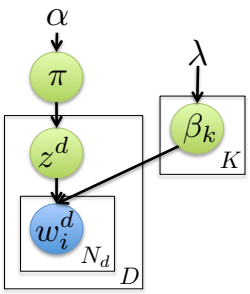
$\pi \sim p(\pi \mid \{z^d\}, \{\beta_k\}, \{w_i^d\})$

For $k = 1, \ldots, K$

$\beta_k \sim p(\beta_k \mid \pi, \{z^d\}, \{w_i^d\})$

For $d = 1, \ldots, D$

$z^d \sim p(z^d \mid \pi, \{\beta_k\}, \{w_i^d\})$

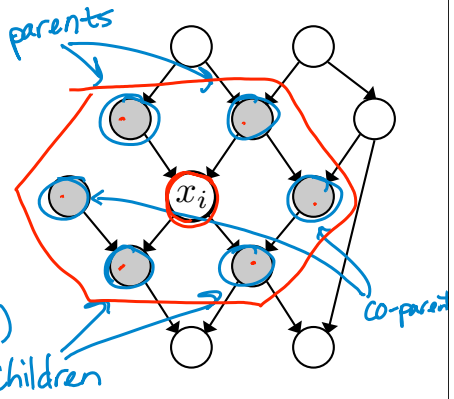


- What form do these complete conditionals take?
  - First a look at statements of conditional independence in directed graphical models

Markov Blanket

- A node is conditionally independent of all other nodes in the graph given its Markov blanket

Markov blanket of x_i = - all parents - all children - all coparents



- Gibbs sampling iterates between full conditionals

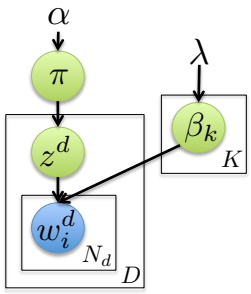
$x_i \sim p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_N)$

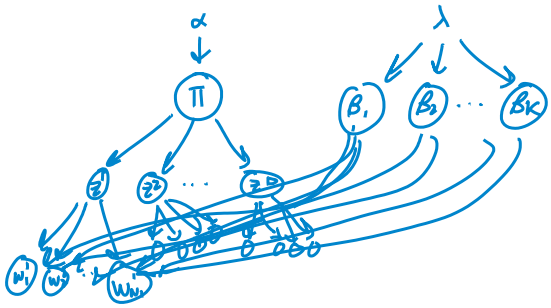
→ simplify to

$x_i \sim p(x_i \mid MB(x_i))$

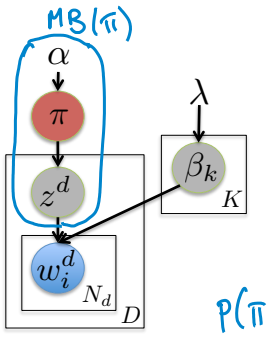
Unplated Document Model

- Recall that the plate notation is really indicating





Complete Conditional for π



- Recall conjugate Dirichlet prior

$$\pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \quad p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}$$

replace w/ $N_k + \alpha_k$

- Likelihood: $z^d \sim \pi \rightarrow \prod_{d=1}^{D} p(z^d \mid \pi)$
- Dirichlet posterior
  - Count occurrences of $N_k = |\{z^d : z^d = k\}|$
  - Then,

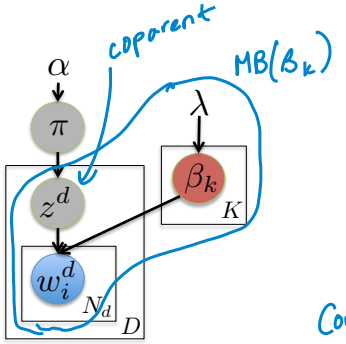
$$p(\pi \mid \{z^d\}, \alpha) \propto \prod_{d=1}^{D} p(z^d \mid \pi)\, p(\pi \mid \alpha)$$

$$\propto \prod_{k=1}^{K} \left[\prod_{d : z^d = k} \pi_k \right] \cdot \pi_k^{\alpha_k - 1}$$

$$\propto \prod_k \pi_k^{N_k} \pi_k^{\alpha_k - 1} = \prod_k \pi_k^{N_k + \alpha_k - 1} = \text{Dir}(N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

