Latent Dirichlet Allocation (LDA)

Mahdi Pakdaman

Intelligent systems Program
University of Pittsburgh

Outline

- Brief Review
- LDA
  - Dirichlet Distribution
  - The Model
  - Theoretical insights
  - Applications
  - Parameter Estimation
- Extensions to LDA
- Summary

How the story began !

- We Model the text corpora to :
  - similarity/relevance judgments, Classification, Summarization, ……
- Represent each document as a vector space
  - A *word* is an item from a vocabulary indexed by {1,…,V}. We represent words using unit-basis vectors. The *v*'th word is represented by a V-vector *w* such that $w^v = 1$ and $w^u = 0$ for $u \neq v$

    $\mathbf{w} = (w_1, w_2, \ldots, w_n)$
  - A *document* is a sequence of N words denoted by

    where w_n is the *n*th word in the sequence.
  - A *corpus* is a collection of M documents denoted by

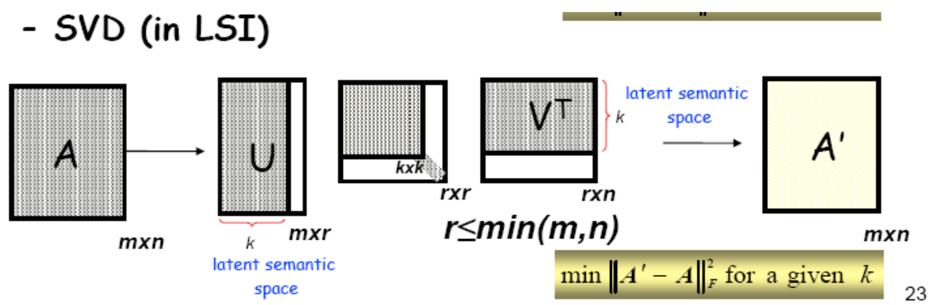
    $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$

The Problem with Vector space representation

- Three problems that arise using the vector space model:
  - The Vectors are very sparse
  - synonymy: many ways to refer to the same object, e.g. car and automobile
    - Will have small cosine but are related
    - leads to poor recall
  - polysemy: most words have more than one distinct meaning, e.g. model, python, chip
    - Will have large cosine but not truly related
    - leads to poor precision

Latent Semantic Space

- LSI maps terms and documents to a "latent semantic space"
- Comparing terms in this space should make synonymous terms look more similar

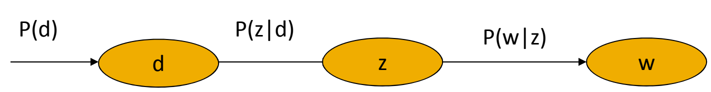


23

pLSI

- Latent Variable model for general co-occurrence data
  - Associate each observation (w,d) with a class variable z Є Z{z_1,…,z_K}

- Generative Model for document-term matrix D
  - Select a doc with probability P(d)
  - Pick a latent class z with probability P(z|d)
  - Generate a word w with probability p(w|z)



pLSI Pitfalls

- It is a generative model for the train data
- It can be easily overfitted to the train data

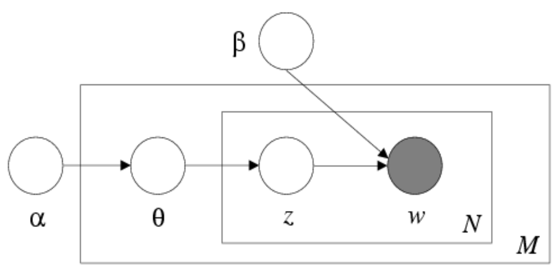
LDA

- Latent variable model

- Dirichlet Prior

- And that's All

Father of Topic Modeling



So, LDA uses Latent variable model, and Dirichlet Distribution priors **and that is All**

Dirichlet Distributions

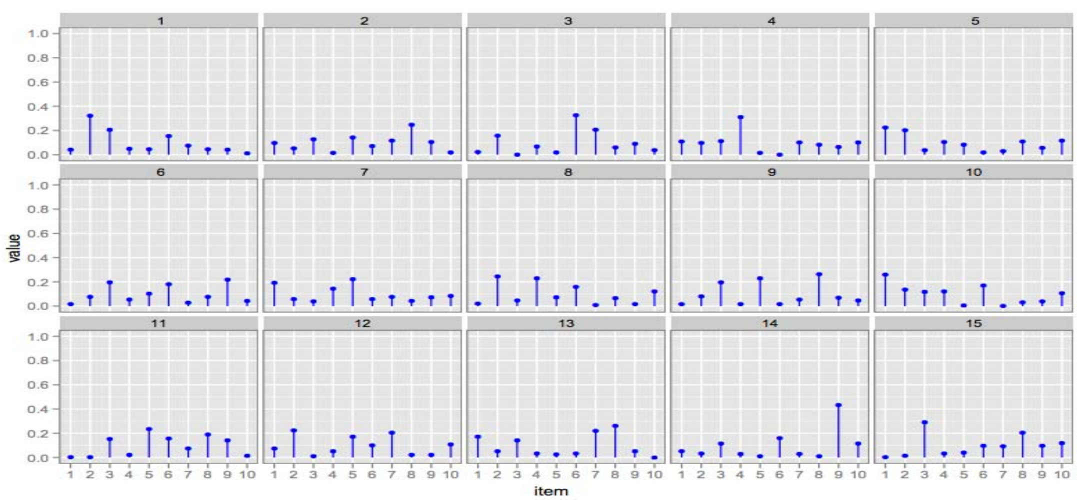
- Dirichlet distribution is the conjugate prior to the multinomial distribution. (This means that if our likelihood is multinomial with a Dirichlet prior, then the posterior is also Dirichlet!)
- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

