Topic Models and Latent Dirichlet Allocation

Kevin Yang and Zachary Wu

27 April 2017

Motivation

We want to find short descriptions of the members of a collection that

- enable efficient processing of large collections
- preserve essential statistical relationships
- that can be used for classification, novelty detection, summarization, and similarity and relevance judgments

Words, documents, and corpora

- Words are discrete units of data and belong to a vocabulary of size V.
- Words are encoded by one-hot vectors of length V.
- A document is a sequence of N words denoted $\mathbf{w} = (w_1, w_2, ...w_N)$
- A corpus is a collection of M documents denoted $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$

Latent Dirichlet Allocation

- Generative probabilistic model
- The K topics $\{z_1, z_2, ...z_K\}$ are distributions over words
- Topics specified by $\beta \in M^{KxV}$ where $\beta_{i,j} = p(w^j = 1 | z^j = 1)$
- Documents are random *mixtures* of the latent topics

Generating a document:

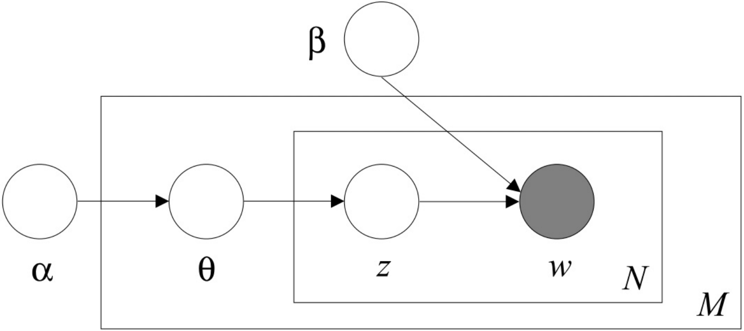
1. Sample a set of mixing probabilities $\theta \sim Dir(\alpha)$, $\alpha, \theta \in \mathcal{R}^K$
2. For each of the N words:
   2.1 Choose a topic $z_n \sim multinomial(\theta)$
   2.2 Choose a word $w_n \sim multinomial(\beta_n)$

Dirichlet distribution

$$
\begin{aligned}
f(\theta|\alpha) &= \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \\
B(\alpha) &= \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}
\end{aligned}
\tag{1}
$$

- Conjugate prior to multinomial and categorical distributions
- K-dimensional Dirichlet random variable takes values in the $(K-1)$-simplex
- Parametrized by K-dimensional vector α

Latent Dirichlet Allocation



Blei et al. 2003

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

Latent Dirichlet Allocation

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

Marginal distribution of a document:

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

Probability of a corpus:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d \tag{4}$$

Exchangeability

A finite set of random variable $\{z_1, ..., z_N\}$ is *exchangeable* if the joint distribution is invariant to permutation. That is, for a permutation π of the integers from 1 to N:

$$p(z_1, ..., z_N) = p(z_{\pi(1)}, ..., z_{\pi(N)})$$

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

Exchangeability

- In LDA, words are generated by topics, and topics within a document are infinitely exchangeable.
- De Finetti's representation theorem: exchangeable observations are conditionally independent given some latent variable.
- In LDA, the words are i.i.d. conditioned on the topics, and the topics are i.i.d. conditioned on α

Therefore:

$$p(\mathbf{w}, \mathbf{z}) = \int (p(\theta)) \left(\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \tag{5}$$

LDA as a continuous mixture of unigrams

Within a document, the words are distributed as:

$$p(w|\theta, \beta) = \sum_{z} p(w|z, \beta)p(z|\theta)$$

The document distribution is then a continuous mixture distribution:

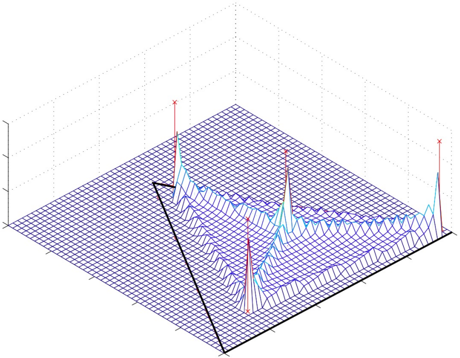
$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} p(w_n|\theta, \beta) \right) d\theta$$

where $p(w_n|\theta, \beta)$ are the mixture components and $p(\theta|\alpha)$ are the mixture weights.

