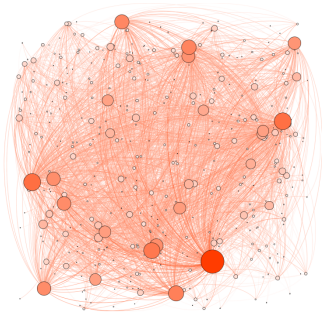
Case Study 4: Collaborative Filtering

Matrix Factorization and Probabilistic LFMs for Network Modeling

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
May 21st, 2015

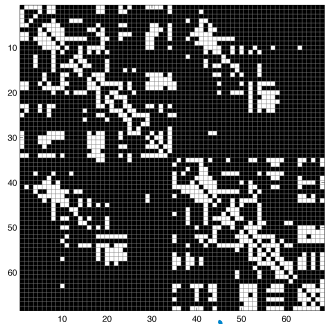
Network Data

- Structure of network data



nodes in network w/ undirected edges

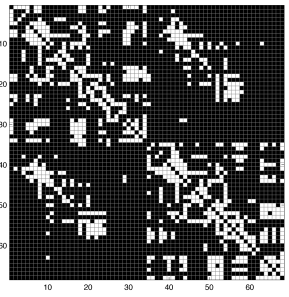
white square = edge between two nodes



Adjacency matrix

Properties of Data Source

- Similarities to Netflix data:
  - Matrix ~valued data (adj. matrix)
  - High-dimensional many nodes
  - Sparse few links between nodes (eg. ppl in a social network)
- Differences
  - Square ← same indices for rows + columns
  - Binary ← yes/no for link (other ext. possible... multiple)



If undirected, then matrix is symmetric

Matrix Factorization for Network Data

- Vanilla matrix factorization approach:

In undirected case, just introduce node (e.g. user) factors L_u

$r_{uv} \approx L_u \cdot L_v$ ← edge between users u + v

In directed case, just introduce sender factors L_u and receiver factors $\tilde{L}_u$

$r_{uv} \approx L_u \cdot \tilde{L}_v$ ← edge from user u to user v

- What to return for link prediction?

Is r_{uv} binary? $L_u \in \mathbb{R}^k$ → no

Many options, but can return top $\tilde{k}$ $r_{uv_1}, r_{uv_2}, \ldots, r_{uv_{\tilde{k}}}$ (just threshold rule)

- Slightly fancier:

More appropriate to have $r_{uv} \in [0, 1]$

use $r_{uv} \approx \sigma(L_u \cdot L_v)$ σ = logistic fcn

Probabilistic Latent Space Models

- Assume features (covariates) of the user X_u or relationship X_{uv} ← could just be $X_u X_v$
- Each user has a "position" in a *k*-dimensional latent space

unobserved (learn) $L_u \in \mathbb{R}^k$, just as in matrix fact.

- Probability of link:

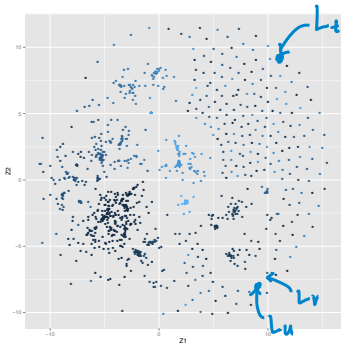
$$\log \text{odds}\ p(r_{uv}=1 \mid L_u, L_v, X_{uv}, \beta)$$

$$= \log \frac{p(r_{uv}=1 \mid L_u, L_v, X_{uv}, \beta)}{p(r_{uv}=0 \mid L_u, L_v, X_{uv}, \beta)}$$

$$= \beta_0 + \beta^T X_{uv} - |L_u - L_v|$$

OR

$$= \beta_0 + \beta^T X_{uv} + |L_u L_v|$$



2D example ($k=2$)