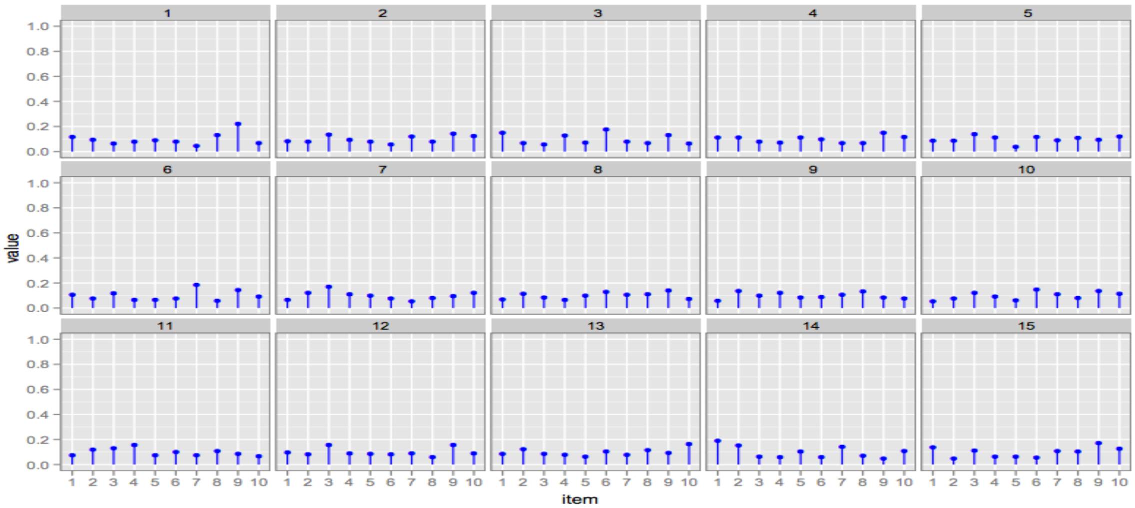
$$\alpha_0 = \sum_{i=1}^{k} \alpha_i \qquad E[\theta_i] = \frac{\alpha_i}{\alpha_0} \qquad Var[\theta_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

- The Dirichlet parameter α_i can be thought of as a prior count of the i^{th} class.
- The parameter α controls the mean shape and sparsity of θ .

