Latent Dirichlet Allocation (LDA)[1]

Hedibert F. Lopes
Insper Institute of Education and Research
São Paulo, Brazil

[1]Slides based on Blei, Ng and Jordan's paper "Latent Dirichlet Allocation" that appeared in 2003 the *Journal of Machine Learning Research*, Volume 3, pages 993-1022.

Paper's abstract

LDA: generative probabilistic model for collections of discrete data (text corpora).

LDA: 3-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics.

Topics: modeled as infinite mixtures over underlying sets of topic probabilities.

In the context of text modeling, the topic probabilities provide an explicit representation of a document.

Notation and terminology

A word is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \ldots, V\}$.

A document is a sequence of N words denoted by $\omega = (w_1, w_2, \ldots, w_n)$, where w_n is the nth word in the sequence

A corpus is a collection of M documents denoted by $D = \{\omega_1, \ldots, \omega_N\}$

Latent Dirichlet allocation

LDA is a generative probabilistic model of a corpus.

Documents are represented as random mixtures over latent topics.

LDA assumes the following generative process for document ω in a corpus D:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
3. For each of the N words w_n:
   3.1 Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   3.2 Choose a word w_n from $p(w_n|z_n, \beta)$.

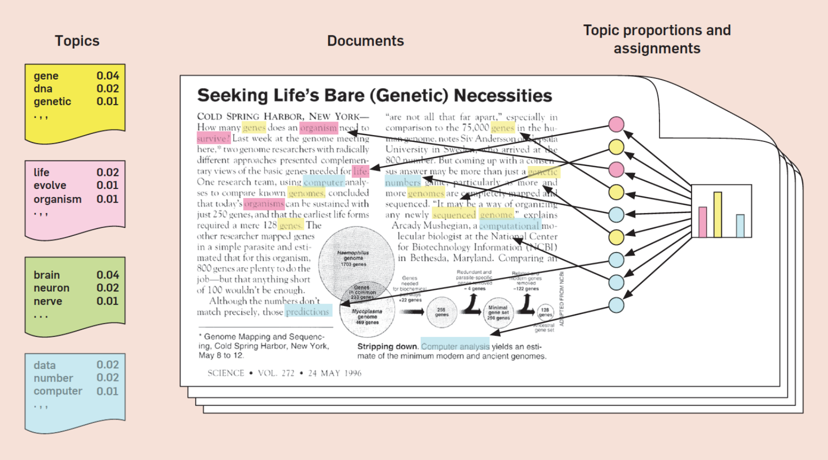
Simplifying assumptions:

- The dimensionality k of the Dirichlet distribution is known and fixed.
- The word probabilities are parameterized by β:

$$\beta_{ij} = Pr(w^j = 1 | z^i = 1)$$

Probabilistic topic models[2]

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

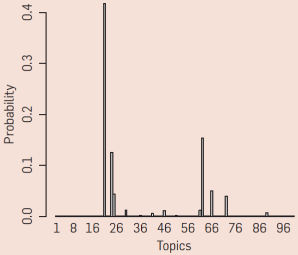


---

[2] David M. Blei (2012) Probabilistic topic models. *Communications of the Association for Computing Machinery (ACM)*, 55(4), 77-84.

Probabilistic topic models

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



“Genetics”	**“Evolution”**	**“Disease”**	**“Computers”**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Likelihood

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

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

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

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

Marginal distribution of a document

Integrating over (θ, z), we obtain the marginal distribution of a document:

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

The probability of a corpus is then:

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

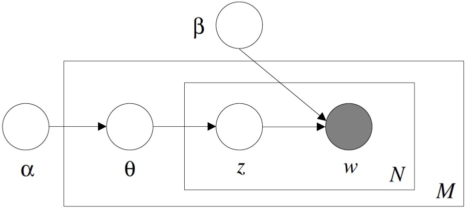


Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Three levels in the LDA representation

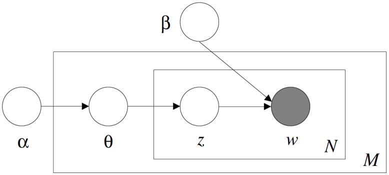


Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Corpus-level:
The parameters α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus.

Document-level:
The variables θ_d are document-level variables, sampled once per document.

Word-level:
Finally, the variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document.

