Latent Dirichlet Allocation

464.561A Models and Technologies for Information Services

Jonghun Park

jonghun@snu.ac.kr

Dept. of Industrial Eng.

Seoul National University

latent Dirichlet allocation (LDA)

- generative probabilistic model for collections of discrete data
  - 3-level hierarchical Bayesian model
- documents are represented as random mixtures over latent topics
  - each topic is characterized by a distribution over words
  - documents can be associated with multiple topics

definitions

- word
  - an item from a vocabulary
  - indexed by $\{1, \ldots, V\}$
  - v-th word: represented as V-vector w s.t. $w^v = 1$ and $w^u = 0$ for $u \neq v$
- document
  - sequence of N words
  - denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$
  - w_n: n-th word in the sequence
- corpus
  - collection of M documents
  - denoted by $\mathcal{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$

generative process for a document

1. choose $N \sim \text{Poisson}(\xi)$
2. choose $\theta \sim \text{Dir}(\alpha)$
3. for each of the N words w_n:
   ① choose a topic $z_n \sim \text{Multinomial}(\theta)$
   ② choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

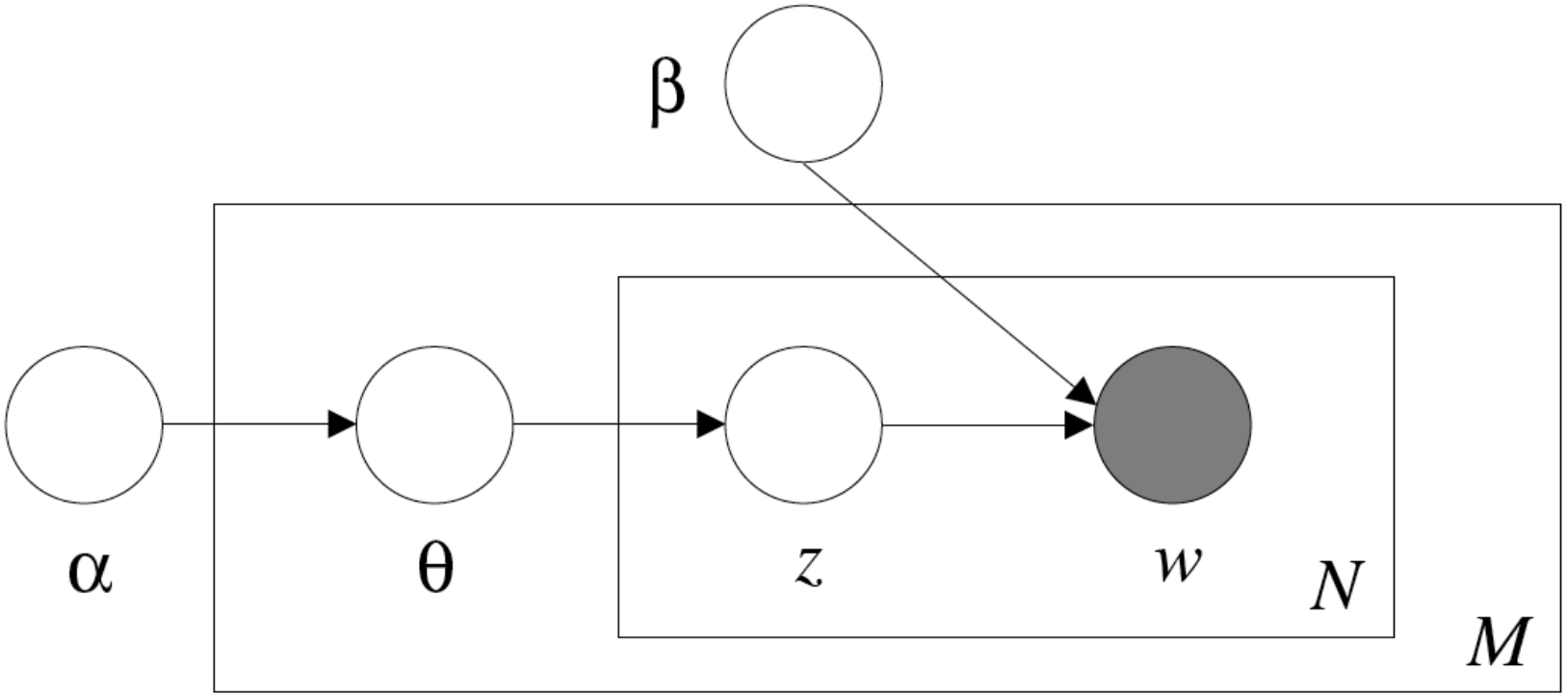
assumptions

- dimension of topic variable z: k
  - = dimension of Dirichlet distribution
- parameter matrix for word probabilities: β
  - size: $k \times V$
    $\beta_{ij} = p(w^j = 1 | z^i = 1)$
- k dimensional Dirichlet rv θ
  - can take values in the (k-1)-simplex