Probabilistic Latent Space Models

- Probability of link:

$$\log \text{odds}\ p(r_{uv}=1 \mid L_u, L_v, x_{uv}, \beta) = \beta_0 + \beta^T x_{uv} - |L_u - L_v|$$

prob. link is high for L_u close to L_v

$$\log \text{odds}\ p(r_{uv}=1 \mid L_u, L_v, x_{uv}, \beta) = \beta_0 + \beta^T x_{uv} + |L_u^T L_v|$$

can modify as $\frac{\|L_u^T L_v\|}{\|L_v\|}$

prob. of link is high if L_u & L_v small

- Bayesian approach:
  - Place prior on user factors and regression coefficients
  - Place hyperprior on user factor hyperparameters
- Many other options and extensions (e.g., can use GMM for L_u → clustering of users in the latent space)

What you need to know…

- Representation of network data as a matrix
  - Adjacency matrix
- Similarities and differences between adjacency matrices and general matrix-valued data
- Matrix factorization approaches for network data
  - Just use standard MF and threshold output
  - Introduce link functions to constrain predicted values
- Probabilistic latent space models
  - Model link probabilities using distance between latent factors

Case Study 5: Mixed Membership Modeling

Clustering Documents Revisited, Latent Dirichlet Allocation

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

Emily Fox

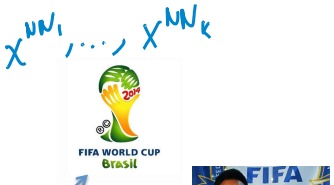
May 21st, 2015

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 x
  - **Output:** Set of k similar articles



x

Task 2: Cluster Documents

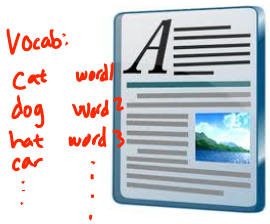
- **Then examined:**
  - Cluster documents based on topic

sports

world news

Document Representation

- Bag of words model



Vocab:
cat word 1
dog word 2
hat word 3
car
:

document *d*

previously ← vector fcn of word counts

$X = [\ ]$ (e.g. tf-idf)

performed operations on this vector

$w_i = 3$ ⇒ i^{th} word in this doc is "hat"

now: representation of a doc

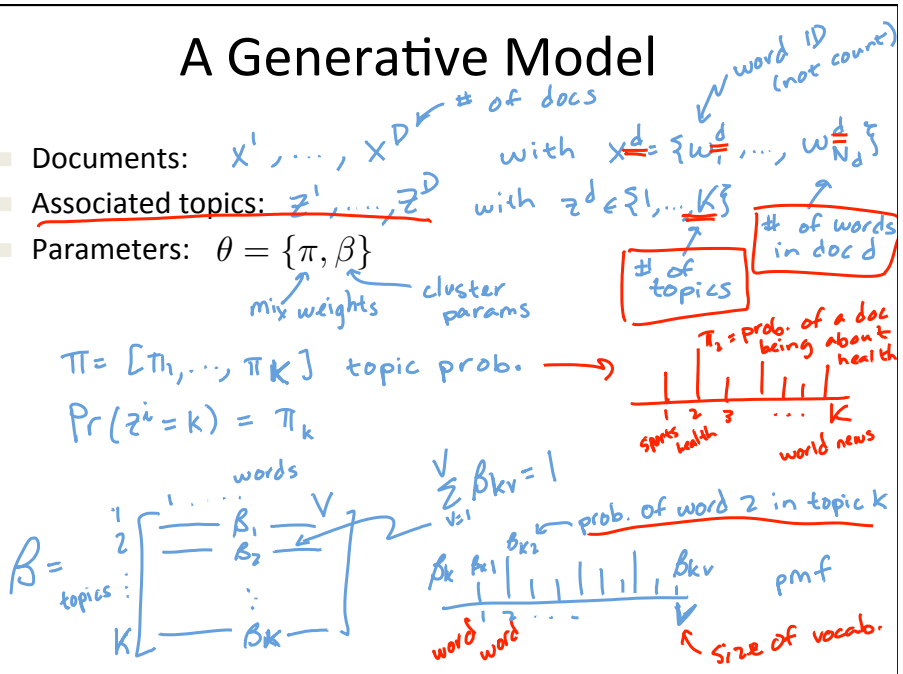
$X = \{w_1, \ldots, w_N\}$ # of words in doc

unordered ~~set~~ list of N words in doc.

