ICCS413 - Lecture 23

Topic Models

Sunsern Cheamanunkul

Adapted from LDA paper by David M. Blei, Andrew Y. Ng, Michael I. Jordan.

Mahidol University
International College

Outline

- Introduction
- Basic Topic Models
  - Unigram
  - Mixture of unigrams
  - Probabilistic Latent Semantic Indexing
  - Latent Dirichlet Allocation
  - Correlated Topic Models
- Example

Overview

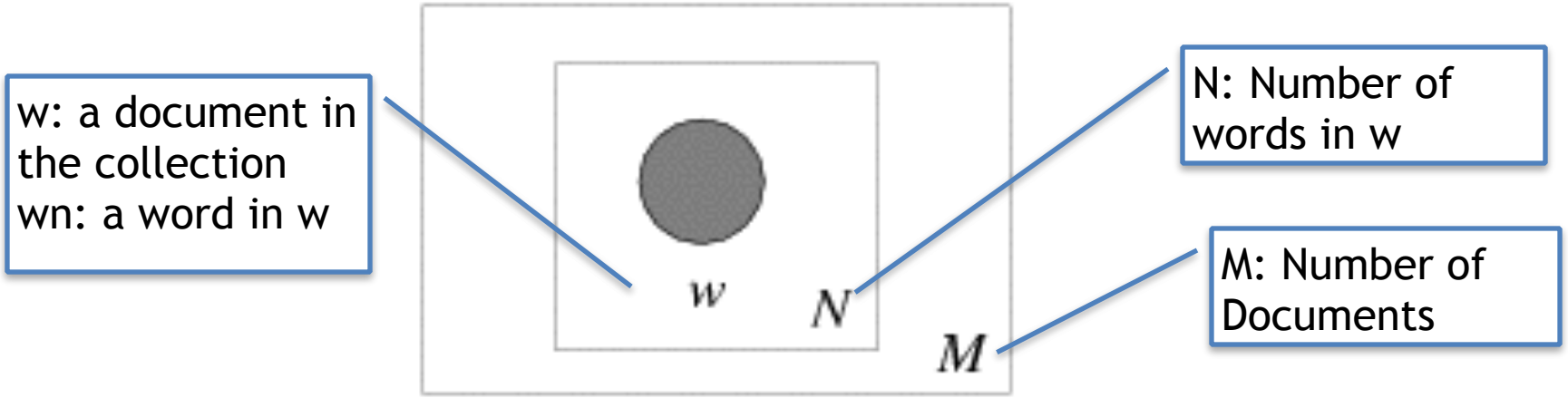
- Motivation:
  - Model the topic/subtopics in text collections
- Basic Assumptions:
  - There are k topics in the whole collection
  - Each topic is represented by a multinomial distribution over the vocabulary (language model)
  - Each document can cover multiple topics
- Applications
  - Summarizing topics
  - Predict topic coverage for documents
  - Model the topic correlations

Basic Topic Models

- Generative Models:
  - Unigram model
  - Mixture of unigrams
  - Probabilistic LSI
  - LDA

