Case Study 4: Collaborative Filtering

GraphLab Review,

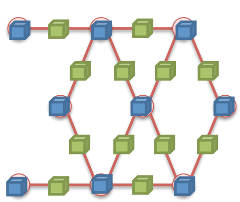
ALS, SGD, Gibbs Sampling for MF in GraphLab

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

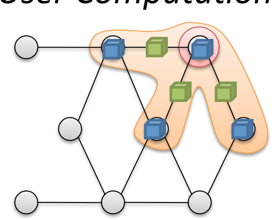
February 27th, 2014

The GraphLab Framework

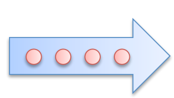
Graph Based
Data Representation



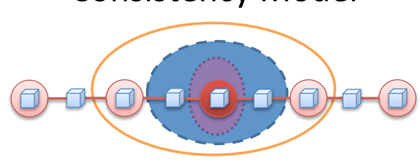
Update Functions
User Computation

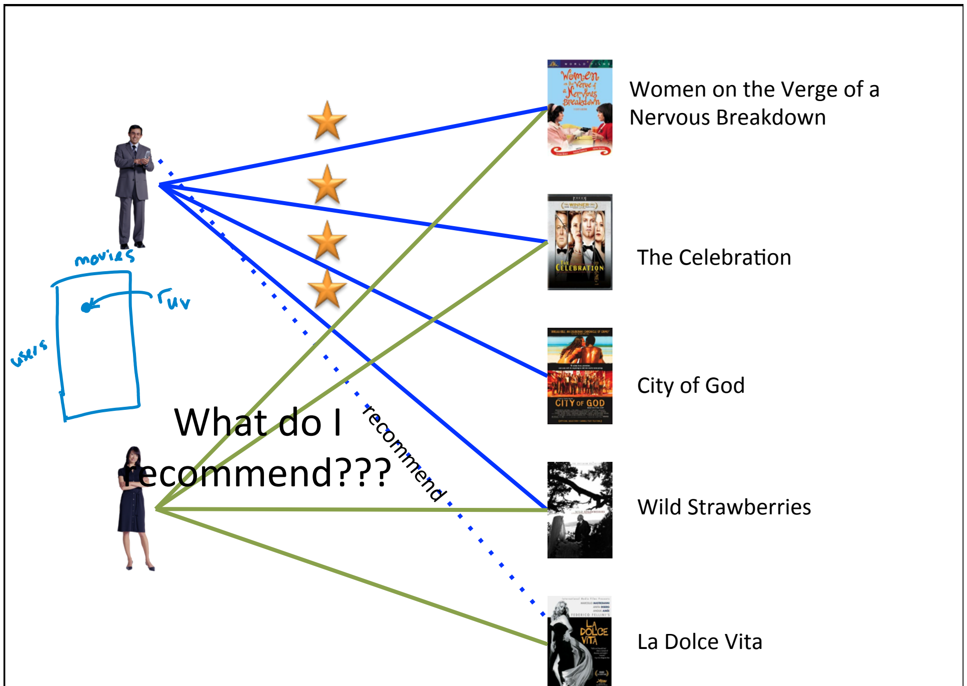


Scheduler

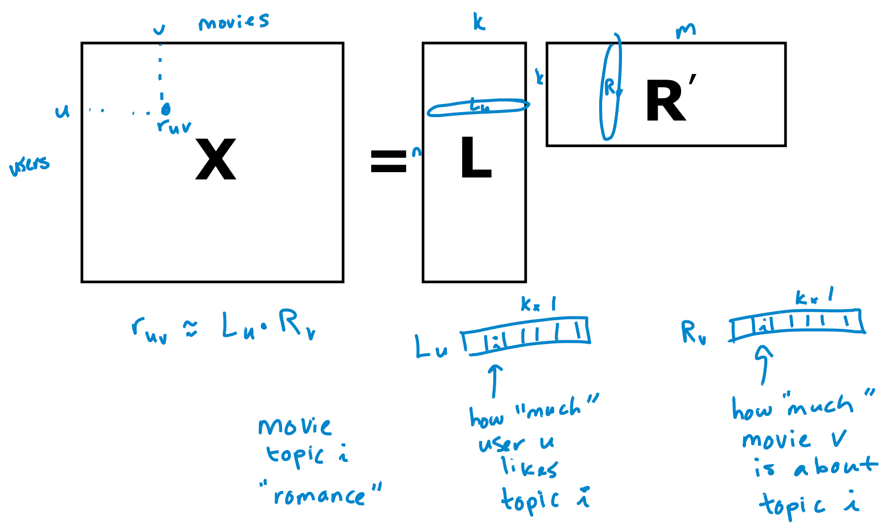


Consistency Model

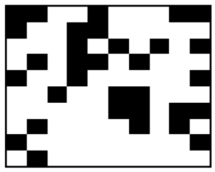




Interpreting Low-Rank Matrix Completion (aka Matrix Factorization)



Matrix Completion as a Graph

X = 

X_{ij} known for black cells
X_{ij} unknown for white cells
Rows index users
Columns index movies

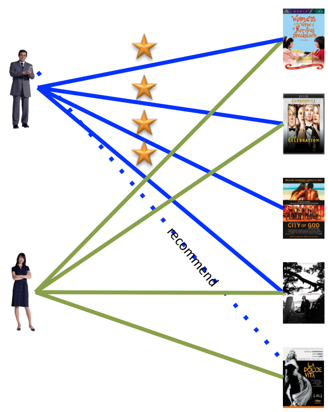
Coordinate Descent for Matrix Factorization: Alternating Least-Squares

$$\min_{L,R} \sum_{(u,v):r_{uv}\neq ?} (L_u \cdot R_v - r_{uv})^2 + \lambda_u \|L\| + \lambda_v \|R\|$$