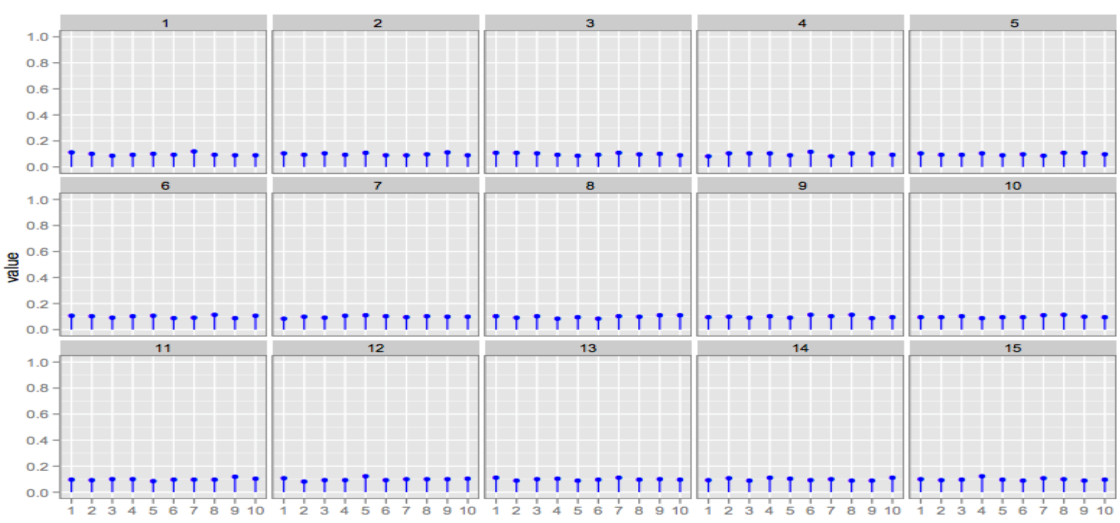
α = 1



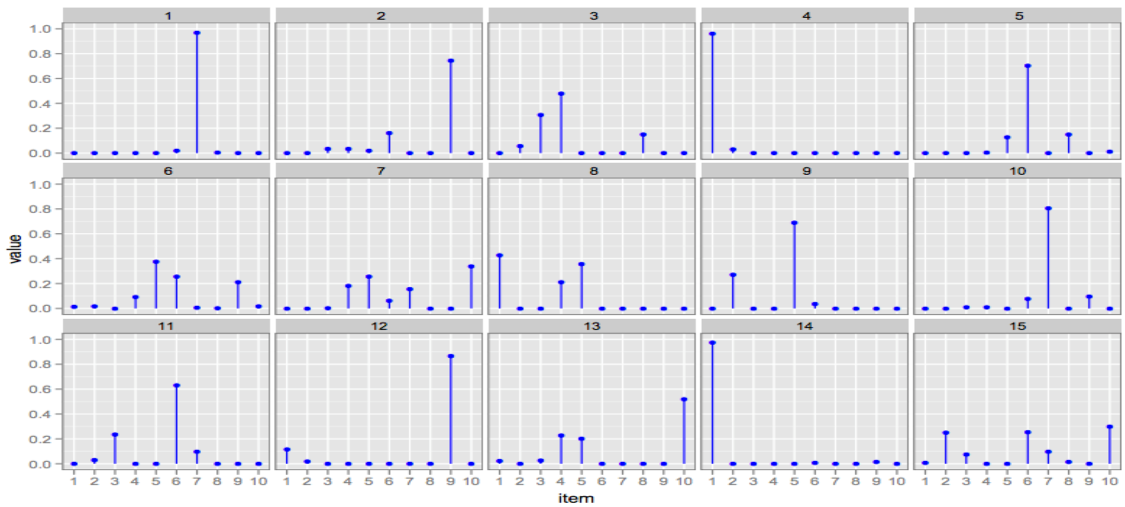
α = 10



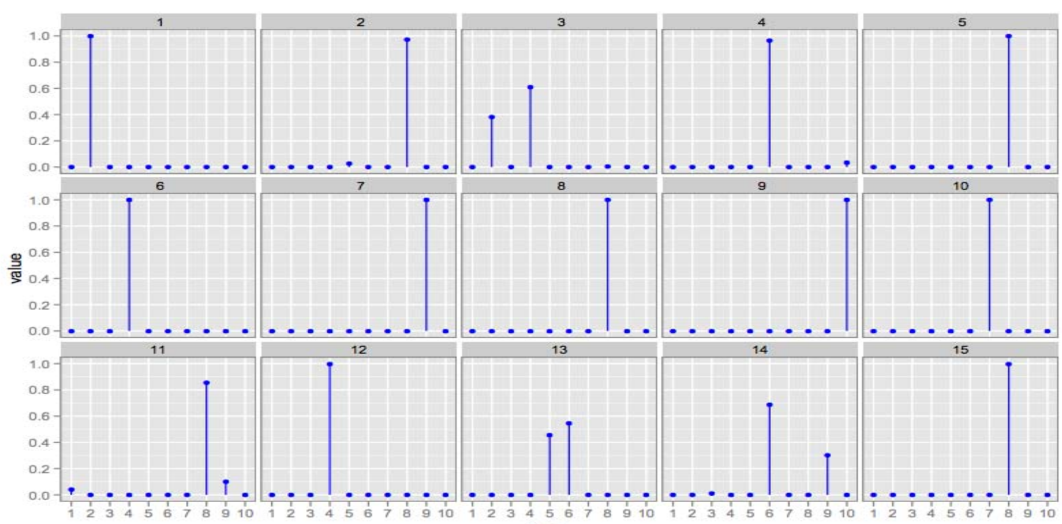
α = 100



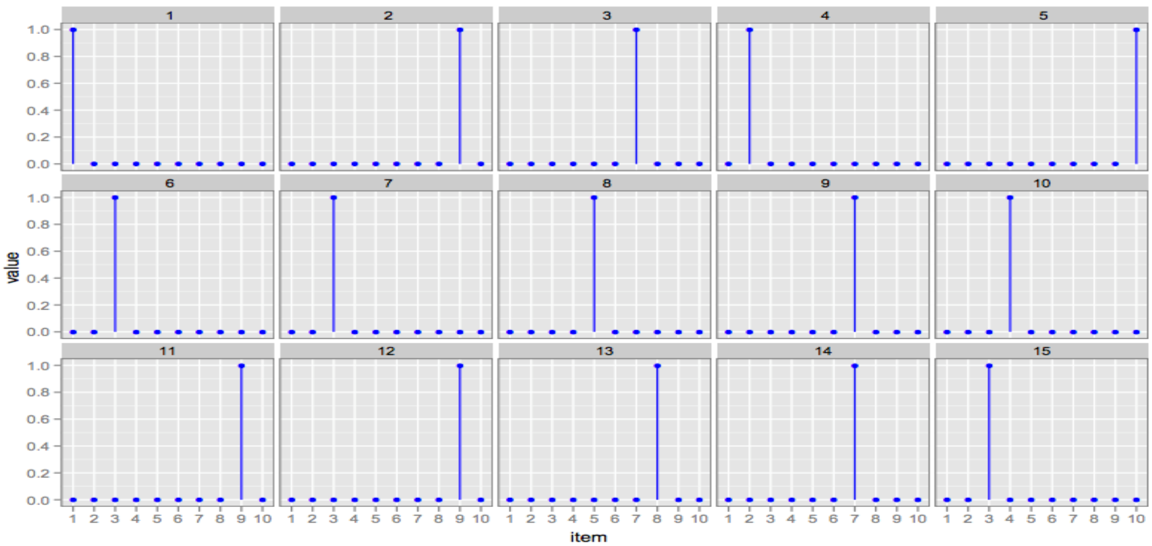
α = 0.1


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0 0.2 0.4 0.6 0.8 1.0
1 2 3 4 5 6 7 8 9 10

