ST720 Data Science

Latent Dirichlet Allocation II

Seung Jun Shin (sjshin@korea.ac.kr)

Department of Statistics

LDA

Journal of Machine Learning Research 3 (2003) 993-1022 Submitted 2/02; Published 1/03

Latent Dirichlet Allocation

David M. Blei BLEI@CS.BERKELEY.EDU
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA

Goal

- The problem of modeling text corpora.
- The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships
- Find a generative model of the text data.

Notation and terminology

- Word ($\mathbf{w}$): Basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \cdots, V\}$. i.e., V-dimensional unit vector.

  The vth word in the vocabulary is represented by a V-dimensional vector $\mathbf{w}$ whose vth element is 1 and 0 for all others.

- Document ($\mathbf{W}$): collection of N words $\mathbf{W} = \{\mathbf{w}_n \mid n = 1 \cdots, N\}$.

- Corpus ($\mathcal{D}$): collection of M documents $\mathcal{D} = \{\mathbf{W}_d \mid 1 \cdots, M\}$.

Dirichlet Distribution

- A k-dimensional Dirichlet random variable $\boldsymbol{\theta}$ can take values in the $(k-1)$-simplex $(\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1)$ and has the following density:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1 \cdots \theta_k$$

  where the parameter $\boldsymbol{\alpha}$ is a k-dimensional vector with components $\alpha_i > 0$.
- When k is 2 Dirichlet reduces to beta distribution.
- Dirichlet distribution is a conjugate prior for Multinomial while Beta is for Binomial.

Model Specification I

- LDA is a generative probabilistic model of $\mathcal{D}$

$$P(\mathcal{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are model parameters to be estimated.

  - $\boldsymbol{\alpha}$ determines probabilistic behavior of topics.
  - $\boldsymbol{\theta}$ determines probabilistic behavior of topic given a document.
  - $\boldsymbol{\beta}$ determines probabilistic behavior of words given a topic.

Model Specification II

- Two principles of LDA.
  - Every document is a mixture of topics: Each document may contain words from several topics in particular proportions.
  - Every topic is a mixture of words.

Model Specification III

- For each document,

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha}), \qquad d = 1, \cdots, M$$

where $\boldsymbol{\theta}_d$ is a document-specific vector of the topic probabilities. (`gamma` in the previous lecture)

- For each word in the document, $\mathbf{w}_{dn}$:
  1. Choose a topic, $\mathbf{z}_{dn}$

$$\mathbf{z}_{dn} \in \mathbb{R}^k \sim \text{Multinomial}(\boldsymbol{\theta}_d)$$

  2. The word $\mathbf{w}_{dn}$ is generated from

$$p(\mathbf{w}_{dn} \mid \mathbf{z}_{dn}, \boldsymbol{\beta})$$

where $\boldsymbol{\beta} = \{\beta_{ij}\}$ with $\beta_{ij} = P(w^j = 1 \mid z = i)$.

Likelihood I

- Joint distribution of $(\boldsymbol{\theta}_d, \mathbf{Z}_d, \mathbf{W}_d)$:

$$
\begin{aligned}
& p(\boldsymbol{\theta}_d, \mathbf{Z}_d, \mathbf{W}_d \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
= \; & p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \times p(\mathbf{Z}_d \mid \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \times p(\mathbf{W}_d \mid \mathbf{Z}_d, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
= \; & p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \times \prod_{n=1}^{N} \left\{ p(\mathbf{z}_{dn} \mid \boldsymbol{\theta}) \times p(\mathbf{w}_{dn} \mid \mathbf{z}_{dn}, \boldsymbol{\beta}) \right\}
\end{aligned}
$$

  where $\mathbf{Z}_d = (\mathbf{z}_{d1}, \cdots, \mathbf{z}_{dN})^T$.

- Integrating over both $\boldsymbol{\theta}_d$ and $\mathbf{Z}_d$, we have

$$
\begin{aligned}
& p(\mathbf{W}_d \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
= & \int p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} \left[\sum_{\mathbf{z}_{dn}} \{ p(\mathbf{z}_{dn} \mid \boldsymbol{\theta}_d) p(\mathbf{w}_{dn} \mid \mathbf{z}_{dn}, \boldsymbol{\beta}) \} \right] d\boldsymbol{\theta}_d
\end{aligned}
$$

Likelihood II

- Corpus $\mathcal{D} = (\mathbf{W}_1, \cdots, \mathbf{W}_M)$ is a collection of M documents, thus its likelihood is

$$
\begin{aligned}
& p(\mathcal{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \prod_{d=1}^{M} p(\mathbf{W}_d \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \prod_{d=1}^{M} \left(\int p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} \left[\sum_{\mathbf{z}_{dn}} \{ p(\mathbf{z}_{dn} \mid \boldsymbol{\theta}_d) p(\mathbf{w}_{dn} \mid \mathbf{z}_{dn}, \boldsymbol{\beta}) \} \right] d\boldsymbol{\theta}_d \right)
\end{aligned}
$$