$w_i \in V$ (vocab)

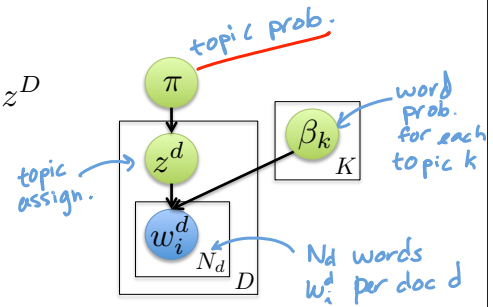
A Generative Model

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



A Generative Model

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



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

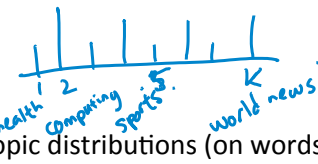
Sample words:
$w_i^d \mid z^d \sim \beta_{z^d}$ $i = 1, \ldots, N_d$
topic of doc d

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

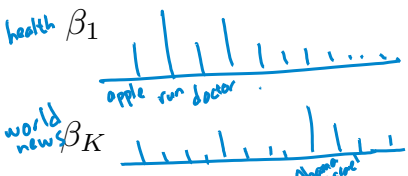
Model In Pictures

- Mixture weights (on topics)

π

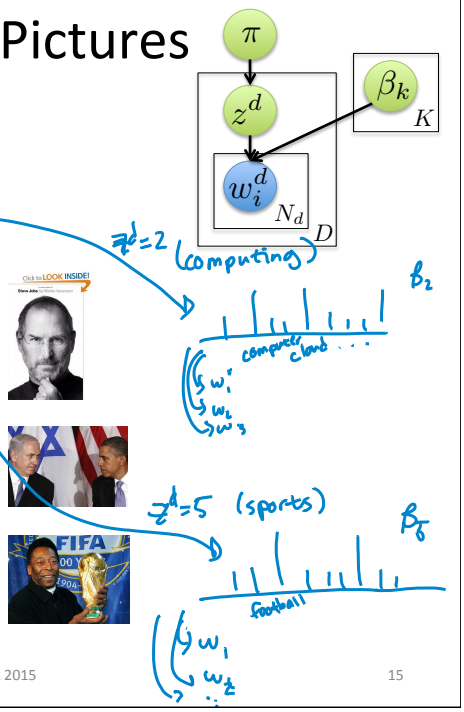


- Topic distributions (on words)



- For each document,

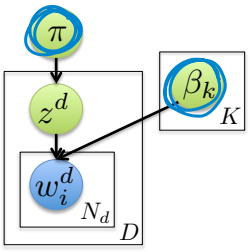
$$z^d \sim \pi$$
$$w_i^d \mid z^d \sim \beta_{z^d}$$