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

  - α is a k-vector with components $\alpha_i > 0$
  - $\Gamma(x)$ is the Gamma function
  - note: k-vector θ lies in the (k-1)-simplex if $\theta \geq 0, \sum_{i=1}^{k} \theta_i = 1$

graphical model representation



joint distribution

- θ: topic mixture
- $\mathbf{z}$: set of N topics
- $\mathbf{w}$: set of N words

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

where

$$p(z_n|\theta) = \theta_i \quad \text{for the unique } i \text{ s.t. } z_n^i = 1$$

document & corpus distributions

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta$$

$$p(\mathcal{D}|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d$$

$$p(w|\theta,\beta) = \sum_{z} p(z|\theta)p(w|z,\beta)$$: word distribution

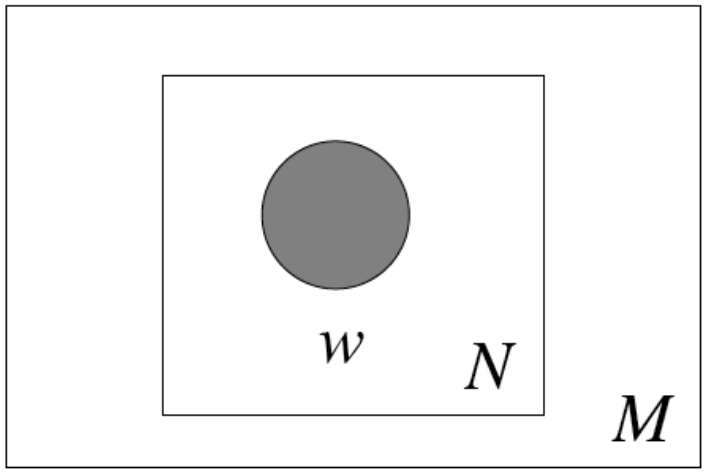
relationship with other latent variable models

- unigram model
- mixture of unigrams
- pLSI (probabilistic latent semantic indexing)

unigram model

- assumes a single multinomial distribution

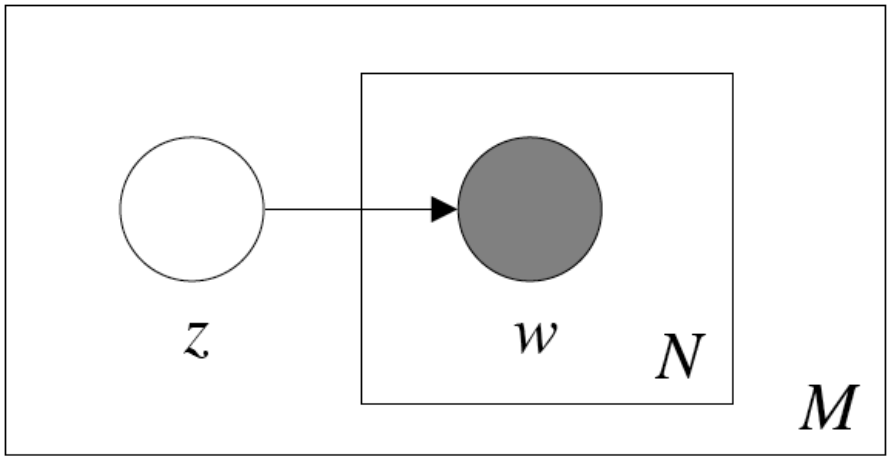
$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$



mixture of unigrams

- each document is generated by first choosing a topic z and then generating N words independently from the conditional multinomial $p(w|z)$

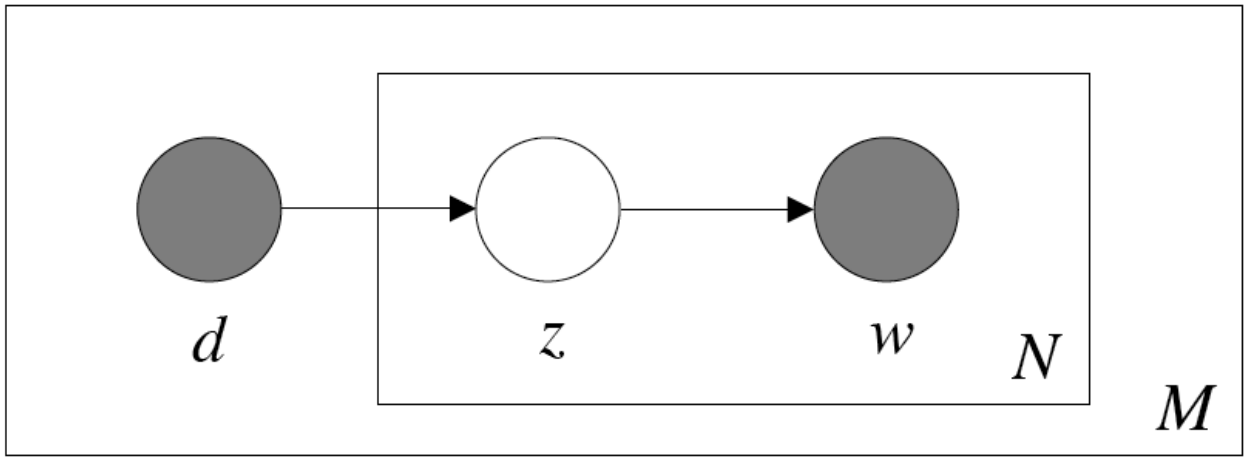
$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n|z)$$