  - Conjugacy: **Posterior** has same form as **prior**

Complete Conditional for β_k



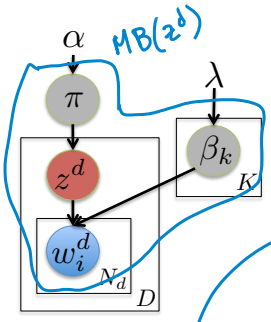
- Again, Dirichlet prior
  $\beta_k \sim Dir(\lambda_1, \ldots, \lambda_V)$
- Consider docs *d* such that $z^d = k$
  - For these observations, $w_i^d \sim \beta_k$
  - Do any other docs depend on β_k? NO
- Then,

Count $m_{v,k} = |\{w_i^d : w_i^d = v, \forall d \text{ s.t. } z^d = k\}|$

$\beta_k \sim Dir(m_{1,k}+\lambda_1, \ldots, m_{V,k}+\lambda_V)$

  - Again, **posterior** has same form as **prior**

Complete Conditional for z^d



- We have $z^d \sim \pi$ "prior"

  $w_i^d \mid z^d, \{\beta_k\} \sim \beta_{z^d}$ "likelihood"
- Calculate the posterior for each value of z^d ("responsibility" of each topic to the doc):

$$r_{dk} = p(z^d = k \mid \{w_i^d\}, \pi, \beta) = \frac{\pi_k p(\{w_i^d\} \mid \beta_k)}{\sum_j \pi_j p(\{w_i^d\} \mid \beta_j)}$$

full cond.

z^d is a discrete r.v.

$\prod_{i=1}^{N_d} \beta_{k,w_i^d}$

- Sample each cluster indicator as

$r_d = [r_{d1}, \ldots, r_{dK}]$

$z^d \sim r_d$

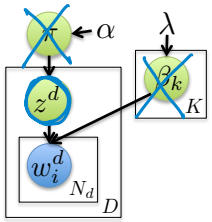
responsibilities

Collapsed Gibbs Sampler

- In conjugate models, can analytically marginalize some variables and only sample remaining

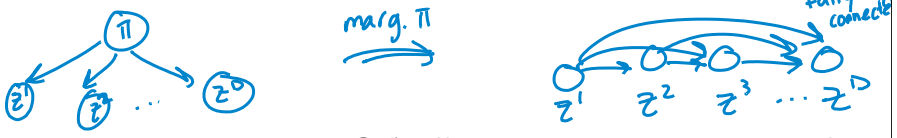
For $d=1,\dots,D$

$z^d \sim p(z^d \mid z^1, \dots, z^{d-1}, z^{d+1}, \dots, z^D)$



- Can improve efficiency if marginalized variables are high-dim
  - Reduced dimension of search space
  - But, often introduces dependences!

opportunities for parallelism are lost !!



Collapsed Sampler Full Conditional

full joint dist.

$$p(\cdot) = p(\pi \mid \alpha) \prod_{d=1}^{D} p(z^d \mid \pi) \left(\prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} \prod_{i=1}^{N_d} p(w_i^d \mid z^d, \beta) \right)$$

$\int_\pi$ $\int_{\beta_k}$ $\prod_{d:z^d=k} \prod_{i=1}^{N_d} p(w_i^d \mid \beta_k)$

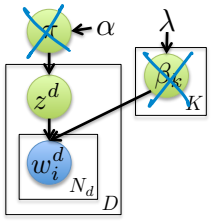
- Derivation

$$p(z^d = k \mid z_{\setminus d}, \{w_i^d\}, \alpha, \lambda) \propto \int_\pi \int_{\beta_1} \cdots \int_{\beta_K} p(\cdot)$$