Unigram Model

- There is only one topic in the collection

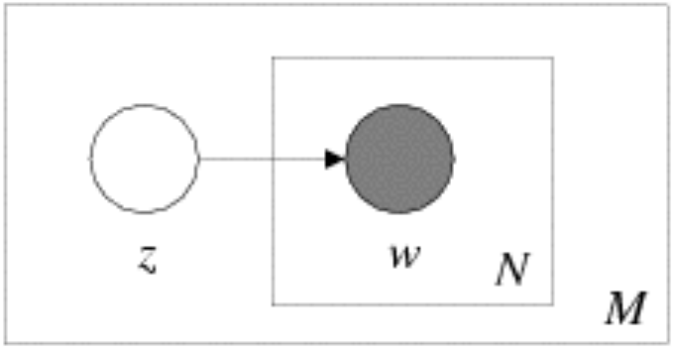


$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n)$$

- Estimation:
  - Maximum likelihood estimation

Mixture of unigrams

- There is k topics in the collection, but each document only cover one topic



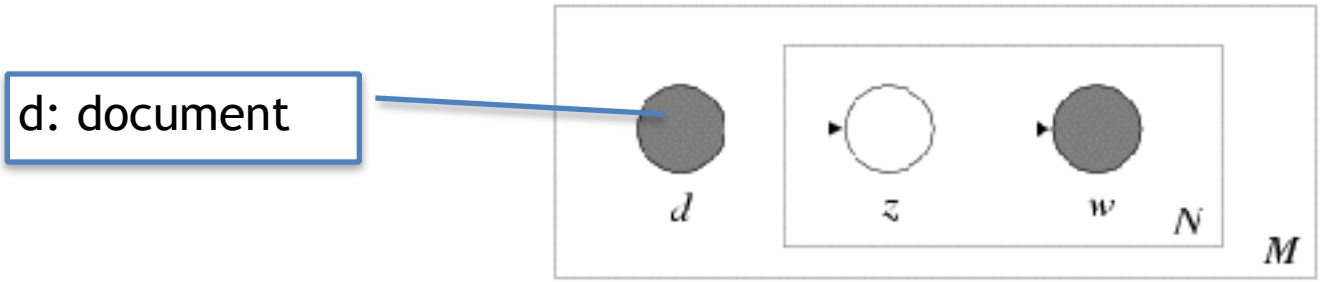
$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z)$$

- Estimation: MLE and EM Algorithm

Probabilistic Latent Semantic Indexing

- (Thomas Hofmann ’99): each document is generated from more than one topics, with a set of document specific mixture weights {p(z|d)} over k topics.
- These mixture weights are considered as fixed parameters to be estimated.
- Also known as aspect model.
- No prior knowledge about topics required, context and term co-occurrences are exploited

PLSI (cont.)



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

- Assume uniform p(d)
- Parameter Estimation:
  - π: {p(z|d)}; θ: {p(w|z)}
  - Maximizing log-likelihood using EM algorithm

Problem of PLSI

- Mixture weights are considered as document specific, thus no natural way to assign probability to a previously unseen document.
- Number of parameters to be estimated grows with size of training set, thus overfits data, and suffers from multiple local maxima.
- Not a fully generative model of documents.

Latent Dirichlet Allocation

- LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words.

Latent Dirichlet Allocation

- (Blei et al ‘03) Treats the topic mixture weights as a k-parameter hidden random variable (a multinomial) and places a Dirichlet prior on the multinomial mixing weights. This is sampled once per document.
- The weights for word multinomial distributions are still considered as fixed parameters to be estimated.
- For a fuller Bayesian approach, can place a Dirichlet prior to these word multinomial distributions to smooth the probabilities. (like Dirichlet smoothing)