- Fix movie factors, optimize for user factors
  - Independent least-squares over users $\min_{L_u} \sum_{v \in V_u} (L_u \cdot R_v - r_{uv})^2 + \lambda_u \|L\|$
- Fix user factors, optimize for movie factors
  - Independent least-squares over movies $\min_{R_v} \sum_{u \in U_v} (L_u \cdot R_v - r_{uv})^2 + \lambda_v \|R\|$
- System may be underdetermined: use regularization
- Converges to local optima

Alternating Least Squares Update Function

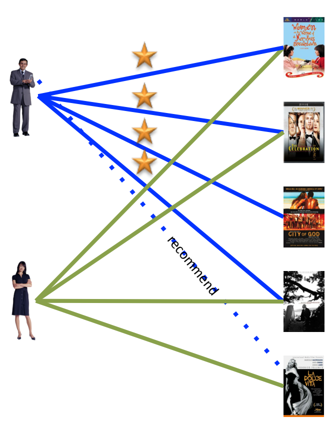
$$\min_{L_u} \sum_{v \in V_u} (L_u \cdot R_v - r_{uv})^2 \qquad \min_{R_v} \sum_{u \in U_v} (L_u \cdot R_v - r_{uv})^2$$



SGD for Matrix Factorization in GraphLab

$$\epsilon_t = L_u^{(t)} \cdot R_v^{(t)} - r_{uv}$$

$$\begin{bmatrix} L_u^{(t+1)} \\ R_v^{(t+1)} \end{bmatrix} \leftarrow \begin{bmatrix} (1-\eta_t\lambda_u)L_u^{(t)} - \eta_t\epsilon_t R_v^{(t)} \\ (1-\eta_t\lambda_v)R_v^{(t)} - \eta_t\epsilon_t L_u^{(t)} \end{bmatrix}$$



Bayesian PMF Example

Full Bayesian approach: place priors on ϕ as well!