Example unigram distribution

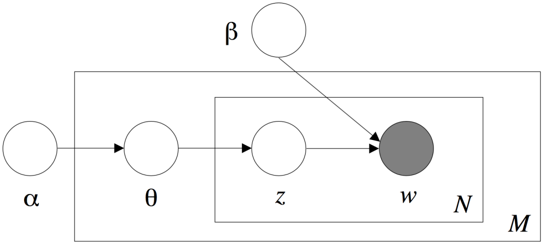
$p(w|\theta, \beta)$ under LDA for 3 words and 4 topics

- Triangle is all possible distributions over 3 words
- 4 topics $p(w|z)$ are marked with red lines
- Surface is an example mixture density given by LDA

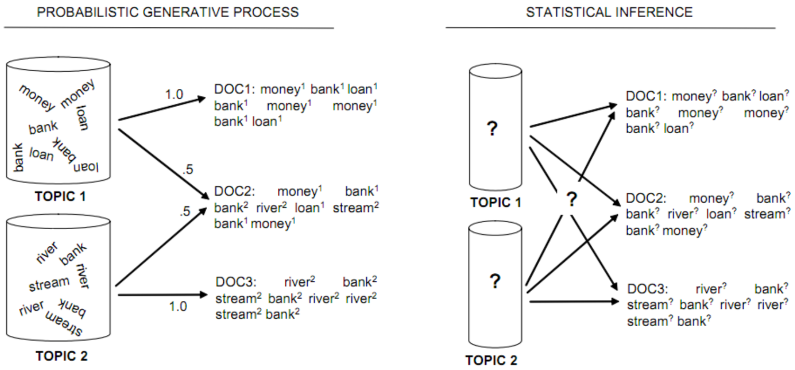


Blei et al. 2003

Model Review



Inference



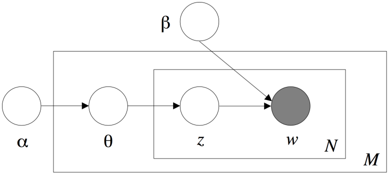
Steyvers and Griffiths, 2007

Inference

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{6}$$

Given a document **w**, compute the posterior distribution of the hidden variables θ (topics distribution) and **z** (random topic variable)

Inference - Numerator



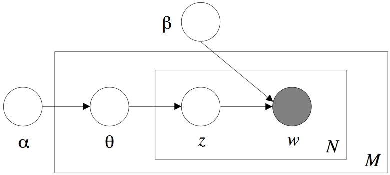
The numerator of the posterior can be decomposed by the hierarchy shown in the graphical model

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\mathbf{w}|\mathbf{z}, \beta)p(\mathbf{z}|\theta)p(\theta|\alpha) \tag{7}$$

The first term represents the probability of observing a document **w** given a topic vector that assigns a topic to each word from the probability matrix β.

$$p(\mathbf{w}|\mathbf{z}, \beta) = \prod_{n=1}^{N} \beta_{z_n, w_n}\theta_{z_n} \tag{8}$$

Inference - Numerator



$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\mathbf{w}|\mathbf{z}, \beta)p(\mathbf{z}|\theta)p(\theta|\alpha) \tag{9}$$

The second term follows directly from the definition of θ_i as $p(z_n|\theta) = \theta_i$ such that $z_n^i = 1$. The third term is the Dirichlet distribution described earlier.

Inference - Numerator and Denominator

The posterior distribution is then

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = \left(\frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \prod_{n=1}^{N} \prod_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{i,j})^{w_n^j z_n^j} \tag{10}$$

where V is the size of the entire vocabulary.
We can then find the denominator by marginalizing over θ and $\mathbf{z}$.

$$p(\mathbf{w}|\alpha, \beta) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \int \left(\theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{i,j})^{w_n^j} \right) d\theta \tag{11}$$

Problem: Intractable!

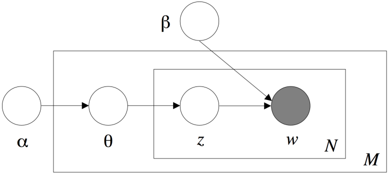
$$p(\mathbf{w}|\alpha,\beta) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\int\left(\theta_i^{\alpha_i-1}\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\beta_{i,j})^{w_n^j}\right)d\theta \tag{12}$$

Inherent problem is coupling of theta and beta.