$\propto p(z^d=k \mid z_{\setminus d}, \alpha)\; p(\{w_i^d\} \mid \{w_i^c : z^c=k, c \neq d\})$

"prior" — depends on # of $z^c=k$ (N_k)

"likelihood" — words in docs assigned to topic k not inc. d

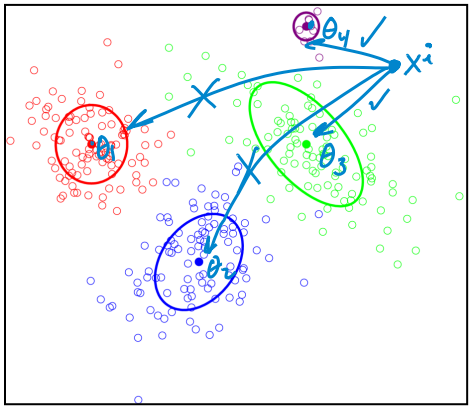


Collapsed Sampler Intuition (MoG)

- Previously, $p(z^i = k \mid x^i, \pi, \theta) \propto \pi_k p(x^i \mid \theta_k)$ ← $\{\mu_k, \Sigma_k\}$
- If you're not told π, θ_k

"prior" Approx. π by counts associated with each cluster

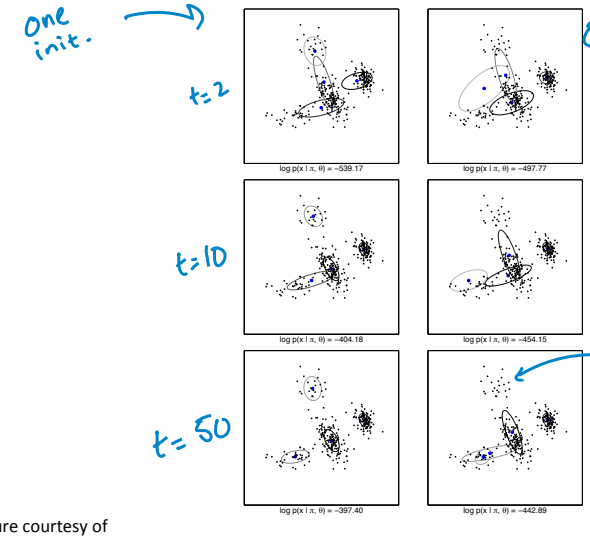
"likelihood" Approx θ_k by obs. already assigned to those clusters