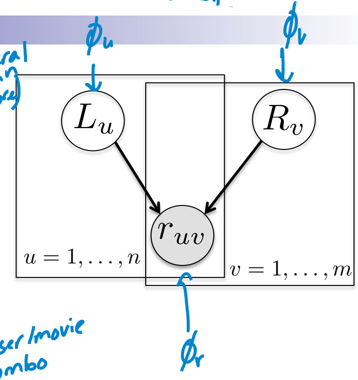
- Latent user and movie factors: (more general than before)

  $L_u \sim N(\mu_u, \Sigma_u) \quad u=1,\dots,n$

  $R_v \sim N(\mu_v, \Sigma_v) \quad v=1,\dots,m$

- Observations $r_{uv} \sim N(L_u' R_v, \sigma_r^2)$
- Hyperparameters:

  $\phi = \{\underbrace{\mu_u, \Sigma_u}_{\phi_u}, \underbrace{\mu_v, \Sigma_v}_{\phi_v}, \underbrace{\sigma_r^2}_{\phi_r}\}$



- Want to predict new movie rating: new user/movie combo; posterior given obs. so far

$$p(r_{uv}^* \mid X, \phi) = \int p(r_{uv}^* \mid L_u, R_v)\, p(L, R \mid X, \phi)\, dL\, dR$$

new rating; obs. ratings

Bayesian PMF Gibbs Sampler

- Outline of Bayesian PMF sampler

1. Init $L^{(1)}, R^{(1)}$
2. For $k=1,\dots,$ Niter
   - (i) Sample hyperparams $\phi^{(k)} = \{\phi_u^{(k)}, \phi_v^{(k)}, \phi_r^{(k)}\}$
   - (ii) For each user $u=1,\dots,n$ sample in parallel

     $L_u^{(k+1)} \sim p(L_u \mid X, R^{(k)}, \phi^{(k)})$
   - (iii) For each movie $v=1,\dots,m$ sample in parallel

     $R_v^{(k+1)} \sim p(R_v \mid X, L^{(k+1)}, \phi^{(k)})$

Very similar to ideas of ALS (systematically)

Bayesian PMF Example

- For user u:

prior / likelihood for user u

$$p(L_u \mid X, R, \phi_u) \propto p(L_u \mid \phi_u) \prod_{v \in V_u} p(r_{uv} \mid L_u, R_v, \phi_r)$$

$$\propto N(L_u \mid \mu_u, \Sigma_u) \prod_{v \in V_u} N(r_{uv} \mid L_u \cdot R_v, \sigma_r^2)$$

$$= N(L_u \mid \tilde{\mu}_u, \tilde{\Sigma}_u) \quad \leftarrow \text{via conjugacy}$$

posterior is in the same family as prior

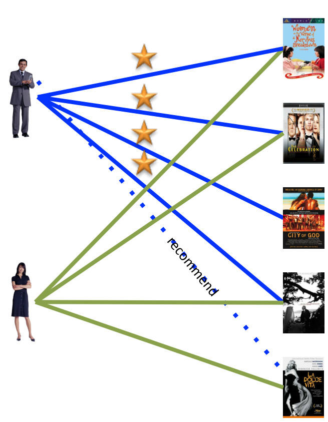
where $\tilde{\Sigma}_u^{-1} = \Sigma_u^{-1} + \sigma_r^{-2} \sum_{v \in V_u} R_v R_v^T$

$\tilde{\mu}_u = \tilde{\Sigma}_u \left(\sigma_r^{-2} \sum_{v \in V_u} r_{uv} R_v + \Sigma_u \mu_u \right)$

- Symmetrically for R_v conditioned on L (breaks down over movies)
- Luckily, we can use this to get our desired posterior samples