Other latent variable models

Unigram model:

The words of every document are drawn independently from a single multinomial distribution:

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

Mixture of unigrams:

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

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

Probabilistic latent semantic indexing (pLSI):

Attempts to relax the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic.

$$p(d, w_n) = p(d) \sum_{z} p(w_n|z) p(z|d).$$

Topic models

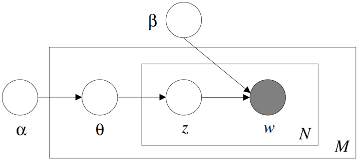
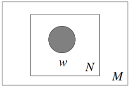
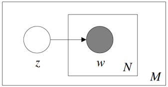


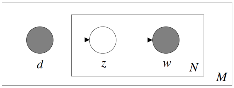
Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.



(a) unigram



(b) mixture of unigrams



(c) pLSI/aspect model

Figure 3: Graphical model representation of different models of discrete data.

Inference

The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given a document:

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

with

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

a function which is intractable due to the coupling between θ and β in the summation over latent topics.

Although the posterior distribution is intractable for exact inference, a wide variety of approximate inference algorithms can be considered for LDA, including Laplace approximation, variational approximation, and MCMC (Jordan, 1999)[3].

[3] Michael Jordan, editor. Learning in Graphical Models. MIT Press, Cambridge, MA, 1999.

Example

16,000 documents from a subset of the TREC AP corpus (Harman, 1992)[4]. They fit a 100-topic LDA model. The top words from some of the resulting multinomial distributions $p(w|z)$ are illustrated in Figure 8 (top).

"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.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Bottom Fig 8: document from TREC AP corpus not used for parameter estimation.

[4]Harman (1992) Overview of the first text retrieval conference (TREC-1). In *Proceedings of the First Text Retrieval Conference (TREC-1)*, pages 1-20.

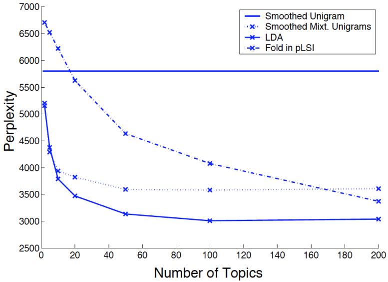
Perplexity

The perplexity is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. More formally, for a test set of M documents, the perplexity is:

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

A lower perplexity score indicates better generalization performance.

TREC AP corpus with 16,333 newswire articles with 23,075 unique terms.
90% for training and 10% for testing.



References (chronological order)

1. Blei, Ng and Jordan (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3, 993-1022.
   http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

2. Griffiths and Steyvers (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 101, Suppl. 1, 5228-5235.
   https://doi.org/10.1073/pnas.0307752101

3. Blei and Lafferty (2006) Correlated topic models. *Advances in Neural Information Processing Systems (NIPS)*.
   http://people.ee.duke.edu/∼lcarin/Blei2005CTM.pdf

4. Blei and Lafferty (2006) Dynamic topic models. In *International Conference on Machine Learning (ICML)*, 2006.
   https://mimno.infosci.cornell.edu/info6150/readings/dynamic_topic_models.pdf

5. Blei and Lafferty (2007) A correlated topic model of science. *The Annals of Applied Statistics (AOAS)*, Vol. 1, No. 1, 17-35.
   https://projecteuclid.org/download/pdfview_1/euclid.aoas/1183143727

6. Blei and McAuliffe (2007) Supervised topic models. In *Proceedings of the Neural Information Processing Systems (NIPS)*, Vol. 21, 1-8.
   https://arxiv.org/pdf/1003.0783.pdf

7. Steyvers and Griffiths (2007) Probabilistic topic models.
   http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf

8. Blei (2012) Probabilistic topic models. *Communications of the Association for Computing Machinery (AMC)*, Vol. 55, No. 4.
   http://www.cs.columbia.edu/∼blei/papers/Blei2012.pdf

9. Taddy (2012) On estimation and selection for topic models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.
   http://proceedings.mlr.press/v22/taddy12/taddy12.pdf

10. Alghamdi and Alfalqi (2015) A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 6, No. 1, 147-153.
    https://thesai.org/Downloads/Volume6No1/Paper_21-A_Survey_of_Topic_Modeling_in_Text_Mining.pdf

11. Airoldi and Bischof (2016) Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association (JASA)*, 111:516, 1381-1403.
    https://doi.org/10.1080/01621459.2015.1051182