Case Study 5: Mixed Membership Modeling

Clustering Documents Revisited, Latent Dirichlet Allocation

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox

March 6th, 2014

Task 2: Cluster Documents

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

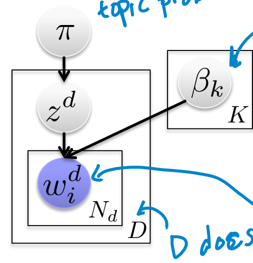
sports

world news

A Generative Model

$\{w_1^D, \dots, w_{N_D}^D\}$ bag of words

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



topic prob.

for topic k, prob. of words in that topic

N_d words in doc d.

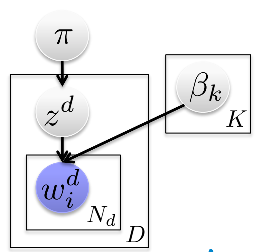
D docs

$z^d \sim \pi$ generate topic

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

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

Model In Pictures



- Mixture weights (on topics)

π — health, computing, ..., world news — K topics

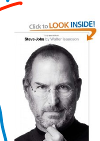
$z^d = 2$

$z^d = K$

- Topic distributions (on words)

health β_1 — apple, run, doctor, laptop — V

world news β_K — obama, israel — V



$z^d = 5$

sports

β_2 — computer, distributed

$w_1^d, w_2^d, w_3^d, w_4^d$ words in doc d

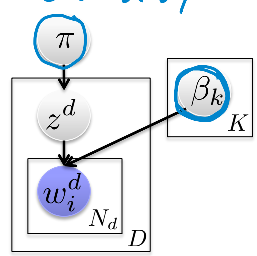
β_K — obama, israel

- For each document,

$$z^d \sim \pi$$
$$w_i^d \mid z^d \sim \beta_{z^d}$$

Bayesian Document Model

- Model parameters $\underline{\pi, \{\beta_k\}}$ unknown ← can use EM as in case study 2



- Bayesian approach

  place priors on parameters

- Need distribution on pmf's

  $\sum_{k=1}^{K} \pi_k = 1$ $\quad \sum_{v=1}^{V} \beta_{kv} = 1$ ← π, β_k live on the simplex

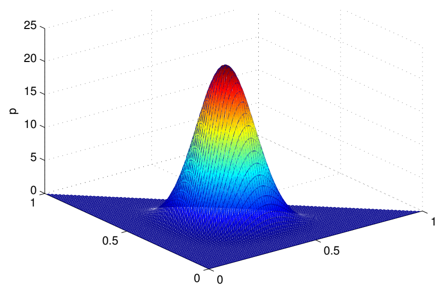
  (1) What is the simplex?
  (2) What is a distribution on the simplex?

 5

Dirichlet Distributions ← distribution on the simplex

- The Dirichlet distribution is defined on the simplex

$\alpha_k = 10 \ \forall k$

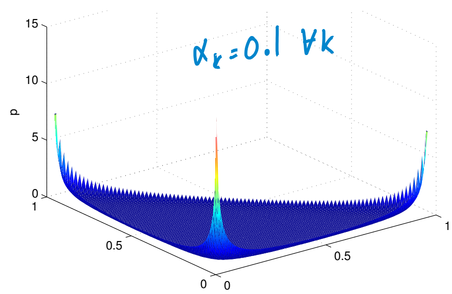


$\pi \sim \underline{Dir(\alpha_1, \dots, \alpha_K)}$

$\Rightarrow \sum \pi_k = 1$ and $\pi_k \geq 0 \ \forall k$

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

$\alpha_k = 0.1 \ \forall k$



Moments: $\mathbb{E}_\alpha[\pi_k] = \dfrac{\alpha_k}{\alpha_0}$

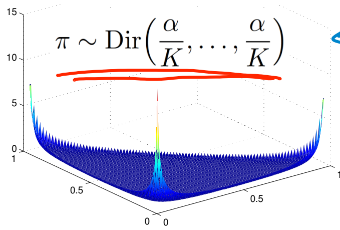
$$\mathrm{Var}_\alpha[\pi_k] = \frac{K-1}{K^2(\alpha_0 + 1)}$$