Example – Uncollapsed Results

sampling $\pi, \{\theta_k\}$, $\{z^i\}$

one init. → another



given $z^1, \ldots, z^N$ low post. prob. of drawing θ_k here

will eventually happen, but maybe not in our lifetime

Figure courtesy of Erik Sudderth

Example – Collapsed Results

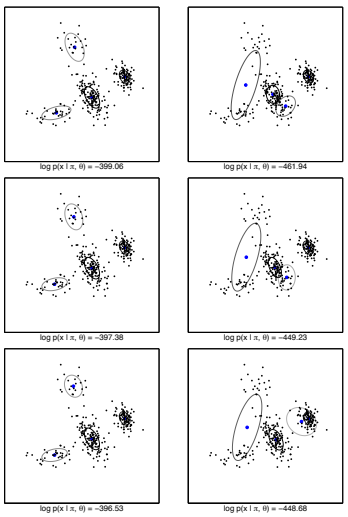


Figure courtesy of Erik Sudderth

Comparing Collapsed vs. Uncollapsed

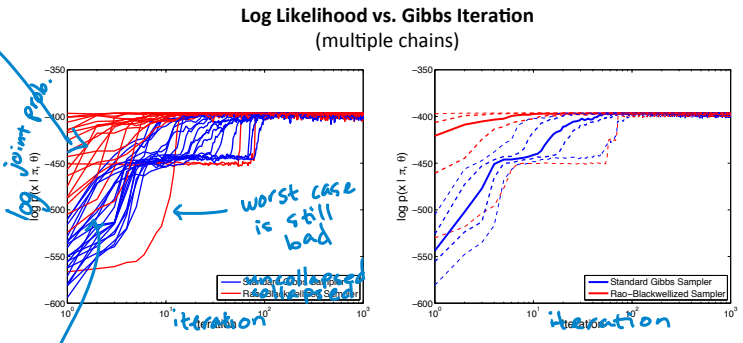
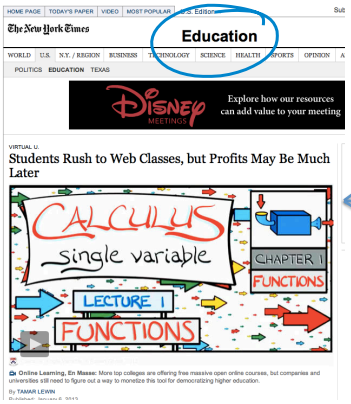
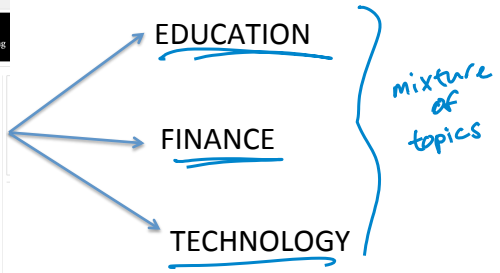


Figure courtesy of Erik Sudderth

Task 3: Mixed Membership Models

- **Now:** Document may belong to multiple clusters

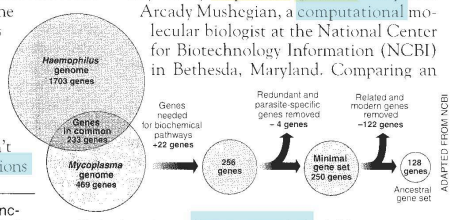




Latent Dirichlet Allocation (LDA)