Bayesian Document Model

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

place priors on parameters



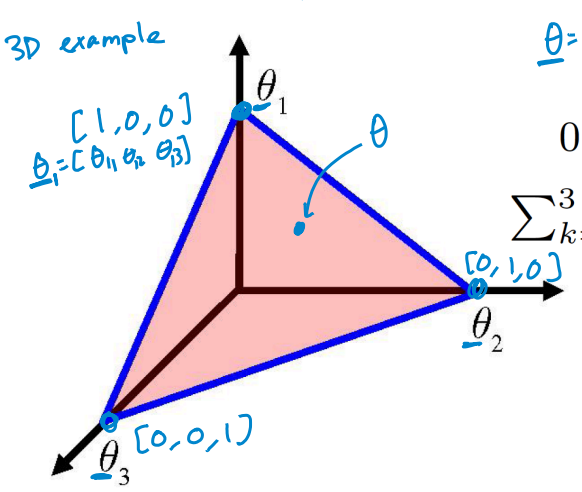
- Need distribution on pmf's

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

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

The Simplex in 3D

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



$\theta = [\theta_1, \theta_2, \theta_K]$

$$0 \le \theta_k \le 1$$

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

Dirichlet Distributions

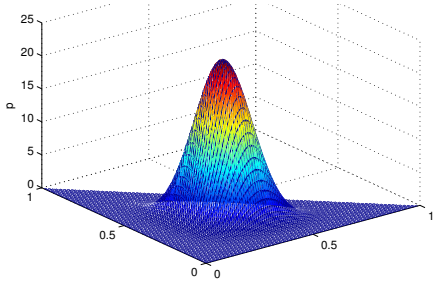
- The Dirichlet distribution is defined on the simplex

$\pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$ — K-dim pmf

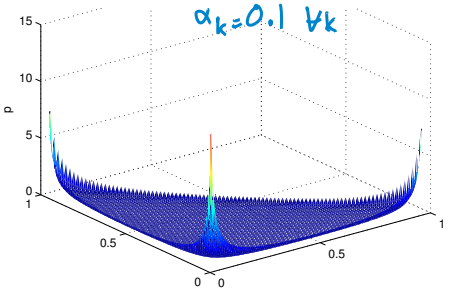
$\Rightarrow \sum_{k=1}^{K} \pi_k = 1$

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

$\alpha_k = 10 \ \forall k$



$\alpha_k = 0.1 \ \forall k$

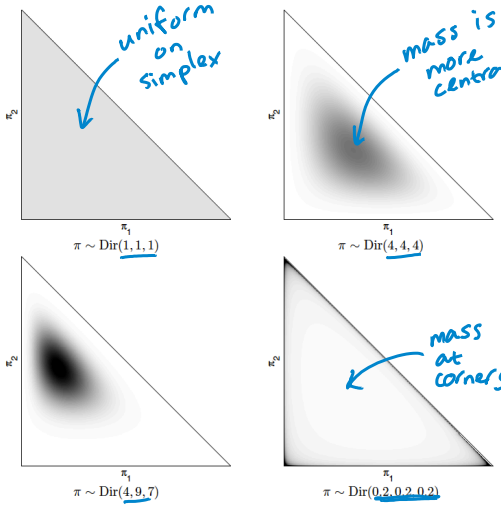


Moments:

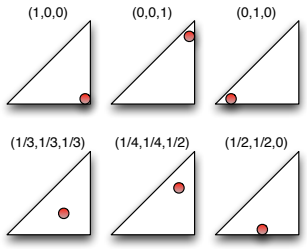
$$\mathbb{E}_\alpha[\pi_k] = \frac{\alpha_k}{\alpha_0}$$

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

Dirichlet Probability Densities



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



draws from Dir

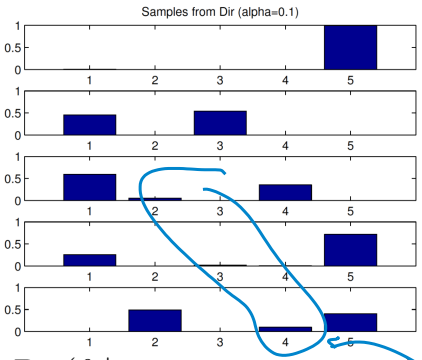
Dirichlet Samples

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

$\alpha_0 = \sum_k \alpha_k$

5D example

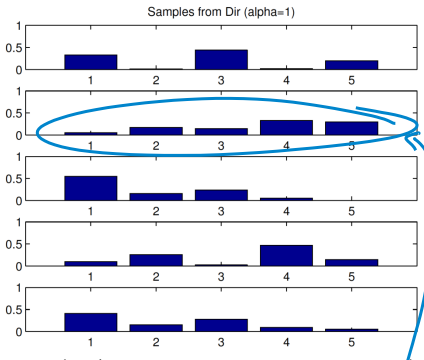
- Samples are ***sparse*** for small values of α_i



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

puts mass at corners

many small weights



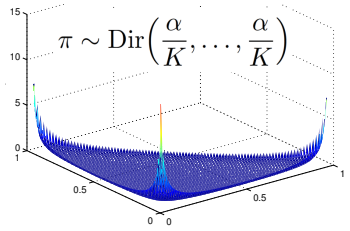
$$\text{Dir}(\pi \mid 1.0, 1.0, 1.0, 1.0, 1.0)$$

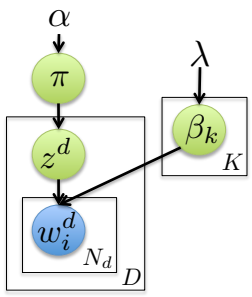
uniform

more balanced

Model Summary

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





  - 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.