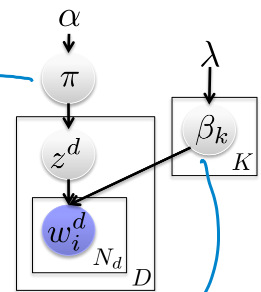
 6

Model Summary

- Prior on model parameters
  - E.g., symmetric Dirichlet for π

$$\pi \sim \mathrm{Dir}\left(\frac{\alpha}{K},\dots,\frac{\alpha}{K}\right)$$





  - Dirichlet prior for topic parameters $\beta_k \sim \mathrm{Dir}(\lambda_1,\dots,\lambda_V)$ $\quad k=1,\dots,K$

- Sample observations as

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

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

©Emily Fox 2014 7

Posterior Inference via Sampling

- Iterate between sampling ← actual obs.

full conditional

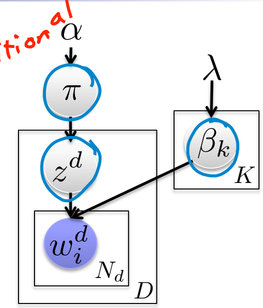
① $\pi \sim p(\pi \mid \{z^d\}, \{\beta_k\}, \{w_i^d\})$

② For $k=1,\dots,K$

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

③ For $d=1,\dots,D$

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

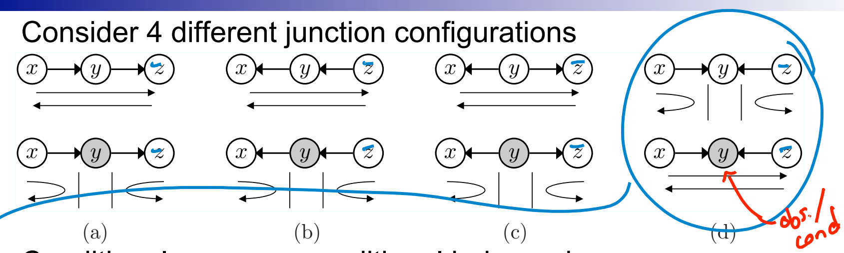


- What form do these complete conditionals take?
  - First a look at statements of conditional independence in directed graphical models

©Emily Fox 2014 8

Conditional Independence in Bayes Nets

- Consider 4 different junction configurations



- Conditional versus unconditional independence:

$$p(x,y,z) = p(x)p(z)p(y|x,z) \overset{\text{int. over } y}{\Rightarrow} p(x,z) = p(x)p(z) \Rightarrow x \perp\!\!\!\perp z$$

$$p(x,z|y) \propto p(x,y,z) = p(x)p(z)p(y|x,z)$$
$$\neq p(x|y)\,p(z|y) \leftarrow x \not\perp\!\!\!\perp z \mid y$$

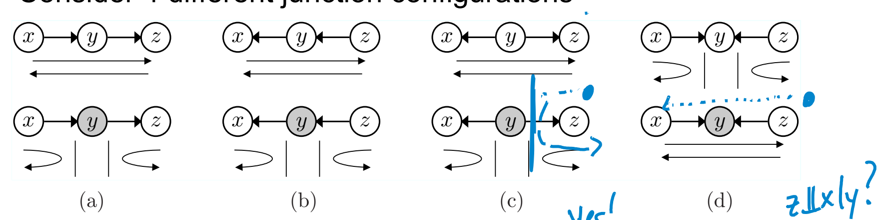
"explaining away": x = earthquake z = burglar , y = car alarm

ind. apriori

If alarm ($y=1$) , an increase in earthquake $p(x|y)$, means $p(z|y)$ lower

Bayes Ball Algorithm

- Consider 4 different junction configurations



- Bayes ball algorithm

Start ball at one end or other.

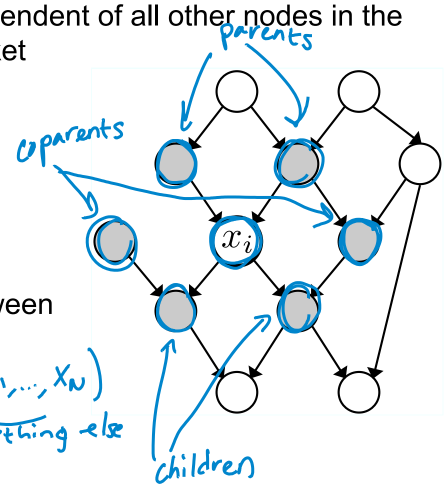
If ball passes to a node (straight arrows) then, not cond./marg. ind.