LDA - generative process

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

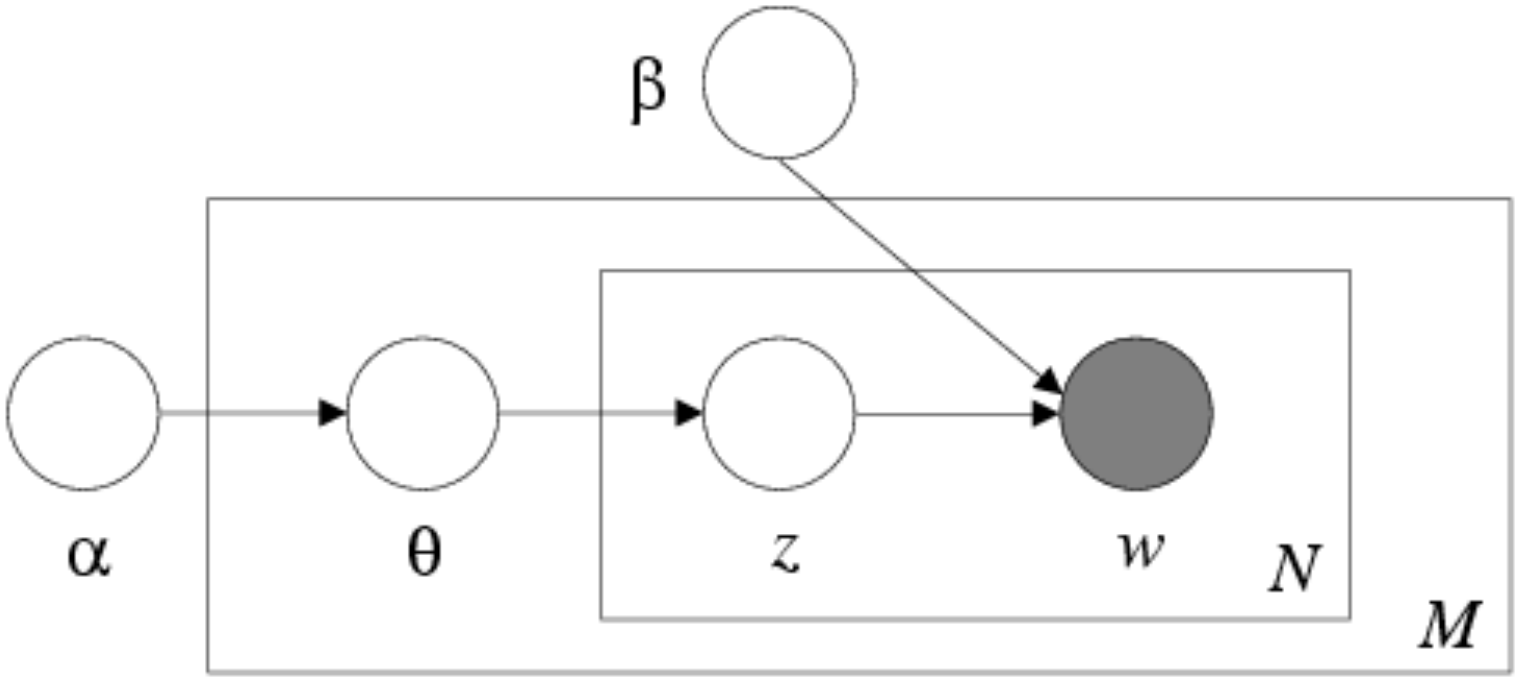
$$[\beta]_{k\times V} \quad \beta_{ij} = p(w^j = 1|z^i = 1)$$

Dirichlet distribution

- A *k*-dimensional Dirichlet random variable θ can take values in the (k-1)-simplex, and has the following probability density on this simplex:

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

The graphical model



The LDA equations

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

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

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

Inference

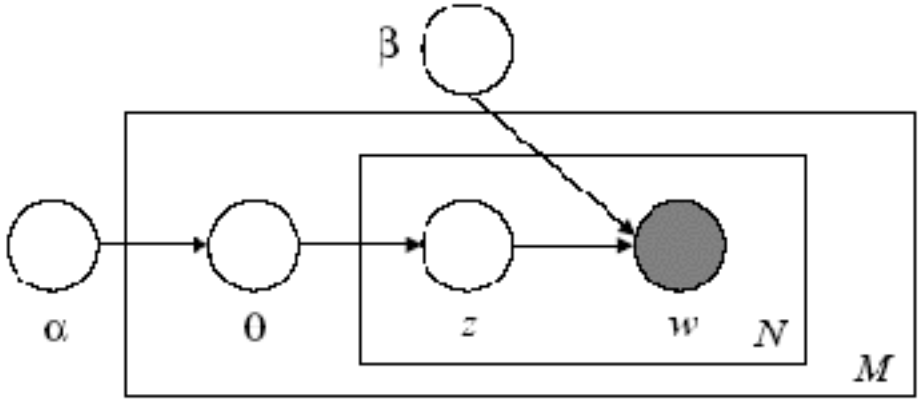
- We want to compute the posterior dist. of the hidden variables given a document:

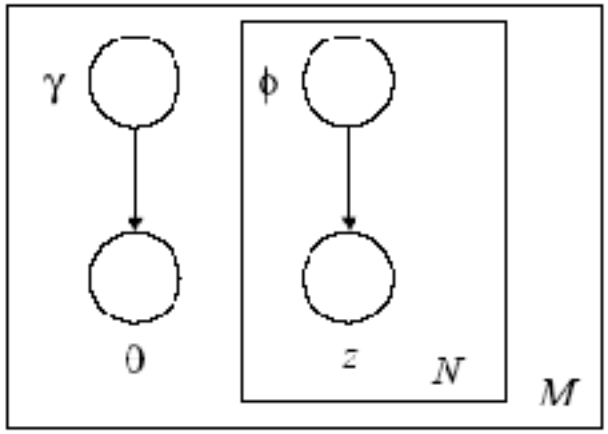
$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

- Unfortunately, this is intractable to compute in general.

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

Variational inference





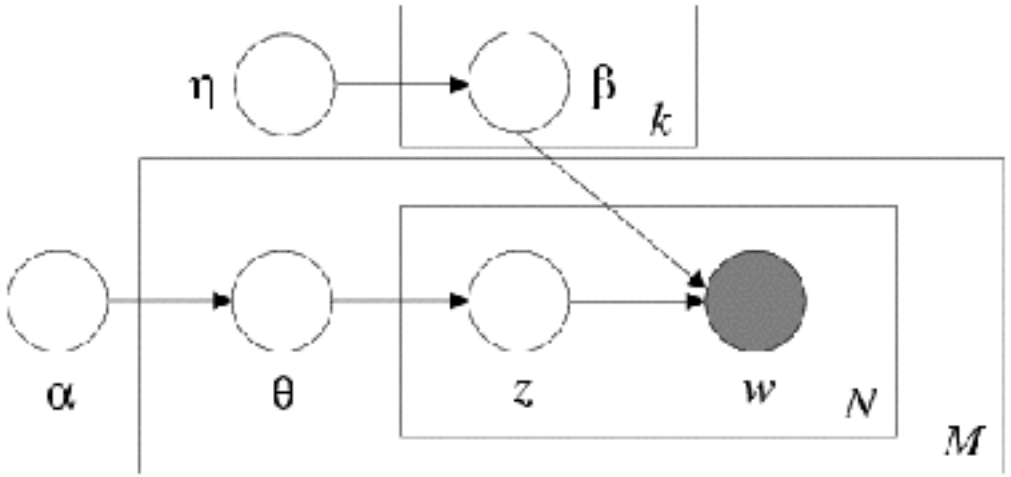
$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \phi) \prod_{n=1}^{N} q(z_n \mid \phi_n)$$

Parameter estimation

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

- Variational EM
  - (E Step) For each document, find the optimizing values of the variational parameters (γ, φ) with α, β fixed.
  - (M Step) Maximize variational distribution w.r.t. α, β for the γ and φ values found in the E step.

Smoothed LDA



- Introduces Dirichlet smoothing on ß to avoid the "zero frequency problem"
- More Bayesian approach
- Inference and parameter learning similar to unsmoothed LDA

"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.