pLSI

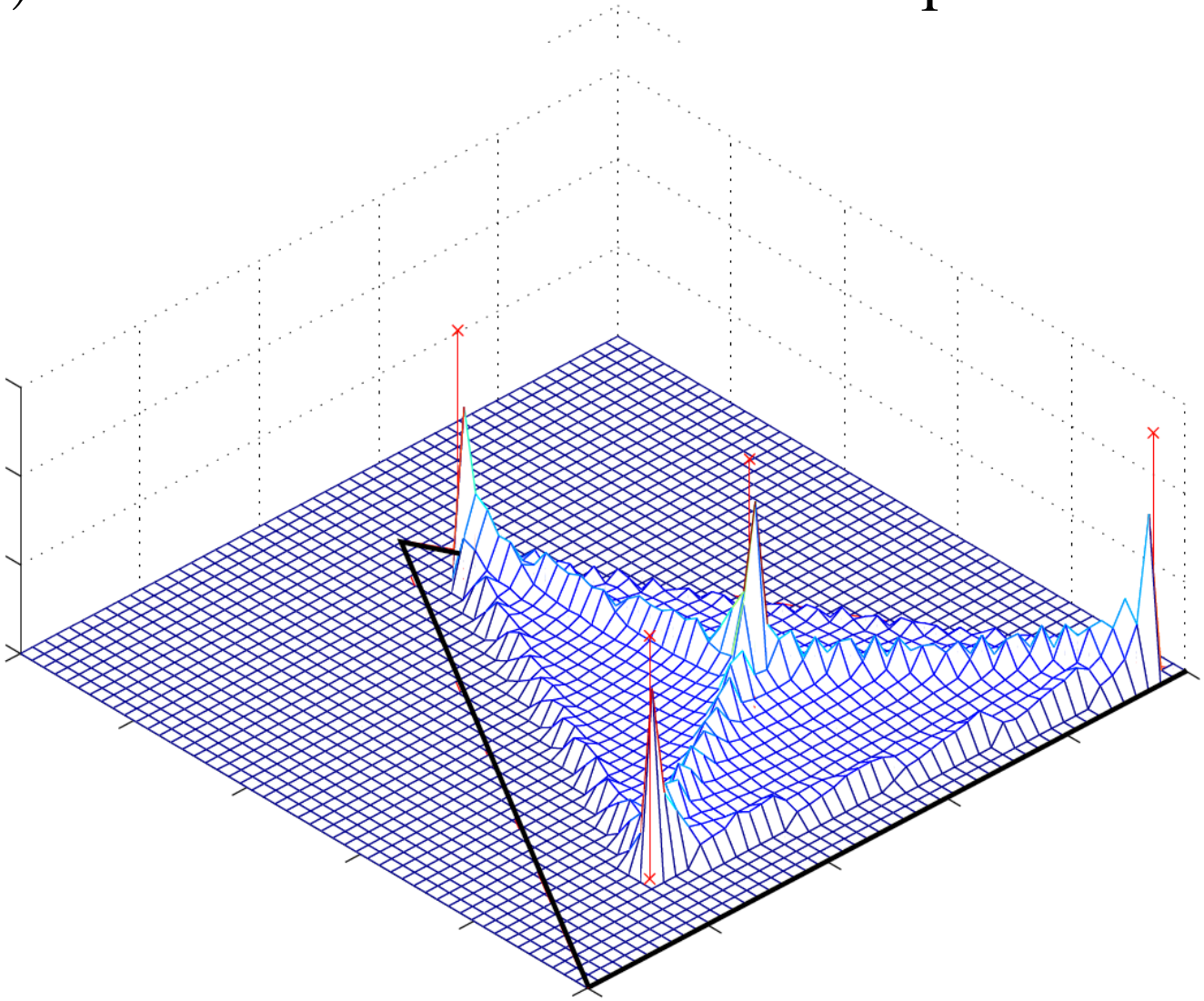
- document d and word w_n are conditionally independent given an unobserved topic z