Model Specification

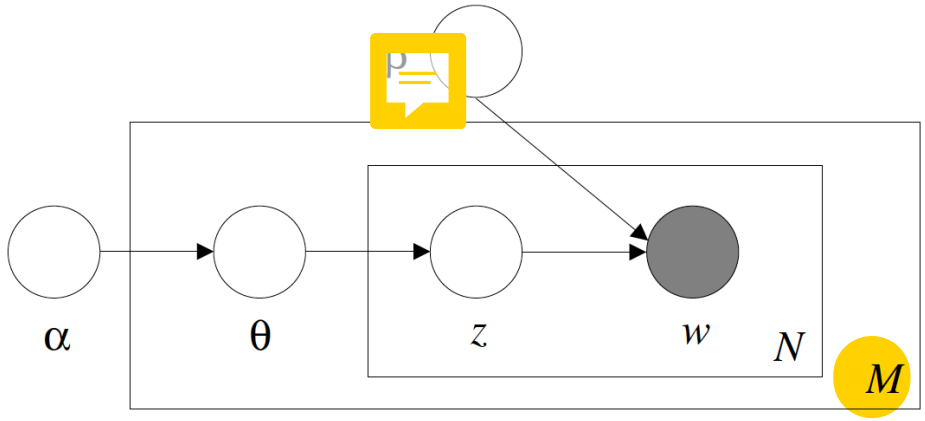


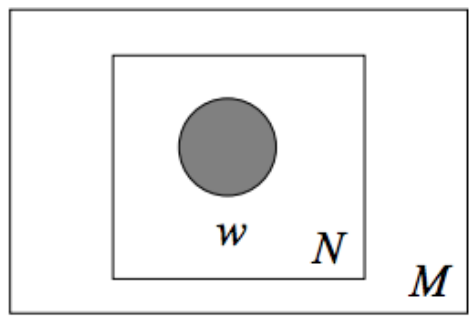
Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

- There are three levels.
- $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is corpus-level parameter.
- $\boldsymbol{\theta}_d$ is document-level.
- $\mathbf{z}_{dn}$ and $\mathbf{w}_{dn}$ are word-level.
- In LDA, documents can be associated with multiple topics.

Relationship with Other Models

- Unigram Model: Words of every document are drawn independently from a single multinomial distribution:

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



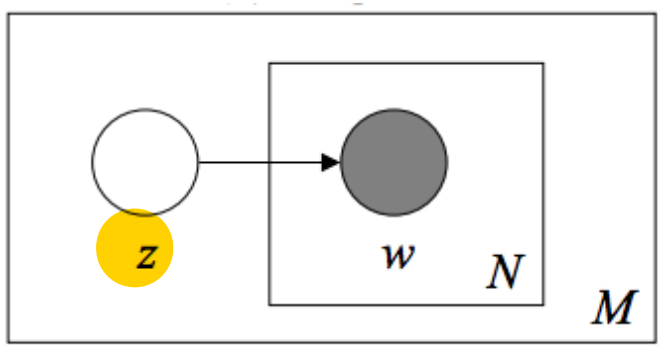
(a) unigram

Relationship with Other Models

- Mixture of Unigrams: Each document is generated by a single topic $\mathbf{z}$ and then generating N words independently from $p(\mathbf{w}_n \mid \mathbf{z})$.

$$P(\mathbf{W}) = \sum_{\mathbf{z}} p(\mathbf{z}) \prod_{i=1}^{N} p(\mathbf{w}_n \mid \mathbf{z})$$

Each document exhibits exactly one topic.



(b) mixture of unigrams

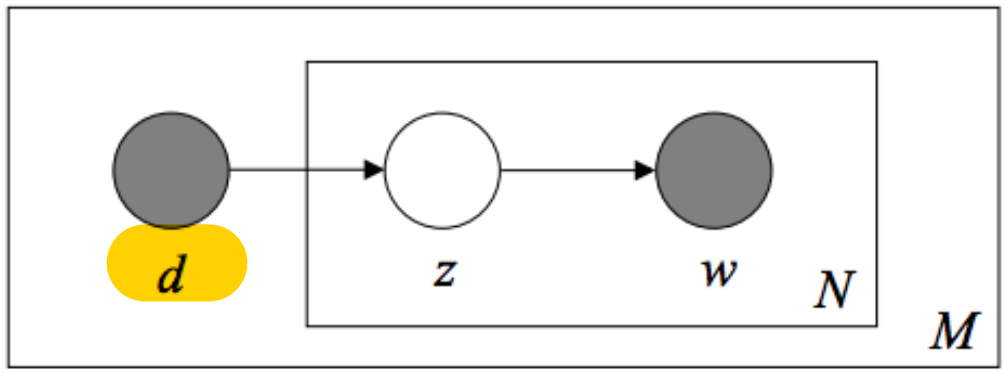
Relationship with Other Models I

- Probabilistic Latent Semantic Indexing (pLSI):
  - Let $\mathbf{d}$ be (dummy) index assigned to training documents.
  - pLSI assumes the conditional independent between $\mathbf{d}$ and $\mathbf{w}_n$ given an unobserved topic $\mathbf{z}$.