Document modeling

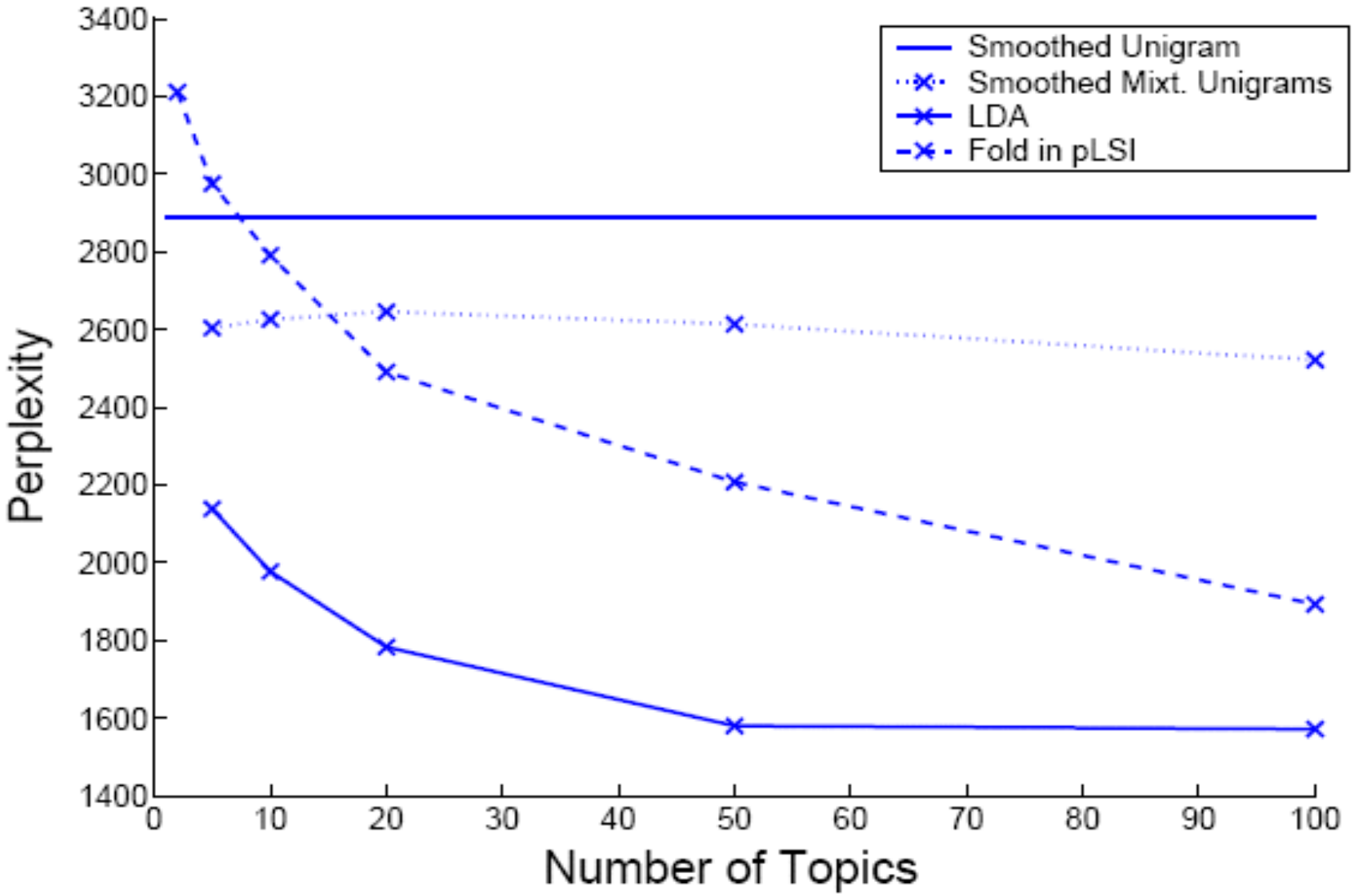
- Unlabeled data – our goal is density estimation.
- Compute the perplexity of a held-out test to evaluate the models – lower perplexity score indicates better generalization.