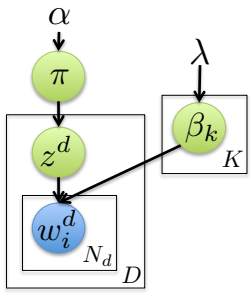
$\pi \sim p(\pi \mid \{z^d\}, \{\beta_k\}, \{w_i^d\})$

For $k = 1, \ldots, K$

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

For $d = 1, \ldots, D$

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

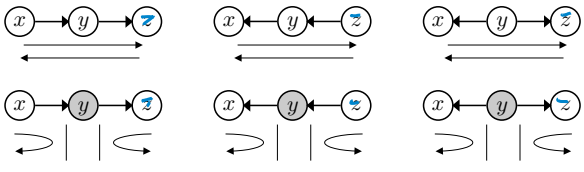


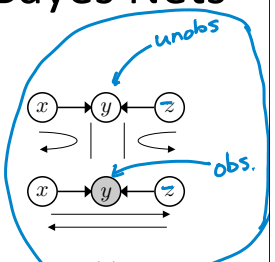
- What form do these complete conditionals take?
  - First a look at statements of conditional independence in directed graphical models

Conditional Independence in Bayes Nets

$p(\text{joint}) = \prod_i p(x_i \mid \text{parents}(x_i))$

- Consider 4 different junction configurations





- Conditional versus unconditional independence:

$x \perp z$

$p(x,y,z) = p(x)p(z)p(y \mid x,z) \xrightarrow{\text{sum } y} p(x,z) = p(x)p(z)$

$p(x,z \mid y) \propto p(x,y,z) = p(x)p(z)p(y \mid x,z)$

$\not\propto p(x \mid y)p(z \mid y)$

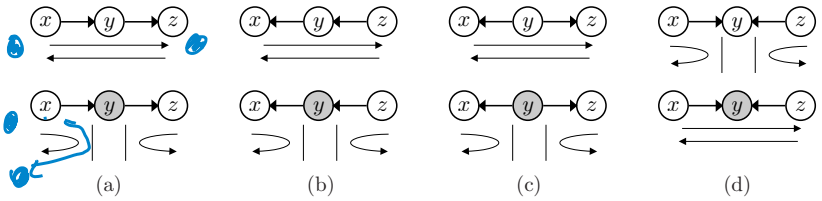
$x \not\perp z \mid y$

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

but if alarm = 1, then ind. a priori increase in prob(z) ⇒ decrease in prob(x)

Bayes Ball Algorithm

- Consider 4 different junction configurations



- Bayes ball algorithm

Start ball at one end or the other.

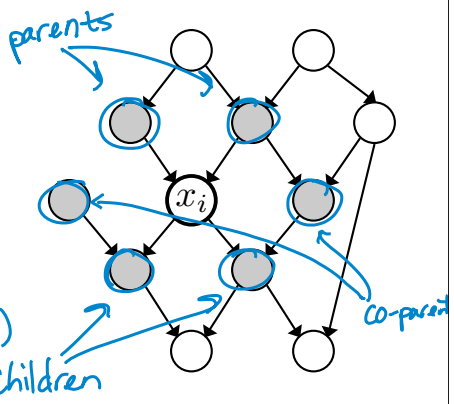
If ball passes to a node (straight arrows) then, not cond./marg. independent.

If ball bounces back (wall + curved arrows), then 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

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

→ simplify to

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

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

Reading for *Next Lecture*

- **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