PMF Gibbs Sampling in GraphLab

$$p(L_u \mid X, R, \phi_u) = N(\tilde{\mu}_u, \tilde{\Sigma}_u) \qquad \tilde{\Sigma}_u = \Sigma_u^{-1} + \sigma_r^{-2} \sum_{v \in V_u} R_v R_v^T \qquad \tilde{\mu}_u = \tilde{\Sigma}_u \left(\sigma_r^{-2} \sum_{v \in V_u} r_{uv} R_v + \Sigma_u \mu_u \right)$$



GraphLab

Release 2.2 available now

http://graphlab.org

Documentation… Code… Tutorials… (more on the way)

GraphChi 0.1 available now

http://graphchi.org

What you need to know…

- Data-parallel versus graph-parallel computation
- Bulk synchronous processing versus asynchronous processing
- GraphLab system for graph-parallel computation
  - Data representation
  - Update functions
  - Scheduling
  - Consistency model
- ALS, SGD and Gibbs for matrix factorization/PMF in GraphLab

Reading

- Papers under “Case Study IV: **Parallel Learning with GraphLab**”

- Optional:
  - Parallel Splash BP
    http://www.ml.cmu.edu/research/dap-papers/dap-gonzalez.pdf

Acknowledgements

- Slides based on Carlos Guestrin’s GraphLab talk

Case Study 5: Mixed Membership Modeling

Clustering Documents Revisited, Latent Dirichlet Allocation

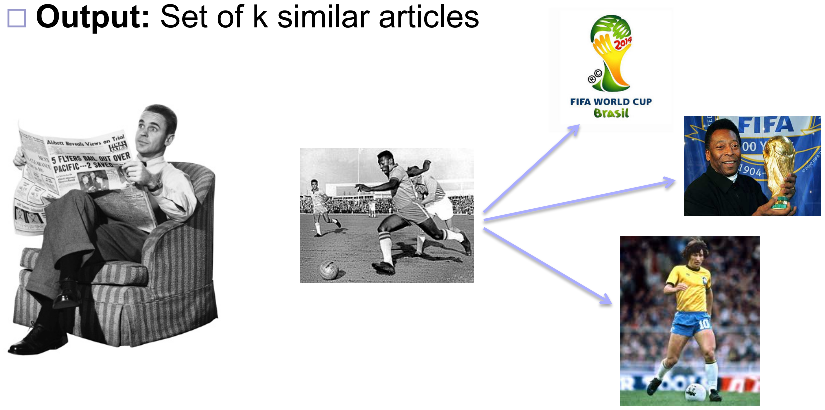
Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
February 27th, 2014

Document Retrieval

- **Goal:** Retrieve documents of interest
- **Challenges:**
  - Tons of articles out there
  - How should we measure similarity?

Task 1: Find Similar Documents

- **First considered:**
  - **Input:** Query article
  - **Output:** Set of k similar articles



Task 2: Cluster Documents

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

Document Representation

- Bag of words model

document *d*

previously: $x^d = [\ \ ]$ ← vector fcn of word counts (e.g. tf-idf)

performed operations on this vector

now: $x^d = \{w_1^d, \dots, w_{N_d}^d\}$

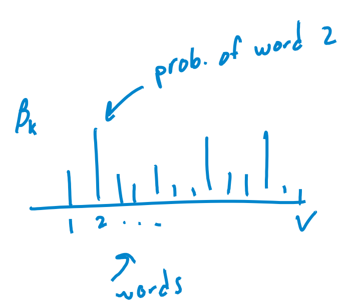
unordered set of N_d word indices with $w_i^d \in V$ vocab.