Problem: Intractable!

$$p(\mathbf{w}|\alpha,\beta) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)} \prod_{i=1}^{k} \int \left(\theta_i^{\alpha_i-1}\right) \left(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\beta_{i,j})^{w_n^j}\right) d\theta \tag{12}$$

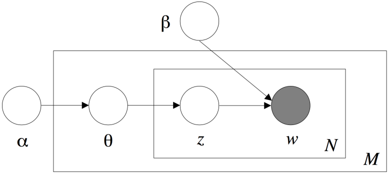
Inherent problem is coupling of theta and beta.

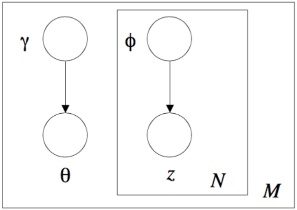


Unfortunately, this coupling cannot be removed...

Solution: Variational Inference

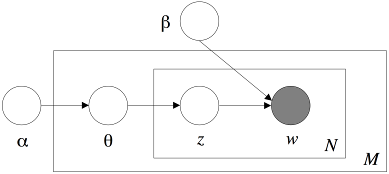
As usual, modify the original graphical model by removing the difficult edges and nodes

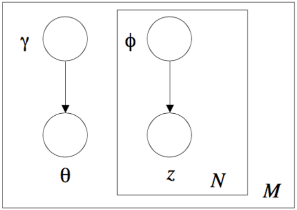




Solution: Variational Inference

As usual, modify the original graphical model by removing the difficult edges and nodes





Variational Distribution

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{13}$$

Variational Inference

Using Jensen's inequality to get an adjustable lower bound on the log likelihood, the following optimization problem is obtained

$$(\gamma^*, \phi^*) = \underset{(\gamma,\phi)}{\arg\min}\, \mathrm{D}(q(\theta, \mathbf{z}|\gamma, \phi) \,||\, p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)) \tag{14}$$

The optimizing values of the variational parameters are found by minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior.

Variational Inference - Optimization

Computing the derivatives of the KL divergence and setting them equal to zero gives the following update equations

$$\phi_{ni} \propto \beta_{iw_n} \exp\left(\mathsf{E}_q[\log(\theta_i)|\gamma] \right) \tag{15}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{16}$$

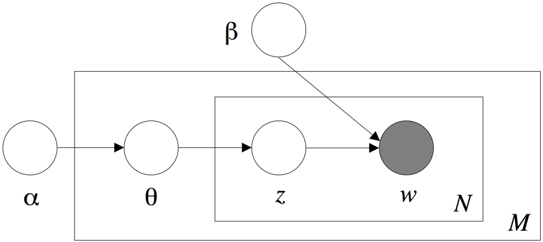
Note that the variational distribution is still conditional on **w** as they are optimized for fixed **w**!

Variational Inference Algorithm

Algorithm 1 Variational Inference Algorithm for LDA

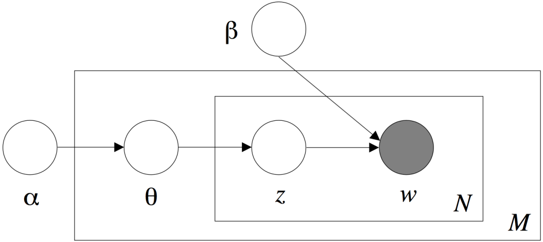
procedure INFERENCE
    initialize $\phi_{ni}^{0} := 1$ for all i and n
    initialize $\gamma_{ni}^{0} := \alpha_i + N/k$ for all i
    repeat
        for $n = 1$ **to** N
            for $i = 1$ **to** k
                $\phi_{ni}^{t+1} := \beta_{iw_n} \exp \Psi(\gamma_i^t)$
            normalize ϕ_n^{t+1} to sum to 1
        $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
    until convergence

Inference: Review



$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{17}$$

Inference: Review



$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{17}$$

Not done yet!

Inference - α and β

Goal: Find parameters α and β that maximize the (marginal) log likelihood of the data:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d | \alpha, \beta) \tag{18}$$

Inference - α and β

Goal: Find parameters α and β that maximize the (marginal) log likelihood of the data:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d | \alpha, \beta) \tag{18}$$

Approach: Use alternating variational EM to maximize the lower bound with respect to variational parameters γ and ϕ, then maximize with respect to model parameters α and β.