If ball bounces back (walls + curved arrows), the nodes are cond./marg. ind.

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

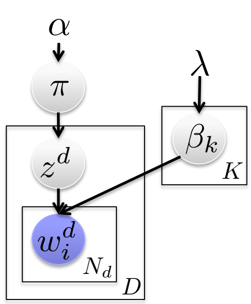
$$X_i \sim p(x_i \mid \underbrace{x_1,\dots,x_{i-1},x_{i+1},\dots,x_N}_{\text{everything else}})$$

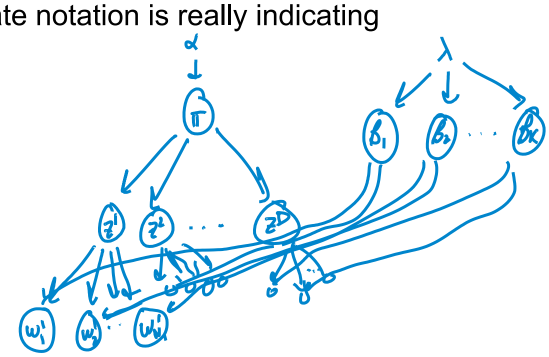
→ simplify to

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

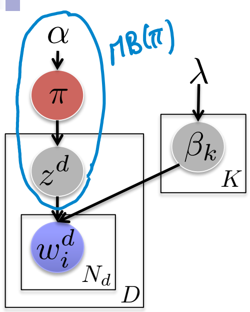
Unplated Document Model

- Recall that the plate notation is really indicating





Complete Conditional for π



- Recall conjugate Dirichlet prior

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

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

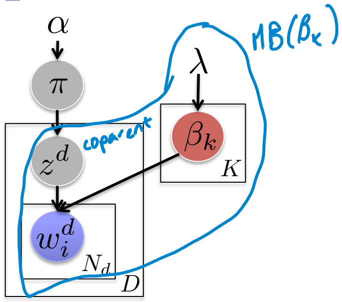
$$\underbrace{p(\pi \mid \{z^d\}, \alpha)}_{\text{full cond. for } \pi} \propto \prod_{d=1}^{D} p(z^d \mid \pi) p(\pi \mid \alpha)$$

$$\propto \prod_{k=1}^{K} \Big(\prod_{d: z^d = k} \underbrace{\pi_k}_{p(z^d = k \mid \pi)} \Big) \cdot \pi_k^{\alpha_k - 1}$$

$$\propto \prod_{k=1}^{K} \pi_k^{N_k + \alpha_k - 1} = \mathrm{Dir}(N_1 + \alpha_1, \ldots, N_K + \alpha_K)$$

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

Complete Conditional for β_k

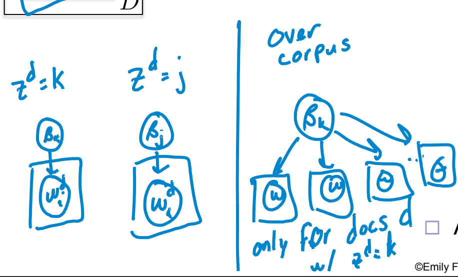


- Again, Dirichlet prior

$$\beta_k \sim \mathrm{Dir}(\lambda_1, \ldots, \lambda_V)$$

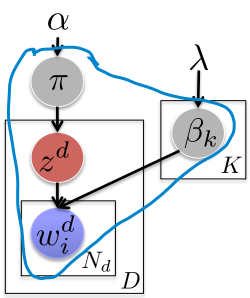
- Consider docs d such that $z^d = k$
  - For these observations, $w_i^d \sim \beta_k$
  - Do any other docs depend on β_k? NO
- Then, count $m_{v,k} = |\{w_i^d : w_i^d = v \;\; \forall d \text{ s.t. } z^d = k\}|$ (word v, topic k)