α = 0.01

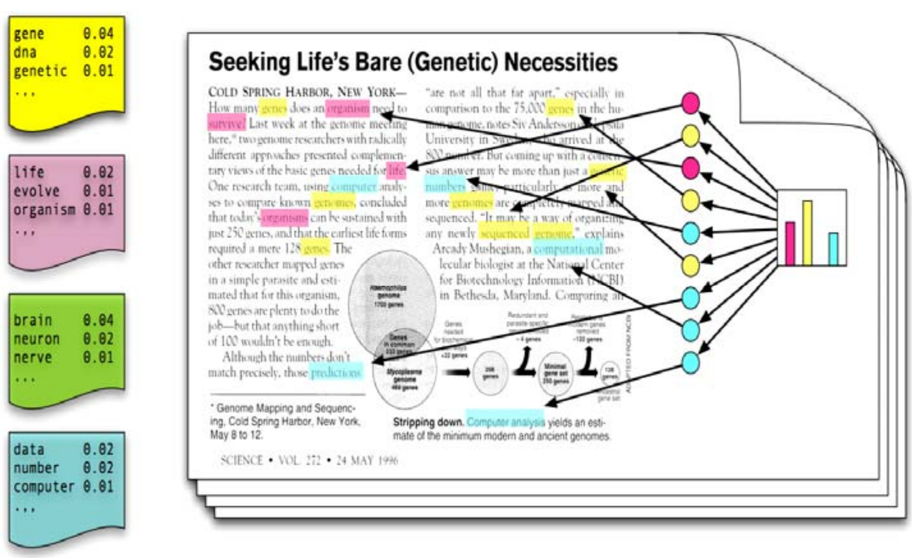

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0 0.2 0.4 0.6 0.8 1.0
1 2 3 4 5 6 7 8 9 10

α = 0.001



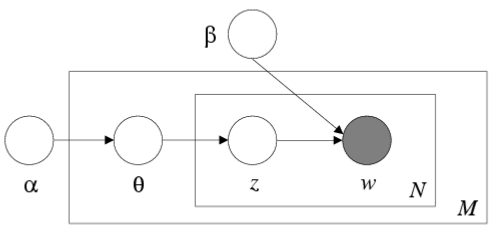
Basic Assumptions

- Each document is a mixture of corpus-wide topics
- Each topic is a distribution over words
- Each word is drawn from one of the topics



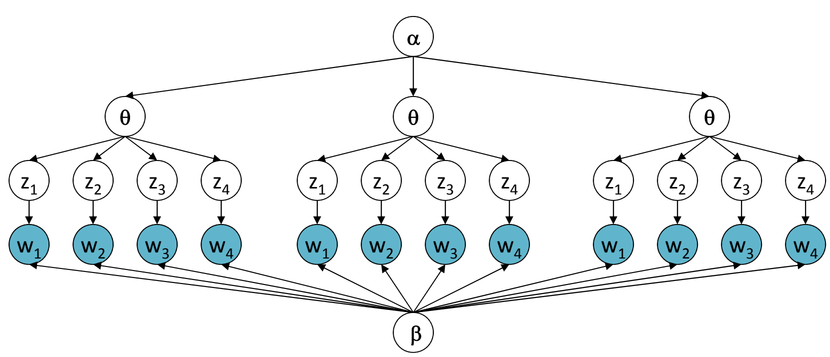
LDA – generative process

- For each document,
- Choose θ~Dirichlet(α)
- For each of the N words wn:
  - Choose a topic z_n» Multinomial(θ)
  - Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

