Latent Dirichlet Allocation: Formulation

Pawan Goyal

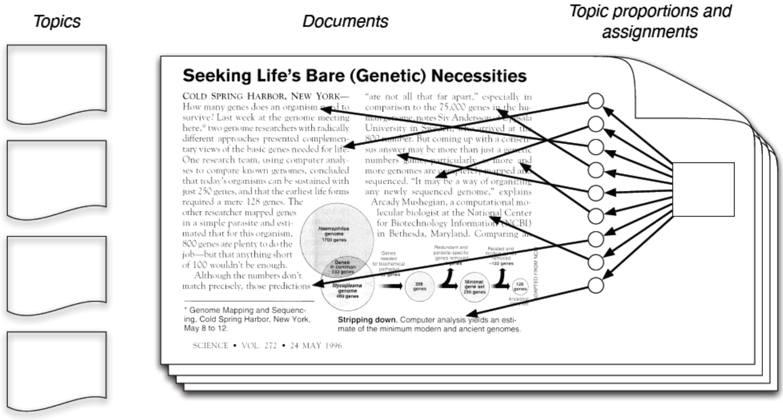
CSE, IIT Kharagpur

Week 9, Lecture 2

Central Problem of LDA

- The documents themselves are observed, while the topic structure - the topics, per-document topic distributions, and the per-document per-word topic assignments - is *hidden structure*.
- The central computational problem is to use the observed documents to infer the hidden topic structure, i.e. *reversing* the generative process.

Goal: The posterior distribution



Infer the hidden variables

Compute their distribution conditioned on the documents

Topics from TASA corpus

37,000 text passages from educational materials (300 topics)

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Topics from TASA corpus

Documents with different content can be generated by choosing different distributions over topics.

- Equal probability to first two topics:

Topics from TASA corpus

Documents with different content can be generated by choosing different distributions over topics.

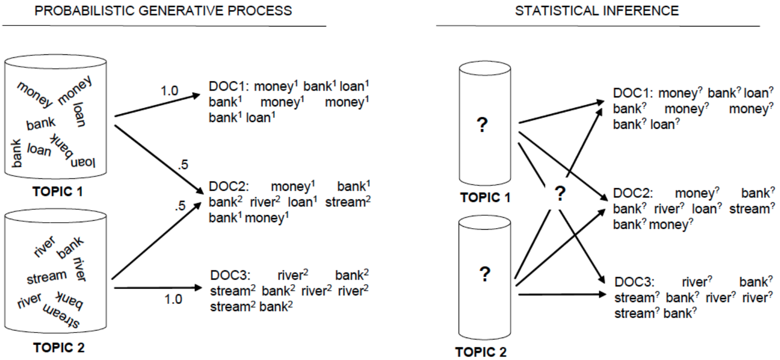
- Equal probability to first two topics: about a person who has taken too many drugs and how that affected color perceptions.
- Equal probability to the last two topics:

Topics from TASA corpus

Documents with different content can be generated by choosing different distributions over topics.

- Equal probability to first two topics: about a person who has taken too many drugs and how that affected color perceptions.
- Equal probability to the last two topics: about a person who experienced a loss of memory, which required a visit to the doctor.

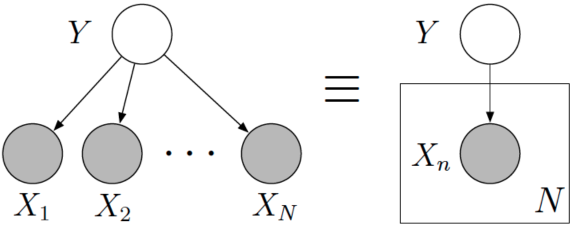
Generative model and statistical inference



Important points

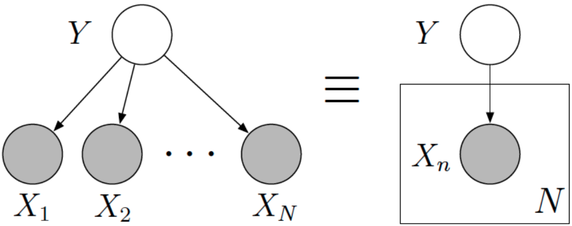
- *bag-of-words assumption:* The generative process does not make any assumptions about the order of words in the documents.
- *capturing polysemy:* The way that the model is defined, there is no notion of mutual exclusivity that restricts words to be part of one topic only. Ex: both ‘money’ and ‘river’ topics can give high probability to the word ‘bank’.