$$p(d, w_n) = p(d) \sum_z p(w_n|z)p(z|d)$$



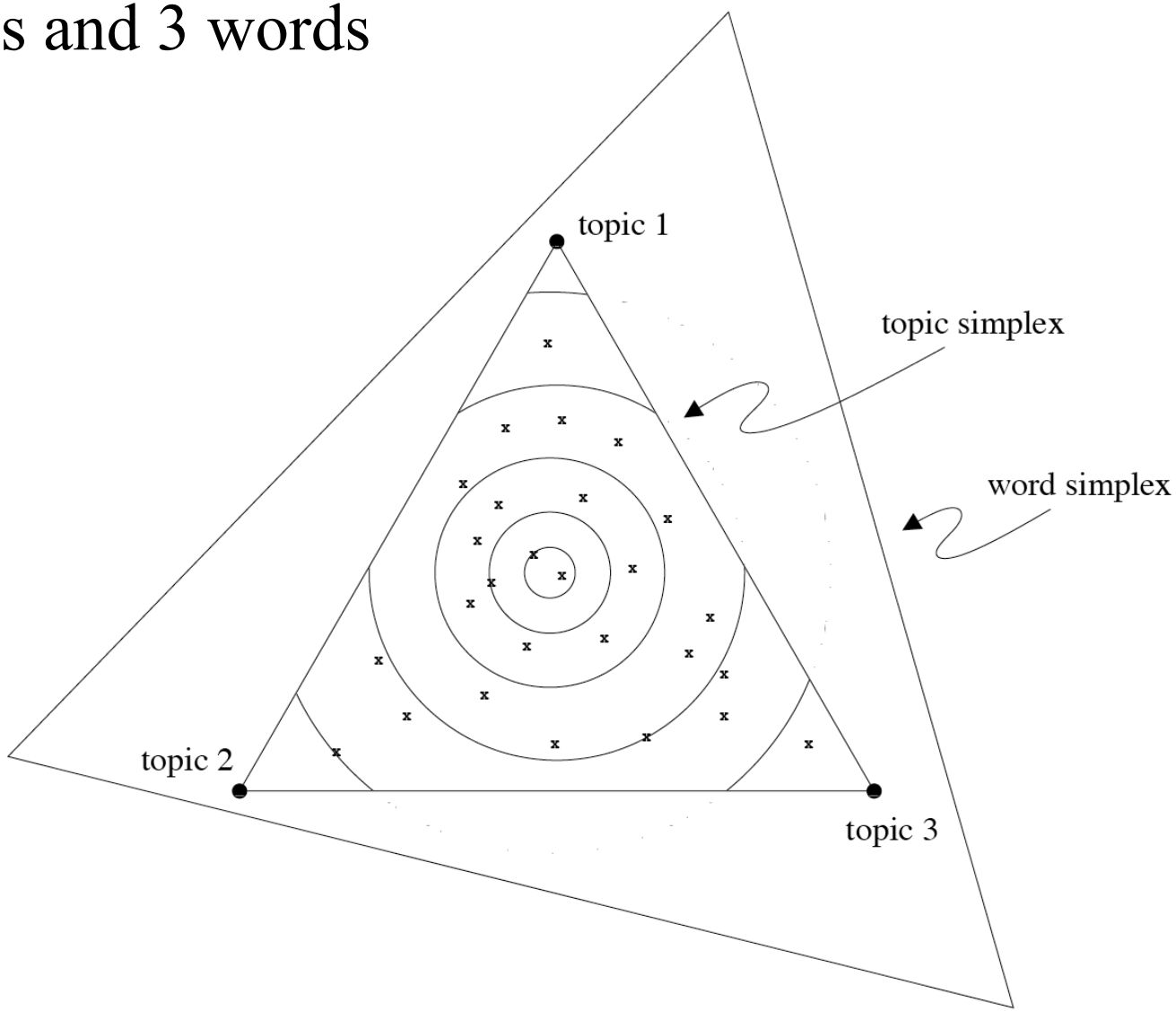
geometric interpretation

$p(w|\theta, \beta)$ under LDA for 3 words and 4 topics



geometric interpretation

3 topics and 3 words



inference

- posterior distribution of the hidden variables given a document

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

- intractable to compute in general =>
  - variational approximation
  - Laplace approximation
  - Markov chain Monte Carlo

parameter estimation

- given $\mathcal{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$, find α and β that maximize the log likelihood of the data:

$$l(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d | \alpha, \beta)$$

- intractable to compute =>
  - variational EM procedure

example

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

example

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

readings

- D. Ramage, S. Dumais, D. Liebling, “Characterizing Microblogs with Topic Models,” ICWSM 2010.

references

- D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, vol. 3, 2003, pp. 993-1022.
- M. Steyvers and T. Griffiths, “Probabilistic Topic Models,” in: In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning, Laurence Erlbaum