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

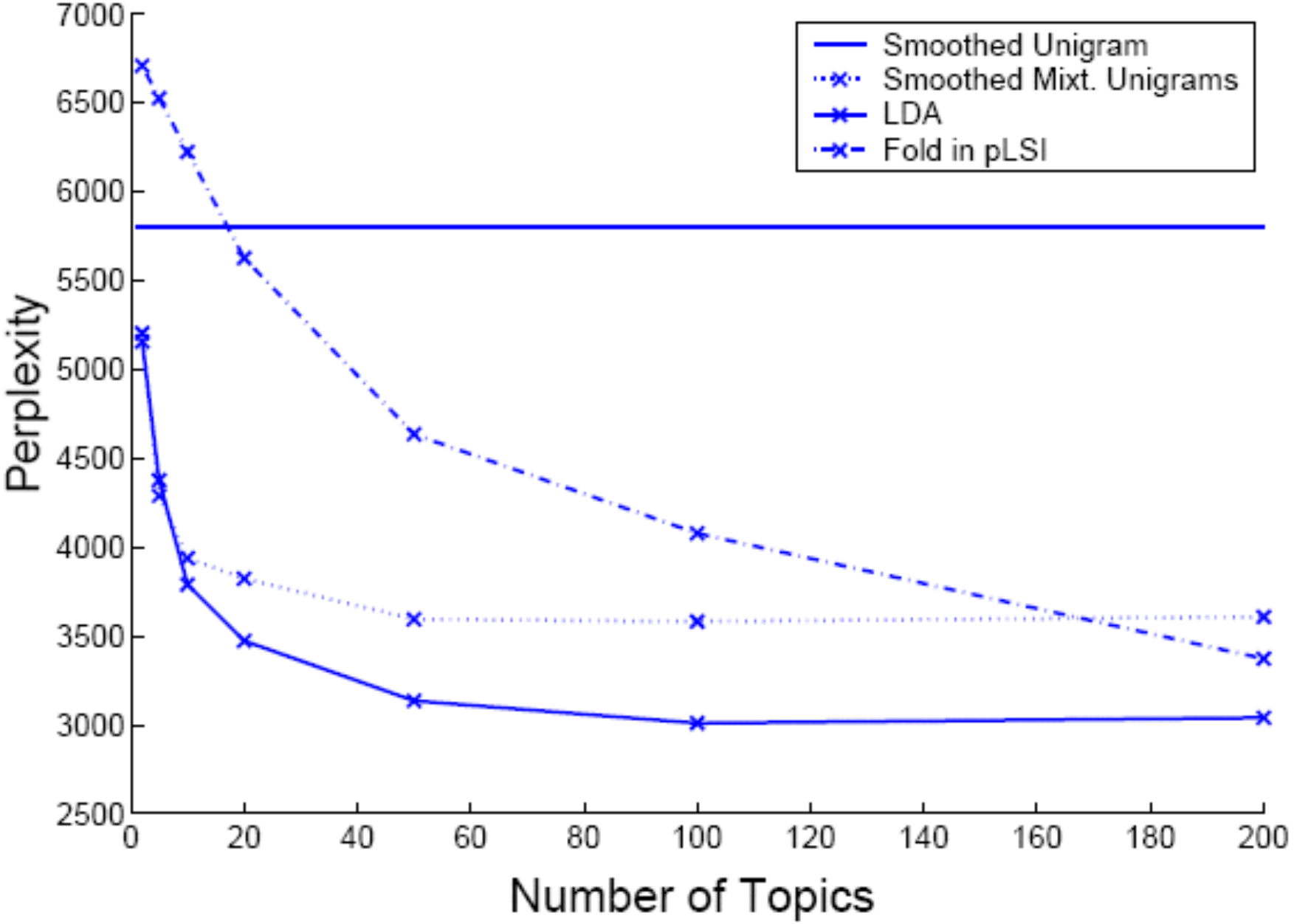
Document Modeling - cont. data used

- C. Elegans Community abstracts
  - 5,225 abstracts
  - 28,414 unique terms
- TREC AP corpus (subset)
  - 16,333 newswire articles
  - 23,075 unique terms
- Held-out data – 10%
- Removed terms – 50 stop words, words appearing once (AP)

nematode


Smoothed Unigram
Smoothed Mixt. Unigrams
LDA
Fold in pLSI
Perplexity
Number of Topics
3400
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
0
10
20
30
40
50
60
70
80
90
100