Graphical Model (Notation)



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

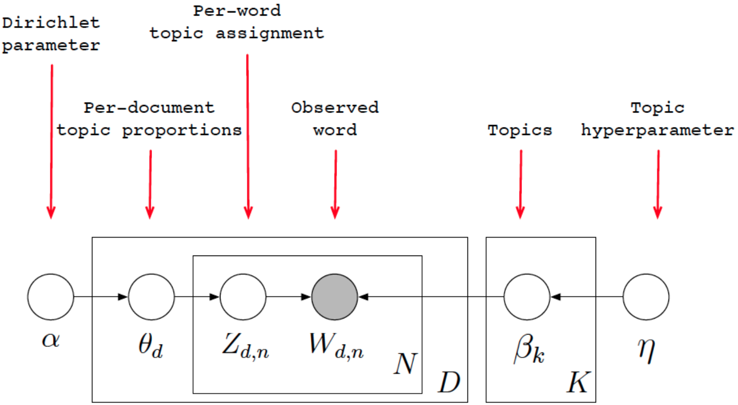
Graphical Model (Notation)



- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

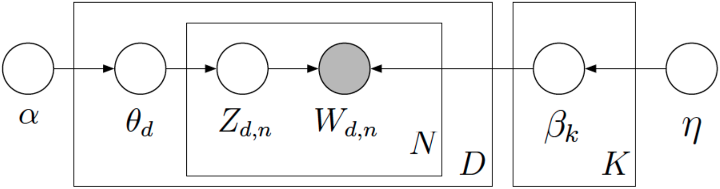
$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n | y)$$

LDA: Graphical Model



Each piece of the structure is a random variable.

Latent Dirichlet Allocation: Generative Model



1. Draw each topic $\beta_i \sim \mathrm{Dir}(\eta)$, for $i \in \{1, \ldots, K\}$.
2. For each document:
    1. Draw topic proportions $\theta_d \sim \mathrm{Dir}(\alpha)$.
    2. For each word:
        1. Draw $Z_{d,n} \sim \mathrm{Mult}(\theta_d)$.
        2. Draw $W_{d,n} \sim \mathrm{Mult}(\beta_{Z_{d,n}})$.

What is Latent Dirichlet Allocation (LDA)?

- ‘Latent’ has the same sense in LDA as in Latent semantic indexing, i.e. capturing topics as latent variables
- The distribution that is used to draw the per-document topic distributions is called a *Dirichlet distribution*. This result is used to allocate the words of the documents to different topics.

What is Latent Dirichlet Allocation (LDA)?

- 'Latent' has the same sense in LDA as in Latent semantic indexing, i.e. capturing topics as latent variables
- The distribution that is used to draw the per-document topic distributions is called a *Dirichlet distribution*. This result is used to allocate the words of the documents to different topics.

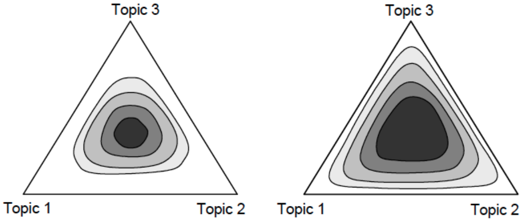
Dirichlet Distribution

The Dirichlet distribution is an exponential family distribution over the simplex, i.e. positive vectors that sum to one