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



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

Complete Conditional for z^d



MB(z^d)

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

$$w_i^d \mid z^d, \{\beta_k\} \sim \beta_{z^d} \quad \text{"likelihood"}$$

- Calculate the posterior for each value of z^d ("responsibility" of each topic to the doc):

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

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

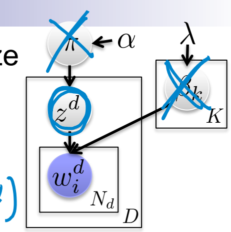
- Sample each cluster indicator as

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

$z^d \sim r_d$

Collapsed Gibbs Sampler

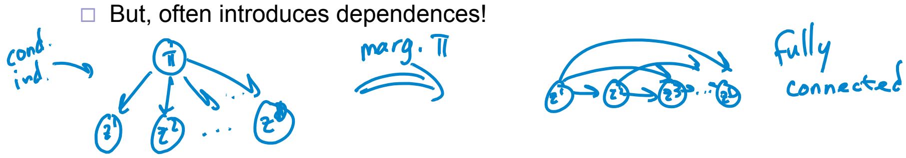
- In conjugate models, can analytically marginalize some variables and only sample remaining



for $d = 1, \ldots, D$

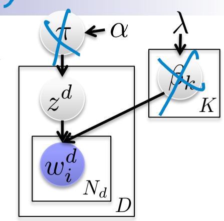
$$z^d \sim p(z^d \mid \underbrace{z^1, \ldots, z^{d-1}, z^{d+1}, \ldots, z^D}_{z_{\setminus d}}, \{w_i^d\}, \lambda, \alpha)$$

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



Collapsed Sampler Full Conditional

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



- Derivation

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

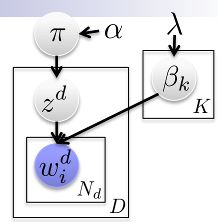
full cond.

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

"prior" "likelihood"

Collapsed Sampler Full Conditional

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



$\pi \sim Dir(\alpha)$
$z^d \sim \pi$

$\Downarrow$

$$p(z^d = k \mid z_{\setminus d}, \alpha) = \frac{N_k^{-d} + \alpha_k}{D - 1 + \sum \alpha_k}$$

# of docs assigned to topic k not counting doc d

$\beta_k \sim Dir(\lambda)$
$w_i^d \mid z^d = k \sim \beta_k$

$\Downarrow$

$$= \prod_{i=1}^{N_d} \frac{m_{w_i^d, k}^{-d} + \lambda_{w_i^d}}{\sum_v \left(m_{v,k}^{-d} + \lambda_v \right)}$$

# of words assigned to topic k taking value w_i^d (not inc doc d)

Collapsed Sampler Intuition (MoG)

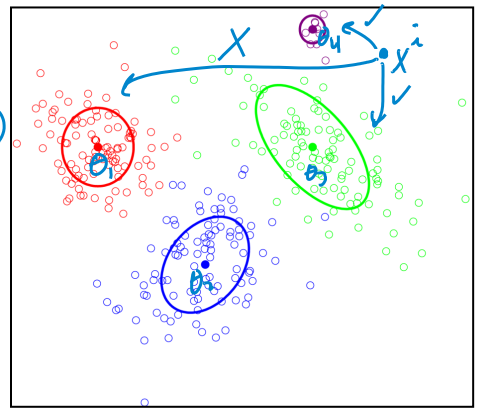
- Previously, $p(z^i = k \mid x^i, \pi, \theta) \propto \pi_k p(x^i \mid \theta_k)$

$\{\mu_k, \Sigma_k\}$

- If you're not told π, θ_k

"prior" Approx π by counts of occupancy of each cluster (plus prior)