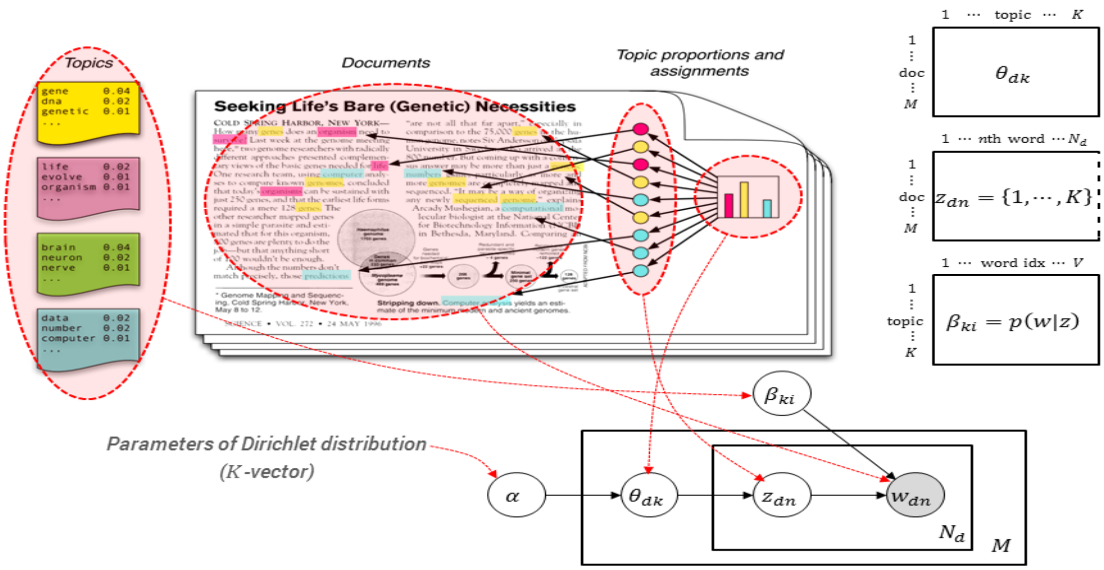


$$[\beta]_{k\times V} \quad \beta_{ij} = p(w^j = 1 | z^i = 1)$$

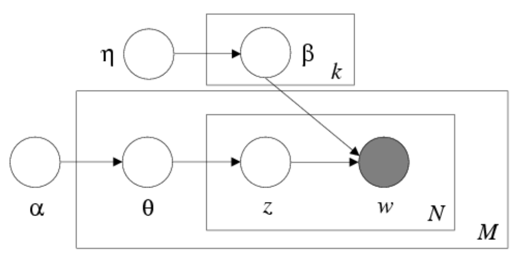
The LDA Model (Unpacked)



LDA Model



Smoothed LDA



- Introduces Dirichlet smoothing on β to avoid the zero frequency word problem
- Fully Bayesian approach

The LDA equations

Joint Probability

$$(2)\ p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

Marginal Distribution of a document

$$(3)\ p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d^k\theta$$

Probability of a corpus

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

More insights on LDA: Exchangeability

- A finite set of random variables $\{x_1, \ldots, x_N\}$ is said to be *exchangeable* if the joint distribution is invariant to permutation. If π is a permutation of the integers from 1 to N:

$$p(x_1, \ldots x_N) = p(x_{\pi(1)}, \ldots, x_{\pi(N)})$$

- An infinite sequence of random is *infinitely exchangeable* if every finite subsequence is exchangeable

bag-of-words Assumption

- Word order is ignored
- "bag-of-words" – exchangeability, not i.i.d
- **Theorem (De Finetti, 1935)** – if $(x_1, x_2, \ldots, x_N)$ are infinitely exchangeable, then the joint probability $p(x_1, x_2, \ldots, x_N)$ has a representation as a mixture:

For some random variable θ

$$p(x_1, x_2, \ldots, x_N) = \int d\theta \, p(\theta) \prod_{i=1}^{N} p(x_i | \theta)$$

LDA and exchangeability

- We assume that words are generated by topics and that those topics are infinitely exchangeable within a document.
- By de Finetti's theorem:

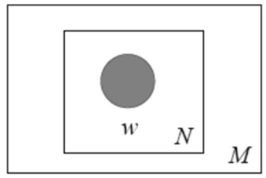
$$p(\mathbf{w}, \mathbf{z}) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n | \theta) \, p(w_n | z_n) \right) d\theta$$

- By marginalizing out the mixture component in eq 2, we get we will get the same distribution over observed and latent variables as above.

Relationship with other latent variable models

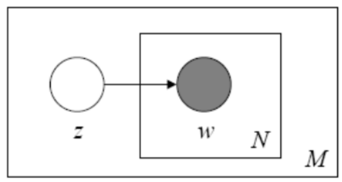
- Unigram model

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



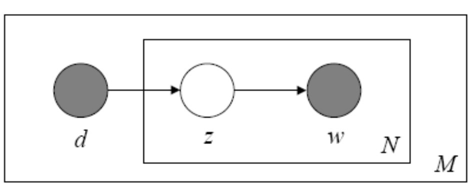
- Mixture of unigrams
  - Each document is generated by first choosing a topic z and then generating N words independently form conditional multinomial
  - k-1 parameters

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



Relationship with other latent variable models (cont.)

- Probabilistic latent semantic indexing
  - Attempt to relax the simplifying assumption made in the mixture of unigrams models
  - In a sense, it does capture the possibility that a document may contain multiple topics
  - kv+kM parameters and linear growth in M

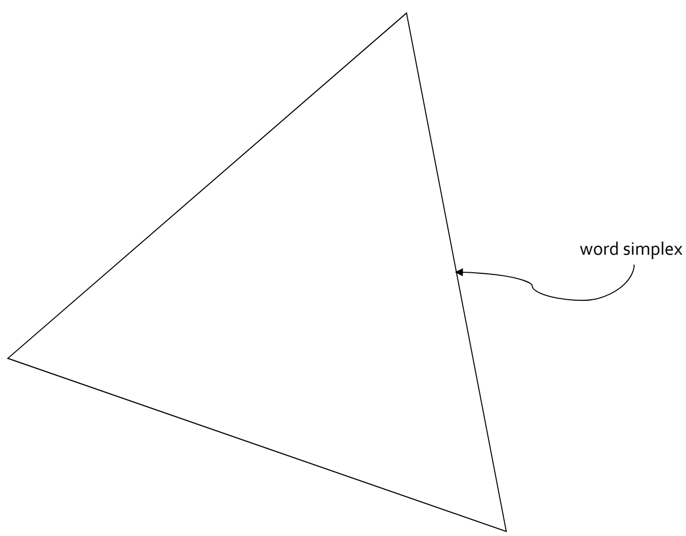


The k+kV parameters in a k-topic LDA model do not grow with the size of the training corpus.