A Generative Model

- Documents: $x^1, \dots, x^D$ with $x^d = \{w_1^d, \dots, w_{N_d}^d\}$
- Associated topics: $z^1, \dots, z^D$ with $z^d \in \{1, \dots, K\}$ ← # topics
- Parameters: $\theta = \{\pi, \beta\}$

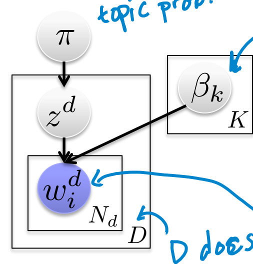
as before $\begin{cases} \pi = [\pi_1, \dots, \pi_K] & \text{topic probabilities} \\ Pr(z^d = k) = \pi_k \end{cases}$

$$\beta = \begin{array}{c|c} & 1 \quad 2 \quad \dots \quad V \\ \hline \begin{array}{c} 1 \\ 2 \\ \vdots \\ K \end{array} & \begin{bmatrix} \text{---}\ \beta_1\ \text{---} \\ \vdots \\ \text{---}\ \beta_K\ \text{---} \end{bmatrix} \end{array}$$



A Generative Model

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

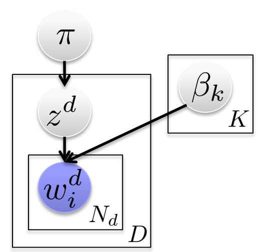


$z^d \sim \pi$ generate topic

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

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

Model In Pictures

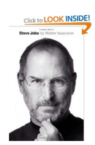


- Mixture weights (on topics)

π

- Topic distributions (on words)

β_1



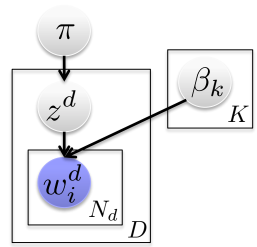
β_K

- For each document,

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

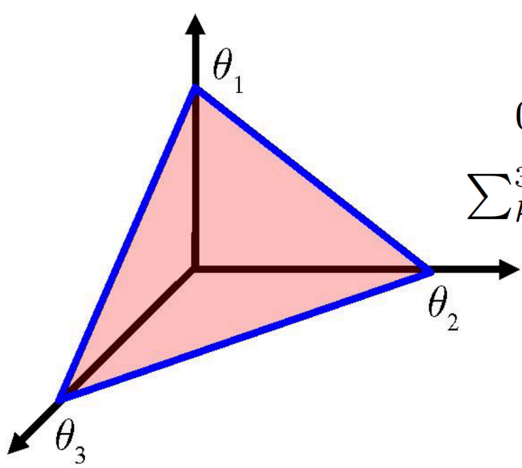
Bayesian Document Model

- Model parameters $\pi, \{\beta_k\}$ unknown
- Bayesian approach
- Need distribution on pmf's



The Simplex in 3D

- The simplex defines the hyperplane of vectors that sum to 1

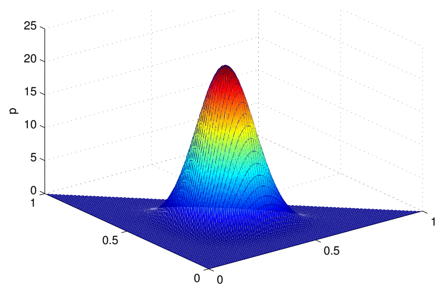


$$0 \leq \theta_k \leq 1$$

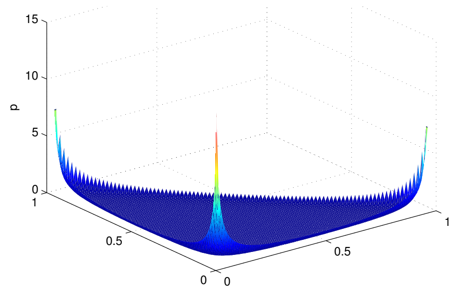
$$\sum_{k=1}^{3} \theta_k = 1$$

Dirichlet Distributions

- The Dirichlet distribution is defined on the simplex



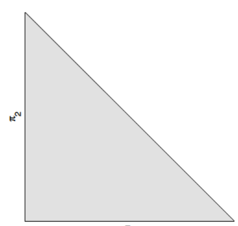
$$p(\pi \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$