$$p(\mathbf{d}, \mathbf{w}_n) = p(\mathbf{d}) \sum_{\mathbf{z}} p(\mathbf{w}_n \mid \mathbf{z}) p(\mathbf{z} \mid \mathbf{d})$$

Relationship with Other Models II

- A document may contain multiple topics.
- For unseen documents, one cannot define $p(\mathbf{z} \mid \mathbf{d})$.
- There are $(kV + kM)$ parameters and therefore linearly growth in M.



(c) pLSI/aspect model

Relationship with Other Models III

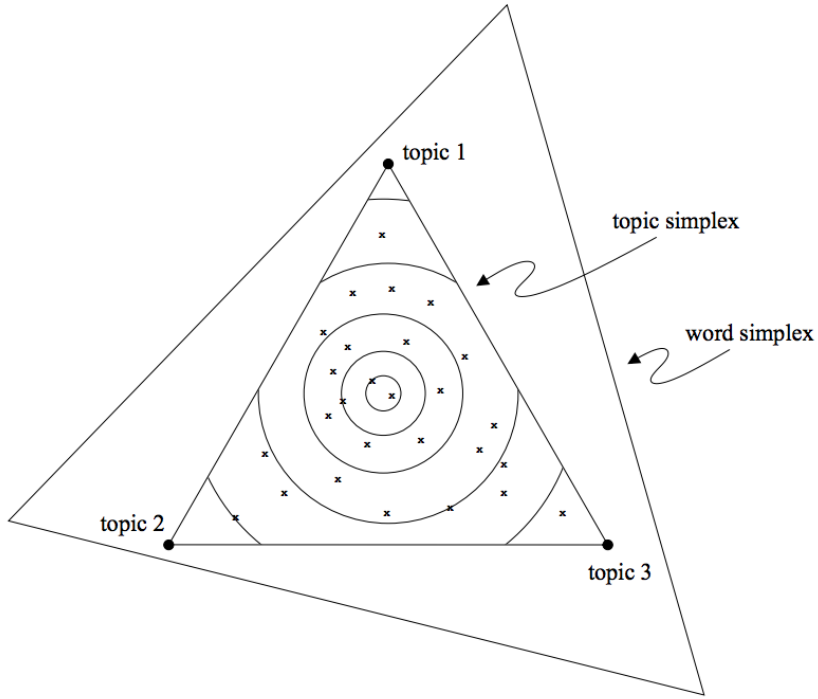


Figure: Graphical Comparison: The mixture of unigrams places each document at one of the corners of the topic simplex. The pLSI model induces an empirical distribution on the topic simplex denoted by x. LDA places a smooth distribution on the topic simplex denoted by the contour lines.

Parameter estimation of LDA I

- Need to maximize

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{d=1}^{M} \log p(\mathbf{W}_d \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- For any distribution $q(\boldsymbol{\theta}, \mathbf{z})$ given,

$$\begin{aligned}
\log p(\mathbf{W} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \log \int \sum_{\mathbf{z}} p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{W} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{W} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) q(\boldsymbol{\theta}, \mathbf{z})}{q(\boldsymbol{\theta}, \mathbf{z})} d\boldsymbol{\theta} \\
&\geq \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{W} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} - \\
&\qquad \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \log q(\boldsymbol{\theta}, \mathbf{z}) d\boldsymbol{\theta} \\
&= E_q[\log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})] - E_q[\log q(\boldsymbol{\theta}, \mathbf{z})] \qquad (1)
\end{aligned}$$

Parameter estimation of LDA II

- The difference is

$$
\begin{aligned}
&\log p(\mathbf{W} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) - E_q[\log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})] + E_q[\log q(\boldsymbol{\theta}, \mathbf{z})] \\
=&E_q[\log q(\boldsymbol{\theta}, \mathbf{z})] - E_q[\log p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})] \\
=&E_q\left[\log \frac{q(\boldsymbol{\theta}, \mathbf{z})}{p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})}\right] \\
=&KL\left[q(\boldsymbol{\theta}, \mathbf{z}) \mid p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})\right]
\end{aligned}
$$

- To get the tightest lower bound of $\log p(\mathbf{W} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$, solve

$$
\operatorname*{argmin}_{q \in \mathcal{F}} KL\left[q(\boldsymbol{\theta}, \mathbf{z}) \mid p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})\right] \tag{2}
$$

where $\mathcal{F}$ denotes a family of distributions.