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

Dirichlet Distribution

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

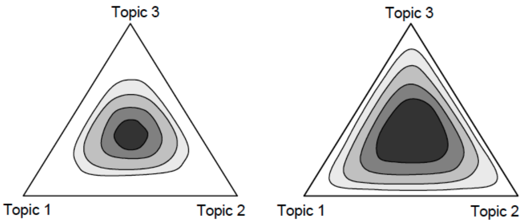


α_is: **hyper-parameters of the model:**

α_j can be interpreted as a prior observation count for the number of times topic j is sampled in a document

Dirichlet Distribution

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

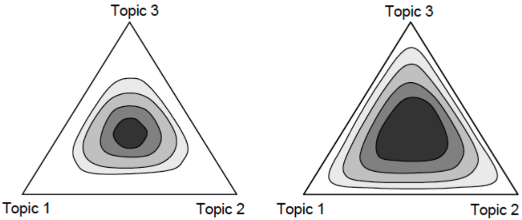


α_is: **hyper-parameters of the model:**

These priors can be interpreted as forces in the topic distributions with higher α moving the topics away from the corners of the simplex

Dirichlet Distribution

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

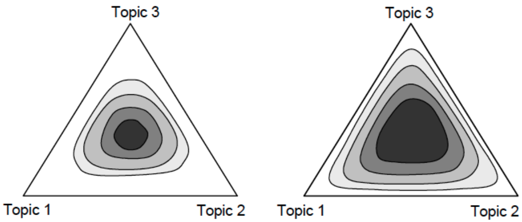


α_is: **hyper-parameters of the model:**

When $\alpha < 1$, there is a bias to pick topic distributions favoring just a few topics

Dirichlet Distribution

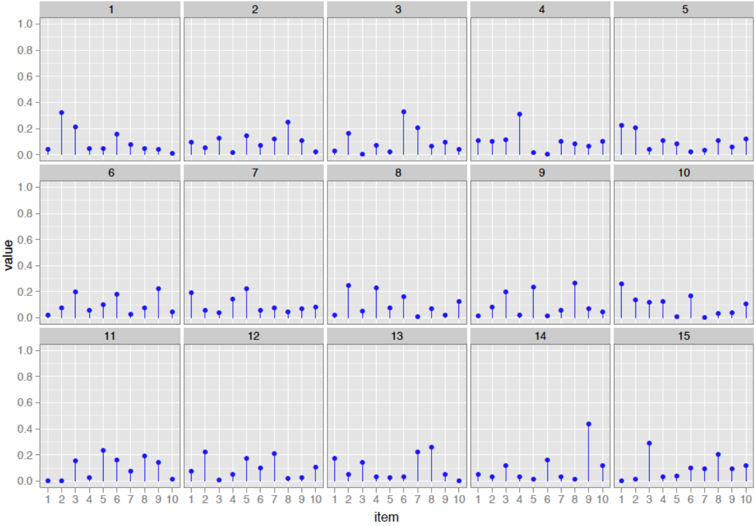
$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$



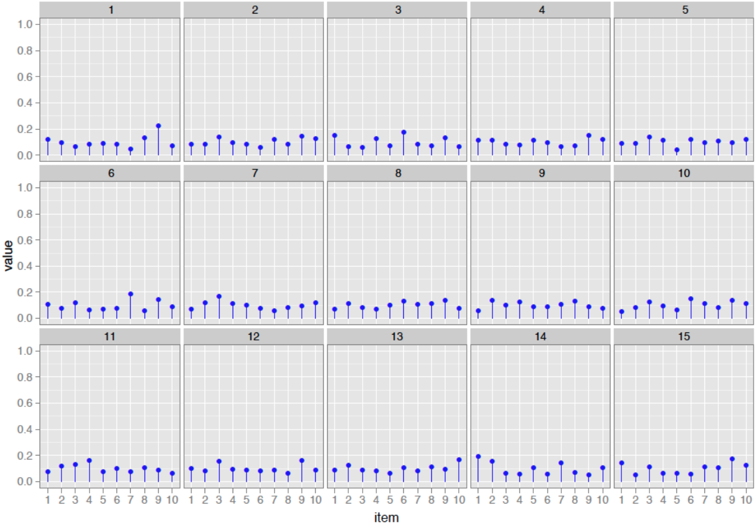
α_is: **hyper-parameters of the model:**