"live" Approx θ_k based on obs. already assigned to cluster k

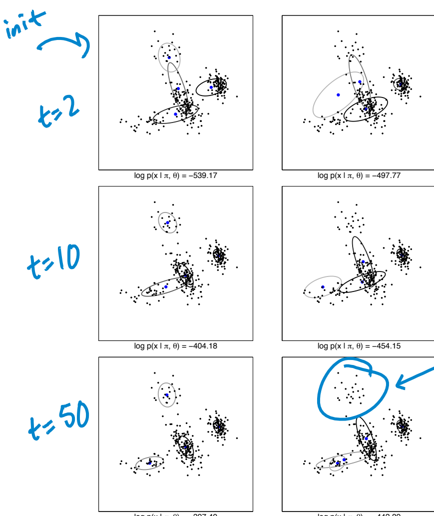


Example – Uncollapsed Results

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

one init

another

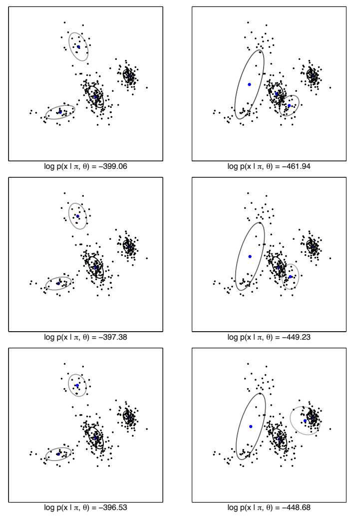


given $z^1, \dots, z^N$, low post prob. of drawing θ_k here

Figure courtesy of Erik Sudderth

will eventually happen, but maybe not in lifetime

Example – Collapsed Results



worst case is still bad

seq. sampling z^i given $z_{\backslash i}$, dependencies can make global moves really hard

trade-off

Figure courtesy of
Erik Sudderth

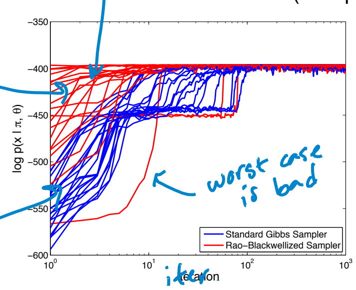
Comparing Collapsed vs. Uncollapsed

in general, red has faster "burn-in"

Log Likelihood vs. Gibbs Iteration
(multiple chains)

red: collapsed

blue = uncollapsed



worst case is bad

iter

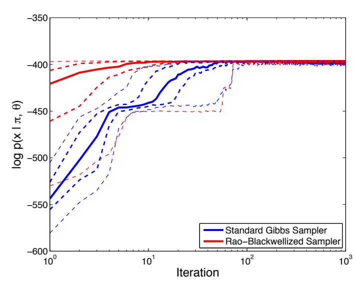


Figure courtesy of
Erik Sudderth

Task 3: Mixed Membership Models

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

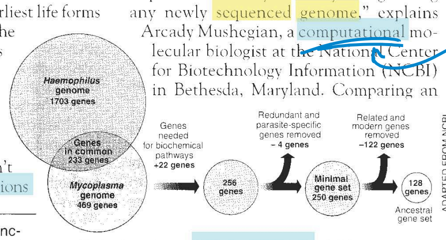
EDUCATION

FINANCE

TECHNOLOGY

mixture of topics

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

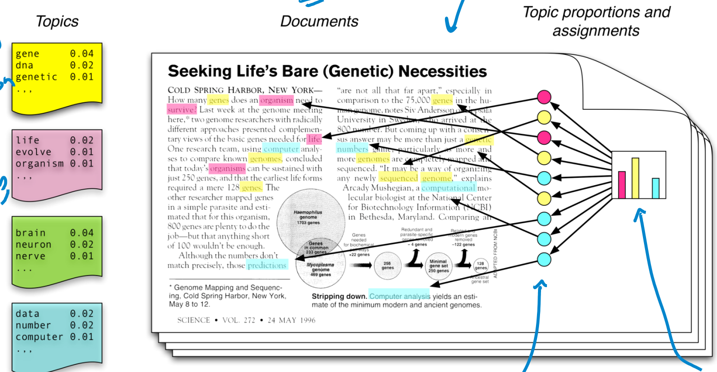
Latent Dirichlet Allocation (LDA)

each topic k is a distribution over words in vocab β_k, just as before

β_k (bar plot over $1\ 2 \cdots V$)