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

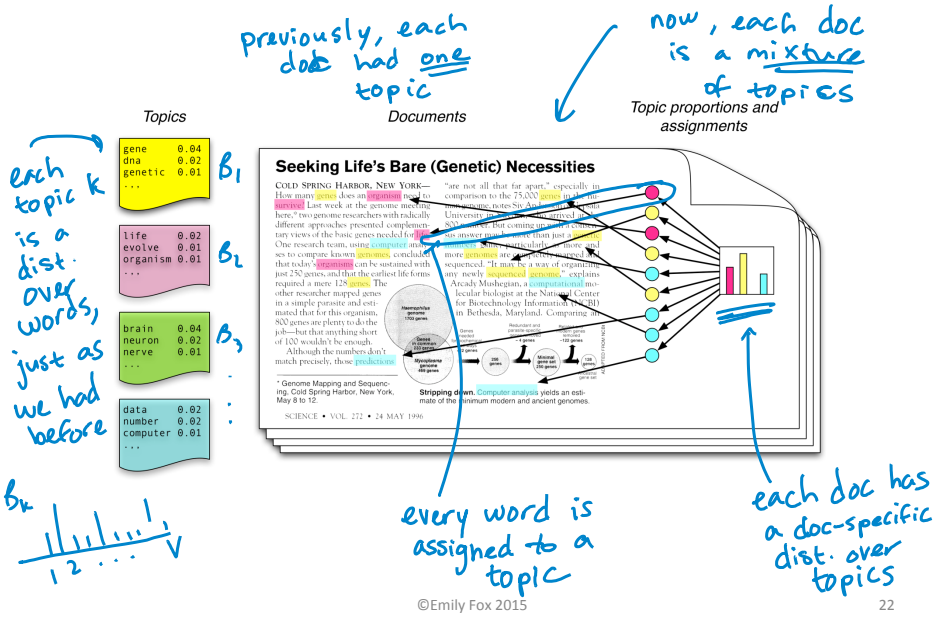


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

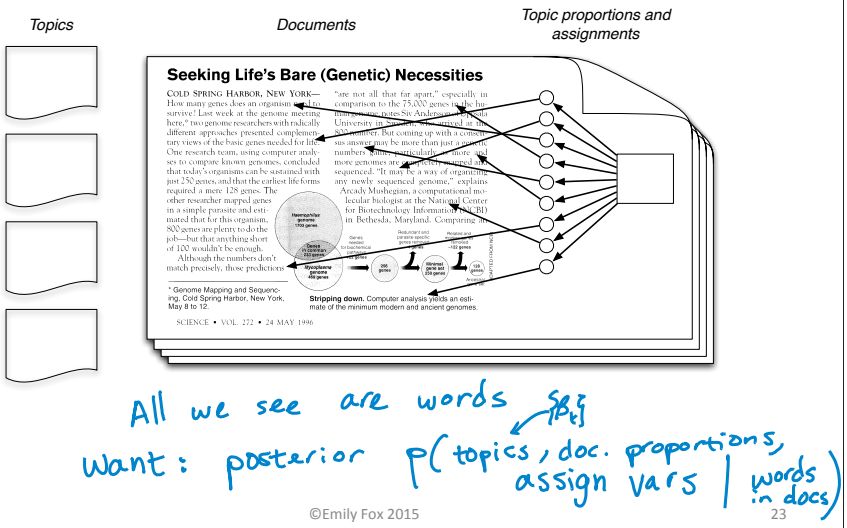
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

words from different topics

Latent Dirichlet Allocation (LDA)



Latent Dirichlet Allocation (LDA)

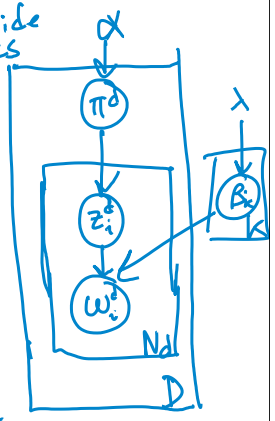


LDA Generative Model

- Observations: $w_1^d, \ldots, w_{N_d}^d$ $d=1,\ldots,D$
- Associated topics: $z_1^d, \ldots, z_{N_d}^d$ ← assign var. per word
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$ (doc-specific topic weights; corpus-wide topics)
- Generative model:

$z_i^d \sim \pi^d \quad d=1,\ldots,D$

$w_i^d \mid z_i^d \sim \beta_{z_i^d} \quad i=1,\ldots,N_d$

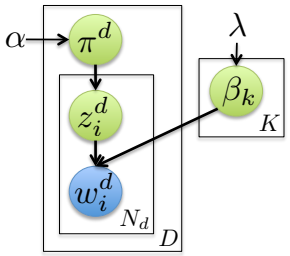


Priors:

$\pi^d \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \quad d=1,\ldots,D$

$\beta_k \sim \text{Dir}(\lambda_1, \ldots, \lambda_V) \quad k=1,\ldots,K$

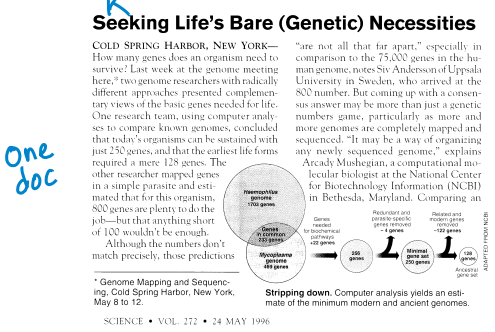
LDA Joint Probability



$$p(\cdot) = \prod_{k=1}^{K} p(\beta_k \mid \lambda) \prod_{d=1}^{D} p(\pi^d \mid \alpha) \left(\prod_{i=1}^{N_d} p(z_i^d \mid \pi^d) p(w_i^d \mid z_i^d, \beta) \right)$$