It is convenient to use a symmetric Dirichlet distribution with a single hyper-parameter $\alpha_1 = \alpha_2 \ldots = \alpha$

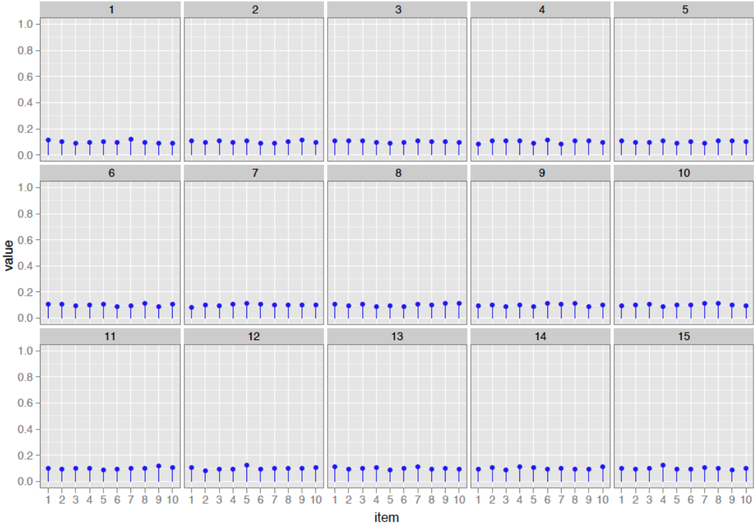
Effect of α: $\alpha = 1$



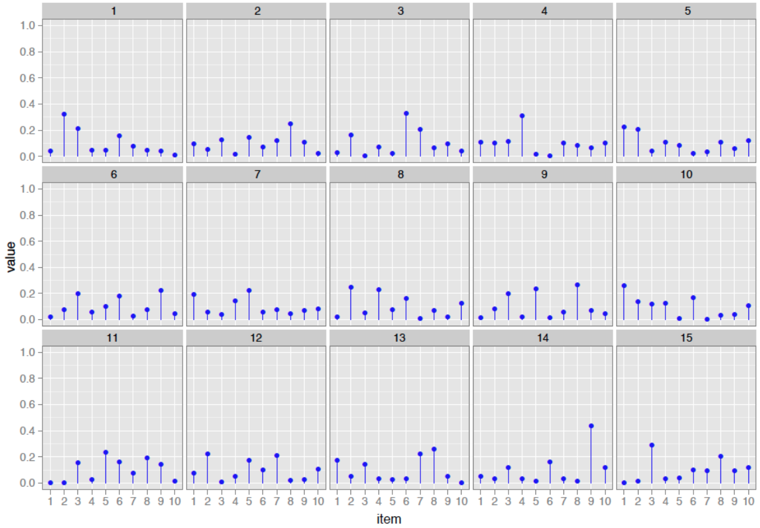
Effect of α: $\alpha = 10$



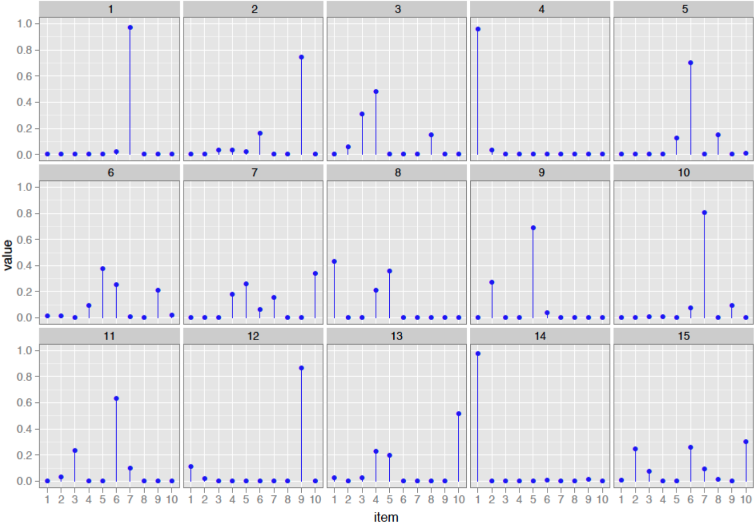
Effect of α: α = 100



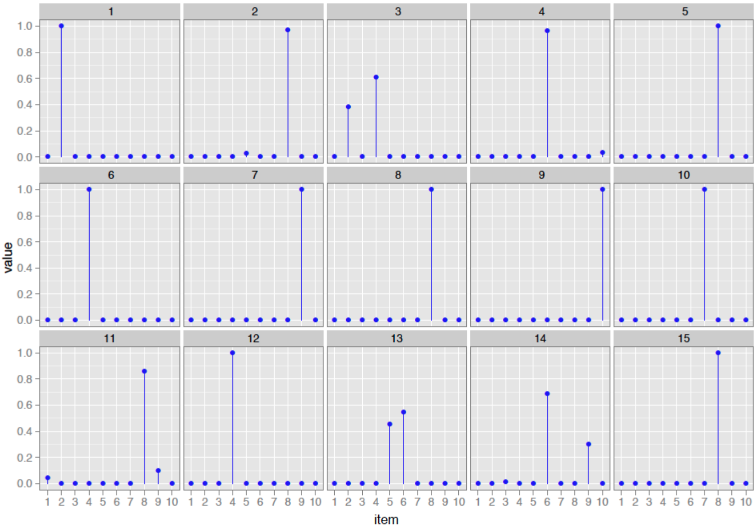
Effect of α: $\alpha = 1$



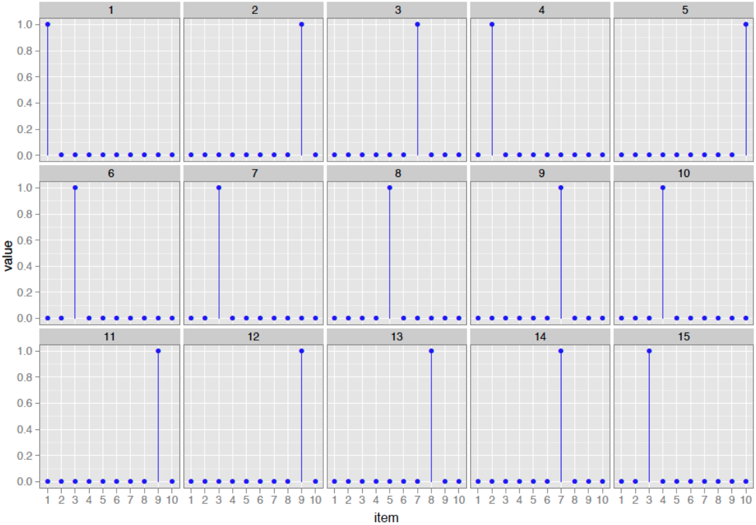
Effect of α: $\alpha = 0.1$



Effect of α: $\alpha = 0.01$



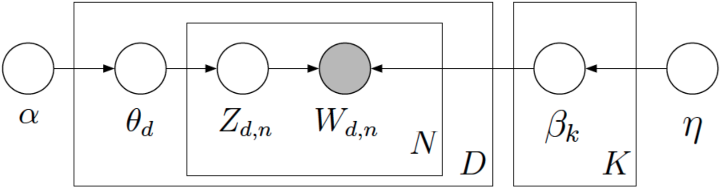
Effect of α: $\alpha = 0.001$



Online Implementations

LDA-C*	A C implementation of LDA
HDP*	A C implementation of the HDP (“infinite LDA”)
Online LDA*	A python package for LDA on massive data
LDA in R*	Package in R for many topic models
LingPipe	Java toolkit for NLP and computational linguistics
Mallet	Java toolkit for statistical NLP
TMVE*	A python package to build browsers from topic models

Latent Dirichlet Allocation: Statistical Inference



- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, etc.