previously, each doc had one topic

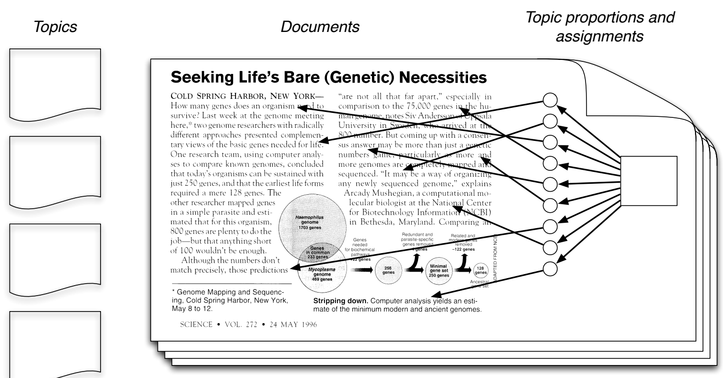
now, each is a mixture of topics



every word is assigned to a topic

each doc has its own prevalence of topics in that doc

Latent Dirichlet Allocation (LDA)



All we see are words — β_k's

Want: posterior p(topics, doc prop. of topics, assign. vars. | words in docs)

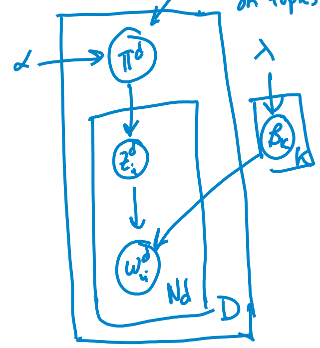
LDA Generative Model

- Observations: $w_1^d, \dots, w_{N_d}^d$
- Associated topics: $z_1^d, \dots, z_{N_d}^d$ ← topic per word in doc d
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model:

$z_i^d \sim \pi^d$ $\quad d=1,\dots,D$
$\quad i=1,\dots,N_d$

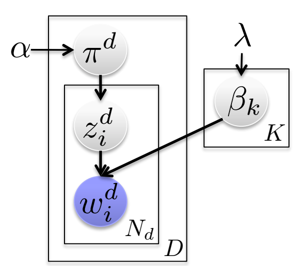
$w_i^d \mid z_i^d \sim \beta_{z_i^d}$

doc-specific weighting on topics



Priors:
$\pi^d \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K)$ $\quad d=1,\dots,D$
$\beta_k \sim \mathrm{Dir}(\lambda_1, \dots, \lambda_V)$ $\quad k=1,\dots,K$

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 ← D
  - 11M words ← V
  - 20K unique terms (stop words and rare words removed)
- **Model:** 100-topic LDA model

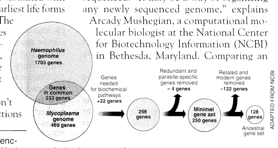
π^d = topic weights for this doc

one doc

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.

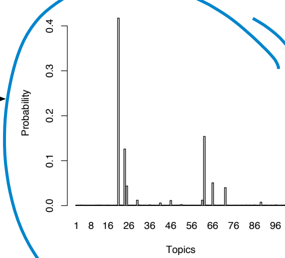
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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



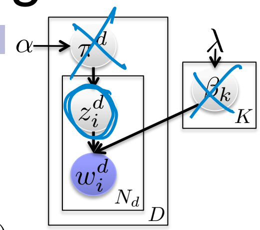
Example Inference – Topic Words

highest prob. words associated w/ each topic

topic 1	topic 2	. . .	
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

shared throughout corpus

Collapsed LDA Sampling

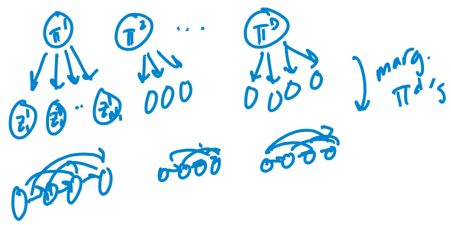


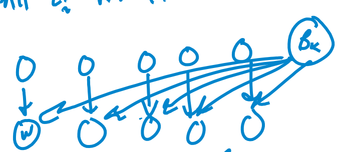
- Marginalize parameters
  - Document-specific topic weights
  - Corpus-wide topic-specific word distributions

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

- Unplate to see dependencies induced

All $z_i^d = k$: (in all of corpus)





⇓ marg. β_k

introducing coupling in all of above

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