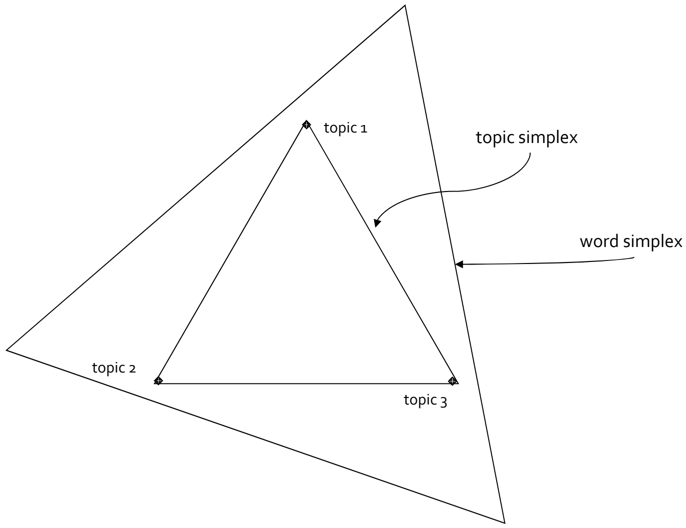
Relationship with other latent variable models (cont.)

- The unigram model find a single point on the word simplex and posits that all word in the corpus come from the corresponding distribution.
- The mixture of unigram models posits that for each documents, one of the k points on the word simplex is chosen randomly and all the words of the document are drawn from the distribution
- The pLSI model posits that each word of a training documents comes from a randomly chosen topic. The topics are themselves drawn from a document-specific distribution over topics.
- LDA posits that each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter

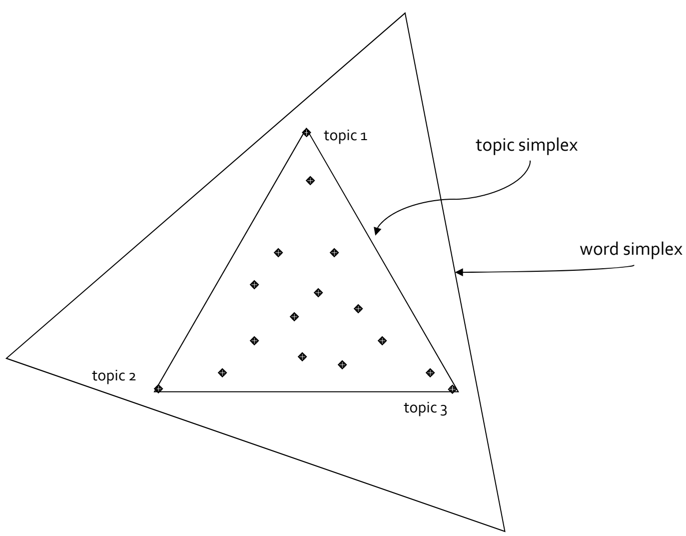
A geometric interpretation



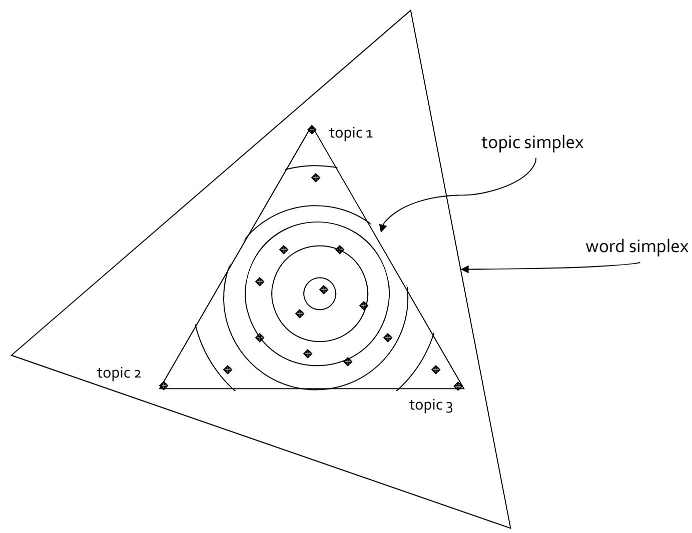
A geometric interpretation



A geometric interpretation



A geometric interpretation



Parameter Estimation

- Exact inference is not feasible
- Approximate methods
  - Gibbs Sampling
  - Variational inference
  - Collapsed Gibbs sampling

Gibbs Sampling

1. Initialize randomly the topic assignments
2. For each document "i" sample its topic mixture

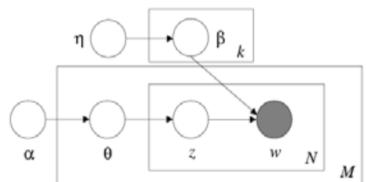
$$p(\theta_i|.) = Dir(\{\alpha_k + \sum_l I(z_{il} = k)\})$$