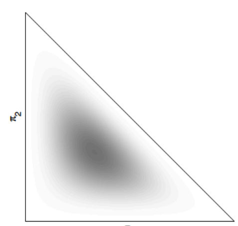
Moments: $\mathbb{E}_\alpha[\pi_k] = \frac{\alpha_k}{\alpha_0}$

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

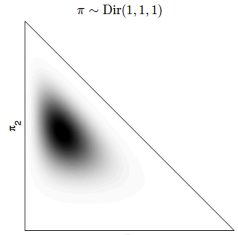
Dirichlet Probability Densities



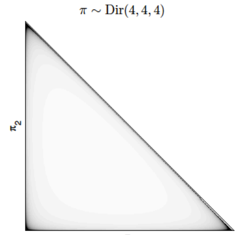
$\pi \sim \mathrm{Dir}(1,1,1)$



$\pi \sim \mathrm{Dir}(4,4,4)$

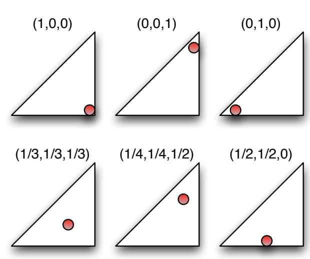


$\pi \sim \mathrm{Dir}(4,9,7)$



$\pi \sim \mathrm{Dir}(0.2,0.2,0.2)$

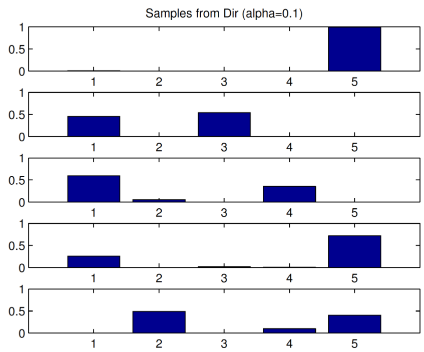
$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$