Example Inference – Topic Weights

- **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



Seeking Life's Bare (Genetic) Necessities

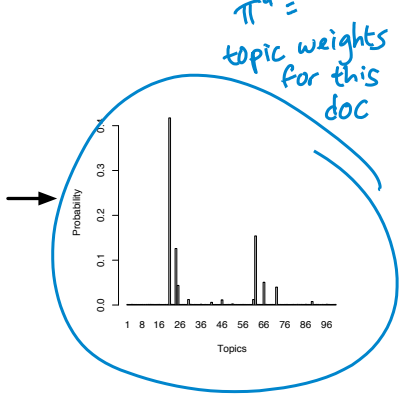
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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



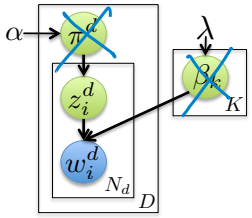
Example Inference – Topic Words

topic 1	topic 2	topic 3	...
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

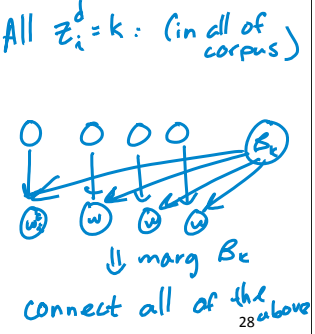
Collapsed LDA Sampling

- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions

$$p(z_i^d = k \mid z_{\setminus id}, \{w_i^d\}, \alpha, \lambda)$$
$$\propto p(z_i^d = k \mid z_{\setminus id}, \alpha) p(w_i^d | z_i^d = k, z_{\setminus id}, w_{\setminus id}, \lambda)$$

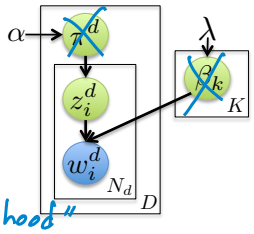


- Unplate to see dependencies induced



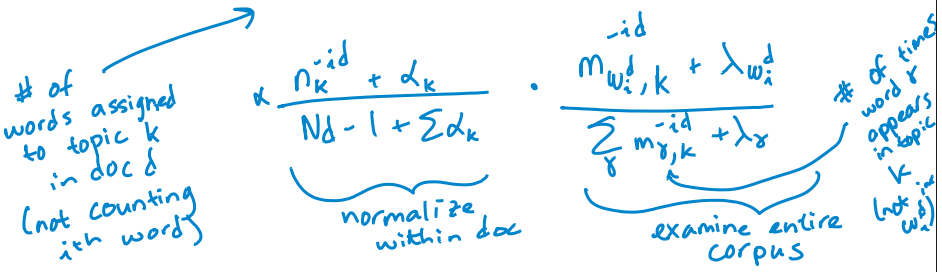
Collapsed LDA Sampling

- Sample topic indicators for each word
  - Algorithm:



$$p(z_i^d = k \mid z_{\setminus id}, \{w_i^d\}, \alpha, \lambda)$$
$$\propto p(z_i^d = k \mid \{z_j^d, j \neq i\}, \alpha) p(w_i^d | \{w_j^c : z_j^c = k, (j,c) \neq (i,d)\}, \lambda)$$

"prior" "likelihood"



What you need to know…

- Bayesian specification of document clustering model
- Rules of conditional and unconditional independence in directed graphical models (Bayes nets)
  - Bayes' ball
  - Markov blanket
- Gibbs sampling for Bayesian document model
- Latent Dirichlet allocation (LDA)
  - Motivation and generative model specification
  - Collapsed Gibbs sampler

Reading

- **Mixed Membership Models: KM Sec. 27.3**
  - Basic LDA:
    Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.
  - Introduction:
    Blei, David M. "Probabilistic topic models." Communications of the ACM, vol. 55, no. 4 (2012): 77-84.
  - Sampling:
    Griffith, Thomas L. and Mark Steyvers. "Finding scientific topics." Proceedings of the National Academy of Sciences of the United States of America, Volume: 101, Supplement: 1 (2004): Pages: 5228-5235

Acknowledgements

- Thanks to Dave Blei for some material in this lecture relating to LDA