3. For each topic "k" sample from posterior of multinomial over vocabulary

$$p(\beta_k|.) = Dir(\{\gamma_v + \sum_i \sum_l I(w_{il} = v, z_{il} = k)\})$$

4. For each document "I" and each word "l" sample its topic assignment

$$p(z_{il} = k|.) \propto exp(\log \theta_{ik} + \log \beta_{kw_{il}})$$



Parameter Estimation (Variational EM)

- Since we have latent variable model we need to use EM

1. Find the expected value of the hidden variables (requires to run inference to compute $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$)

2. Use the expected counts to maximize the likelihood .

Inference and parameter estimation

- The key inferential problem is that of computing the posteriori distribution of the hidden variable given a document

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

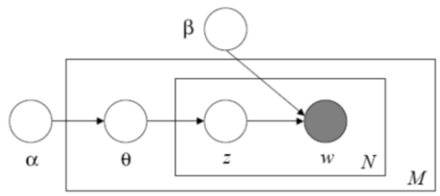
$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

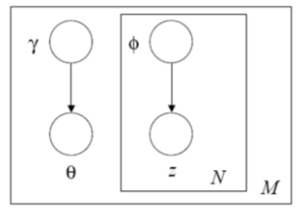
It is intractable to compute in general, due to the coupling between θ and β in the summation over latent topics

Variational Inference

- The basic idea of convexity-based variational inference is to make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood.
- Essentially, one considers a family of lower bounds, indexed by a set of variational parameters.
- A simple way to obtain a tractable family of lower bound is to consider simple modifications of the original graph model in which some of the edges and nodes are removed.

Variational Inference (cont.)





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

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$$

In variational inference, we consider a simplified graphical model with variational parameters γ, ϕ and minimize the KL Divergence between the variational and posterior distributions.

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} KL(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta))$$

LDA: Topic Illustration

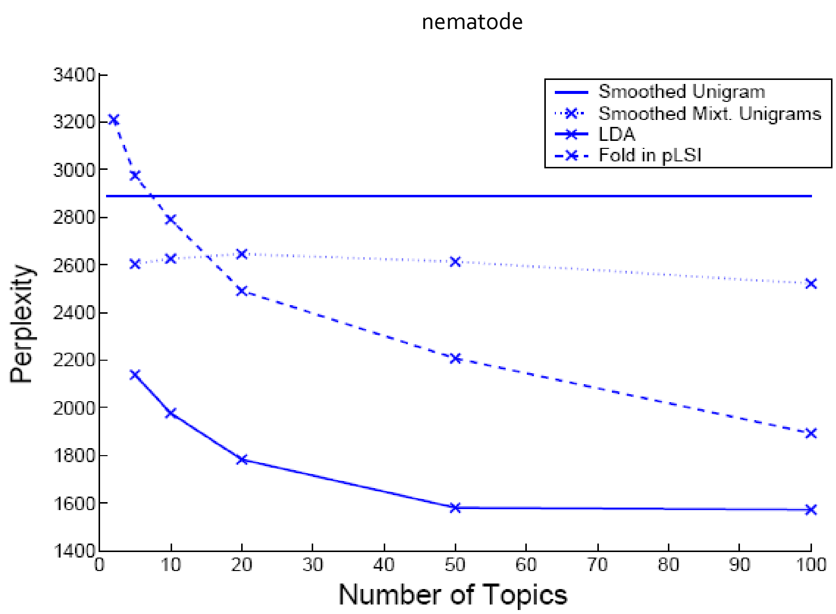
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY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

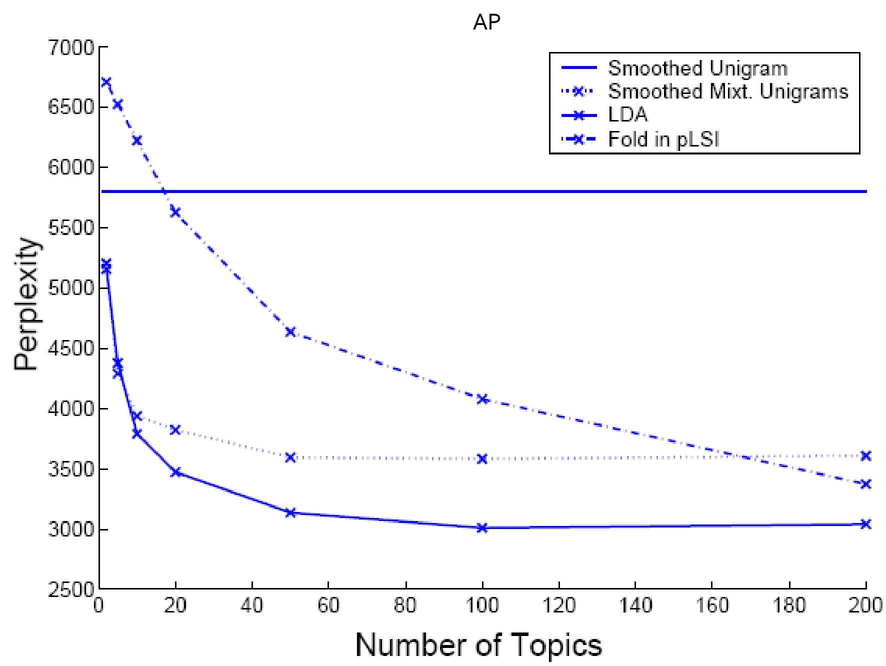
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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.