Parameter estimation of LDA III

- Proposed variational EM algorithm.
  1. (E step) Compute the lower bound of ℓ by computing q that solves (2)
  2. (M step) Maximize the (1) w.r.t $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ given q.
- Repeat Step 1 and 2 until convergence.

Examples: Document Modelling I

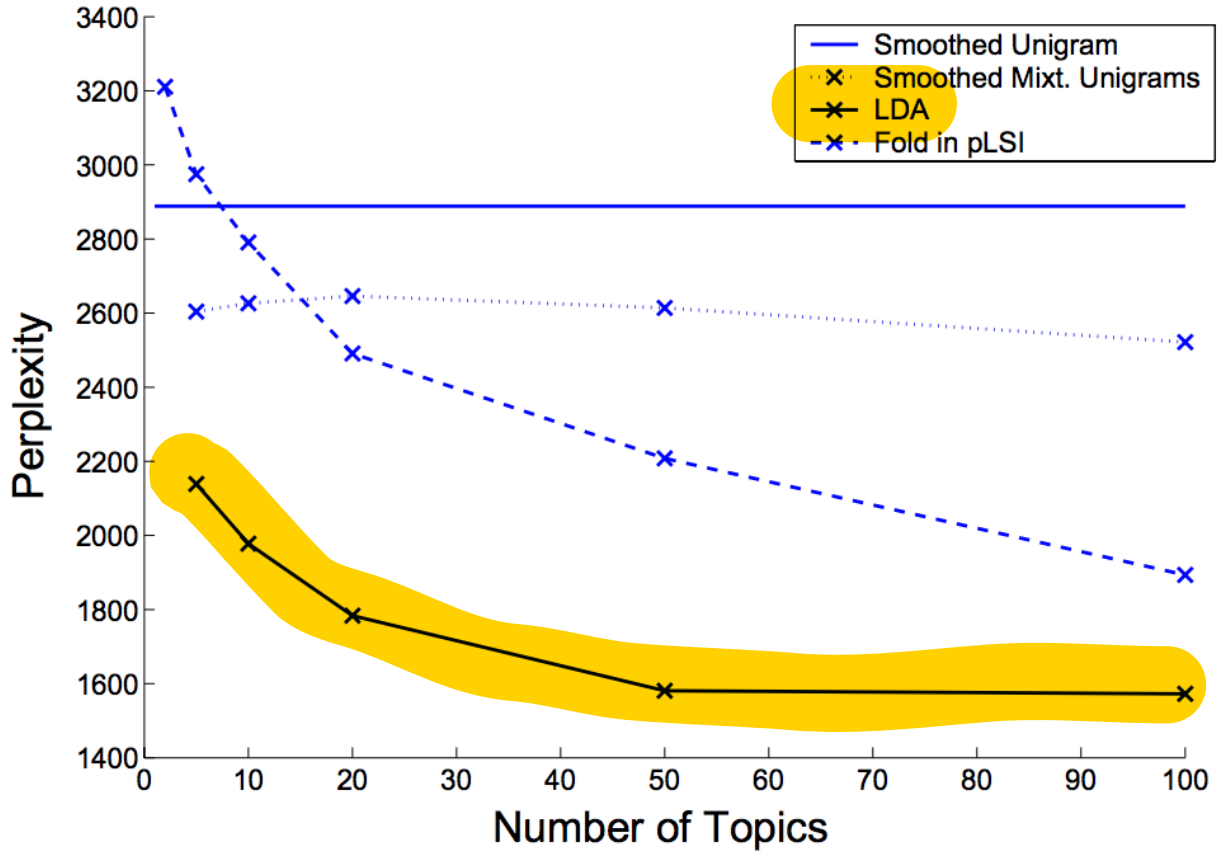
- Essentially density estimation:

$$P(\mathcal{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$$

- Performance measure

$$perplexity(\mathcal{D}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(\mathbf{W}_d)}{\sum_{d=1} N_d} \right\}$$

Examples: Document Modelling II



Examples: Document Modelling III

- LDA with 100 topics is applied to 16,000 documents from TREC AP corpus.

"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

Figure: The top words from some of the resulting multinomial distributions $p(\mathbf{w}_n|\mathbf{z}_n)$. These distributions seem to capture some of the underlying topics in the corpus.

Examples: Document Classification

- $\boldsymbol{\theta}_d$ can be regarded as a feature.
- Classification model based on the feature can be applied.