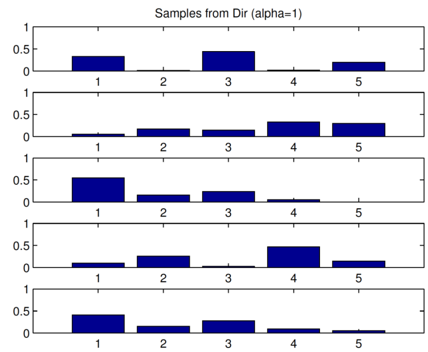
Dirichlet Samples

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

- Samples are ***sparse*** for small values of α_i



$\mathrm{Dir}(\pi \mid 0.1, 0.1, 0.1, 0.1, 0.1)$

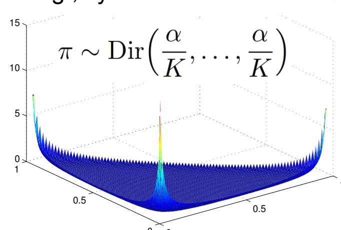


$\mathrm{Dir}(\pi \mid 1.0, 1.0, 1.0, 1.0, 1.0)$

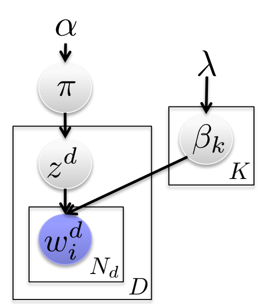
Model Summary

- 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 β_k

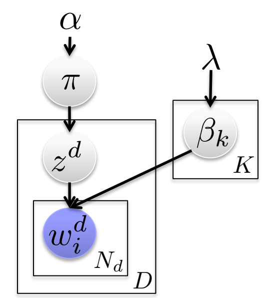


- Sample observations as

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

Posterior Inference via Sampling

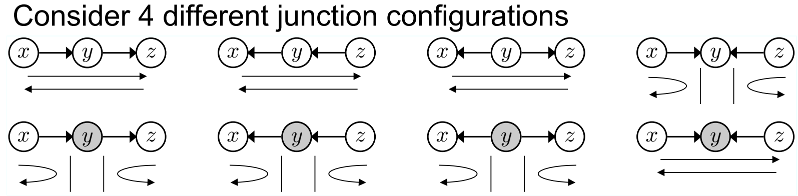
- Iterate between sampling



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

Conditional Independence in Bayes Nets

- Consider 4 different junction configurations

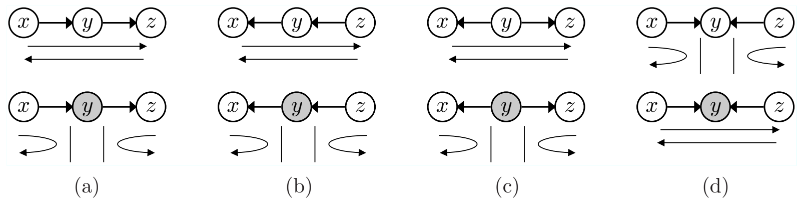


(a) (b) (c) (d)

- Conditional versus unconditional independence:

Bayes Ball Algorithm

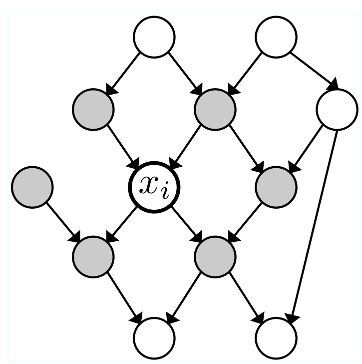
- Consider 4 different junction configurations



- Bayes ball algorithm

Markov Blanket

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

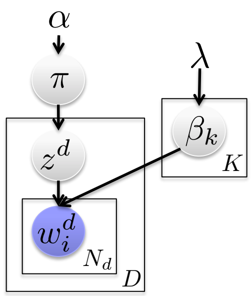


- Gibbs sampling iterates between full conditionals

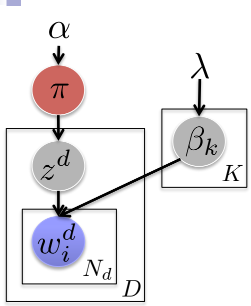
→ simplify to

Unplated Document Model

- Recall that the plate notation is really indicating



Complete Conditional for π

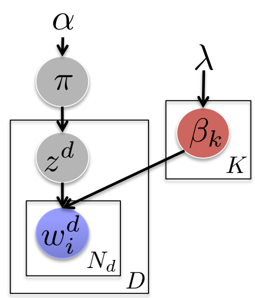


- Recall conjugate Dirichlet prior

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

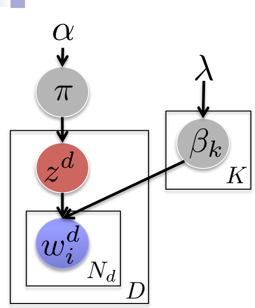
- Likelihood:
- Dirichlet posterior
  - Count occurrences of
  - Then,
  - Conjugacy: **Posterior** has same form as **prior**

Complete Conditional for β_k



- Again, Dirichlet prior
- Consider docs *d* such that
  - For these observations,
  - Do any other docs depend on β_k?
- Then,
  - Again, **posterior** has same form as **prior**

Complete Conditional for z^d



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

$$w_i^d \mid z^d, \{\beta_k\} \sim \beta_{z^d}$$

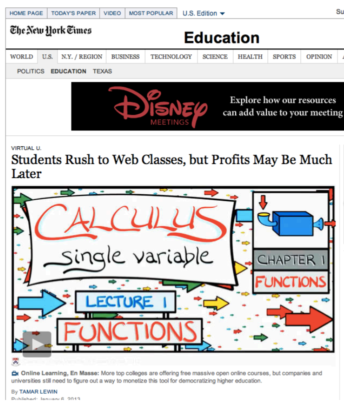
- Calculate the posterior for each value of z^d ("responsibility" of each topic to the doc):

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

- Sample each cluster indicator as

Task 3: Mixed Membership Models

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



EDUCATION

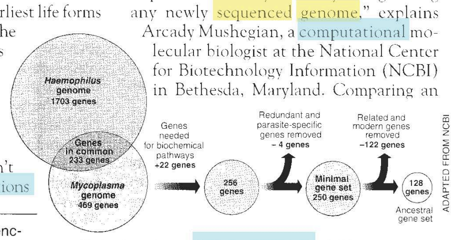
FINANCE

TECHNOLOGY

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.