Application: Document modeling

- Unlabeled data – our goal is density estimation.
- Compute the *perplexity* of a held-out test to evaluate the models – lower perplexity score indicates better generalization.

$$perplexity(D_{test}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d} \right\}$$



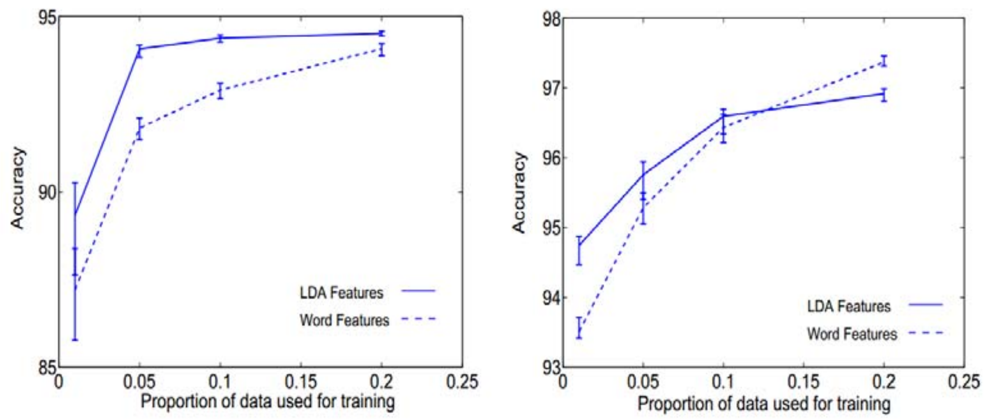


Document Modeling – cont. Results

- Both pLSI and mixture suffer from overfitting.
- pLSI – overfitting due to dimensionality of the p(z|d) parameter.

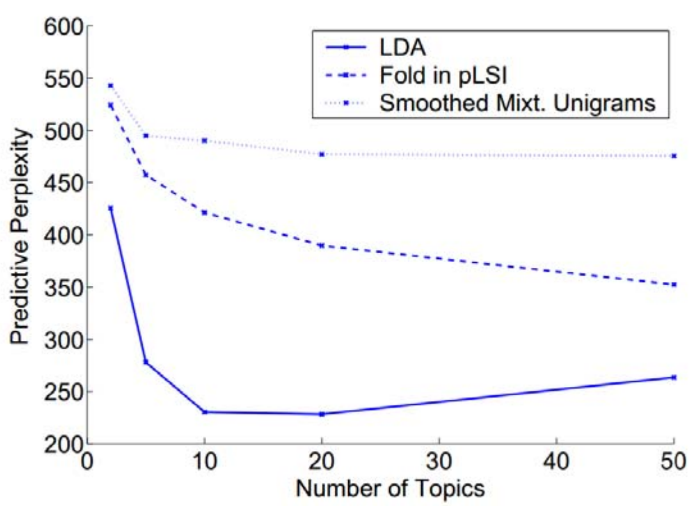
Num. topics (k)	Perplexity	
	Mult. Mixt.	pLSI
2	22,266	7,052
5	2.20×10^{8}	17,588
10	1.93×10^{17}	63.800
20	1.20×10^{22}	2.52×10^{5}
50	4.19×10^{106}	5.04×10^{6}
100	2.39×10^{150}	1.72×10^{7}
200	3.51×10^{264}	1.31×10^{7}

Application: Classification (Dimensionality Reduction)



(a) (b)

Application: Recommendation Systems



Number of Topics K

- Cross validation, using log likelihood on the test set
- Use variational lower bound as a proxy for log p(D|K)
- Use non-parametric Bayesian Methods (The et al. 2006)
- Use annealed importance sampling to approximate the evidence (Wallach et al. 2009)

LDA Extensions

- Correlated Topic Model (Blei and Lafferty 2007)
- Supervised LDA (Blei and McAlliffe 2010)
- Dynamic Topic Model (Blei and Lafferty 2006)
- LDA-HMM (Griffiths et al. 2004)
- ……

Summary

- LDA is a flexible generative probabilistic model for collection of discrete data.
- Arguably, could be considered as the best possible model based on the Bag of Word assumption
- Can be viewed as a dimensionality reduction technique
- Exact inference is intractable, however it is possible to use approximate inference instead
- Can be used in other collection, e.g. images, collaborative filtering, …
- There are lots of extensions and applications to LDA

References

- Michael W. Berry, Zlatko Drmac, Elizabeth R. Jessup. Matrices, Vector Spaces, and Information Retrieval, SIAM Review, 1999.
- Thomas Hofmann, *Probabilistic Latent Semantic Indexing*, Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99), 1999
- **Latent Dirichlet allocation.** D. Blei, A. Ng, and M. Jordan. Journal of Machine Learning Research, 3:993-1022, January 2003.
- **Latent Dirichlet allocation,** presentation Slides, David M. Blei , Princeton University
- **Finding Scientific Topics.** Griffiths, T., & Steyvers, M. (2004). Proceedings of the National Academy of Sciences, 101 (suppl. 1), 5228-5235.
- **Latent Dirichlet allocation** presentation Slides, Ido Abramovich, Hebrew University
- LSI, PLSI, LDA presentation slides, Alexander Yates et al. , Temple University

Thanks