AP


Smoothed Unigram
Smoothed Mixt. Unigrams
LDA
Fold in pLSI
Perplexity
Number of Topics
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
0
20
40
60
80
100
120
140
160
180
200

Document Modeling – cont. Results

- Both pLSI and mixture suffer from overfitting.
- Mixture – peaked posteriors in the training set.
- Can solve overfitting with variational Bayesian smoothing.

Num. topics (k)	*Perplexity*	
	.Mult. Mixt	pLSI
2	22,266	7,052
5	***x 108 2.20***	17,588
10	***x 1.93 1017***	63.800
20	***x 1.20 1022***	***x 105 2.52***
50	***x 4.19 10106***	***x 106 5.04***
100	***x 2.39 10150***	***x 107 1.72***
200	***x 3.51***	***x 107 1.31***

Document Modeling - cont. Results

- Both pLSI and mixture suffer from overfitting.
- pLSI – overfitting due to dimensionality of the p(z|d) parameter.
- As k gets larger, the chance that a training document will cover all the topics in a new document decreases

Num. topics (k)	*Perplexity*	
	.Mult. Mixt	pLSI
2	22,266	7,052
5	***x 108 2.20***	17,588
10	***x 1.93 1017***	63.800
20	***x 1.20 1022***	***x 105 2.52***
50	***x 4.19 10106***	***x 106 5.04***
100	***x 2.39 10150***	***x 107 1.72***
200	***x 3.51***	***x 107 1.31***

Other uses

Corr-LDA:

TREE, LIGHT, SUNSET, WATER, SKY

GM-Mixture:

CLOSE-UP, TREE, PEOPLE, MUSHROOMS, LICHEN

GM-LDA:

WATER, SKY, TREE, PEOPLE, GRASS

Corr-LDA:

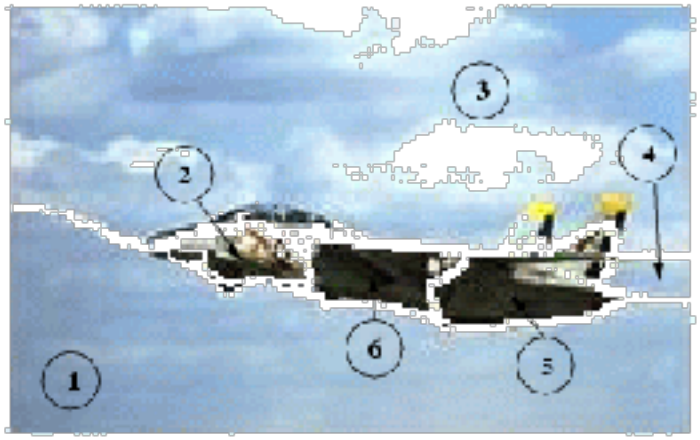
TREE, WATER, GRASS, FLOWERS, BIRDS

GM-Mixture:

TREE, WATER, GRASS, SKY, FIELD

GM-LDA:

WATER, SKY, TREE, PEOPLE, GRASS



Corr-LDA:

1. PEOPLE, TREE
2. SKY, JET
3. SKY, CLOUDS
4. SKY, MOUNTAIN
5. PLANE, JET
6. PLANE, JET

GM-LDA:

1. HOTEL, WATER
2. PLANE, JET
3. TUNDRA, PENGUIN
4. PLANE, JET
5. WATER, SKY
6. BOATS, WATER

Summary

- Based on the exchangeability assumption
- Can be viewed as a dimensionality reduction technique
- Exact inference is intractable, we can approximate instead
- Can be used in other collection – images and caption for example.