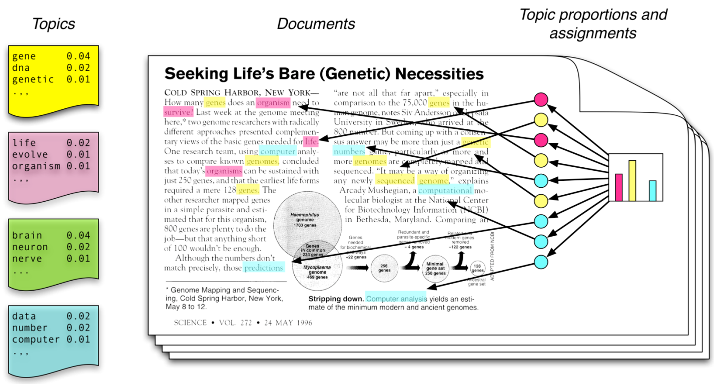
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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



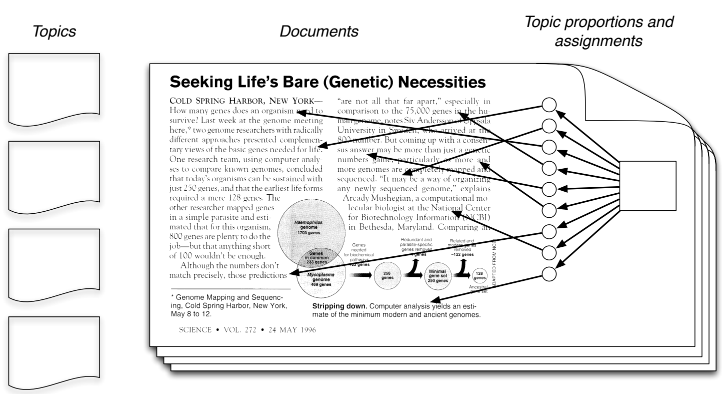
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

Latent Dirichlet Allocation (LDA)



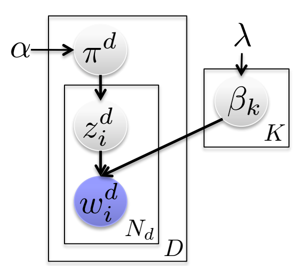
Latent Dirichlet Allocation (LDA)



LDA Generative Model

- Observations: $w_1^d, \ldots, w_{N_d}^d$
- Associated topics: $z_1^d, \ldots, z_{N_d}^d$
- Parameters: $\theta = \{\{\pi^d\}, \{\beta_k\}\}$
- Generative model:

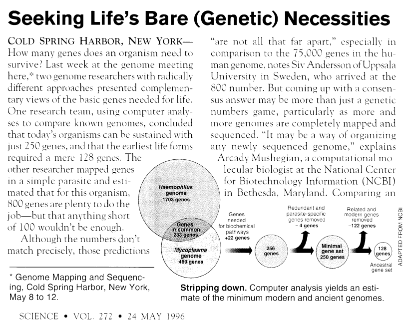
LDA Joint Probability



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

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

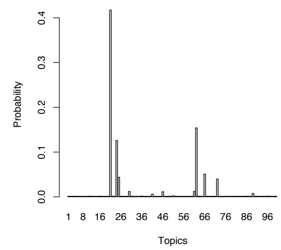
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

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

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

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 of the LDA material