Inference - Variational EM

1. E-Step: Find optimizing values of variational parameters $[\gamma^*, \phi^*]$ as shown previously

2. Maximize the resulting lower bound on log likelihood with respect to model parameters $[\alpha^*, \beta^*]$

Inference - M-Step

Update for conditional multinomial parameter β has the following closed-form solution

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N} \phi^*_{dni} w^j_{dn} \tag{19}$$

Update for Dirichlet parameter α is not as straightforward (implemented with Newton-Raphson method to determine the optimal α)

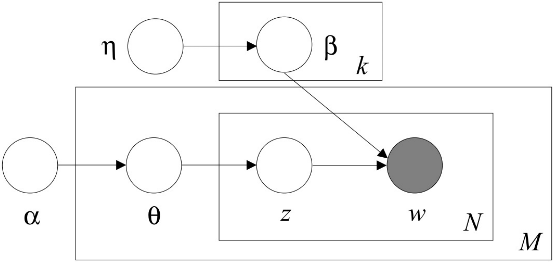
Inference - M-Step

Algorithm 2 M-Step for Determining α and β

procedure M-STEP
    for $d = 1$**to M**
        for $i = 1$**to K**
            for $j = 1$**to V**
                $\beta_{ij} := \phi_{dni} w_{dnj}$
            normalize β_i to sum to 1
    estimate α

Smoothing

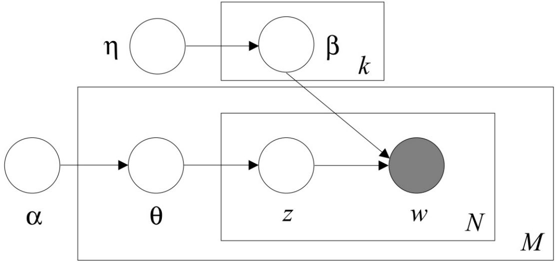
Problem: when a new document has terms not seen in training, LDA assigns the new document 0 probability.
Solution: Draw each row of β from an exchangeable Dirichlet distribution ($\alpha_i = \eta \ \forall \ i \in \{1, ...V\}$)



Blei et al. 2003

Smoothing

This requires adding an aditional variational parameter λ.



Blei et al. 2003

$$q(\beta_{1:K}, \mathbf{z}_{1:M}, \theta_{1:M} | \lambda, \phi, \gamma) = \prod_{i=1}^{K} Dir(\beta_i, \lambda_i) \prod_{d=1}^{M} q_d(\theta_d, \mathbf{z}_d | \phi_d, \gamma_c)$$

With update

$$\lambda_{ij} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}$$

LDA results example

Topics

Top 10 terms

algorithms	organic chemistry	neuroscience	solar system	HIV/AIDS
computer	chemistry	cortex	orbit	infection
methods	synthesis	stimulus	dust	immune
number	oxidation	fig	jupiter	aids
two	reaction	vision	line	infected
principle	product	neuron	system	viral
design	organic	recordings	solar	cells
access	conditions	visual	gas	vaccine
processing	cluster	stimuli	atmospheric	antibodies
advantage	molecule	recorded	mars	hiv
important	studies	motor	field	parasite

Blei et al. 2009

LDA results example

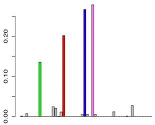
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions



Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

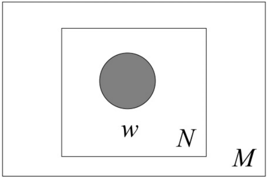
Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

Blei et al. 2009

Other latent variable models

Unigram model



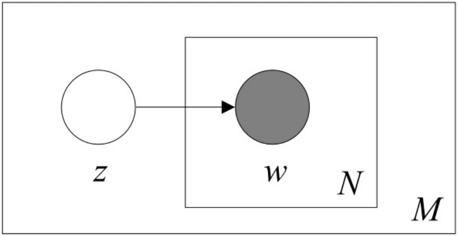
Blei et al. 2003

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

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

Other latent variable models

Mixture of unigrams



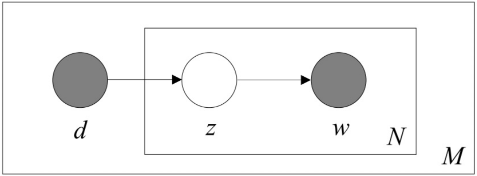
Blei et al. 2003

Each document has a single topic z that determines the multinomial from which the words are drawn.

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

Other latent variable models

Probabilistic latent semantic indexing



Blei et al. 2003

A document index d and a word w_n are conditionally independent given an unobserved topic z:

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

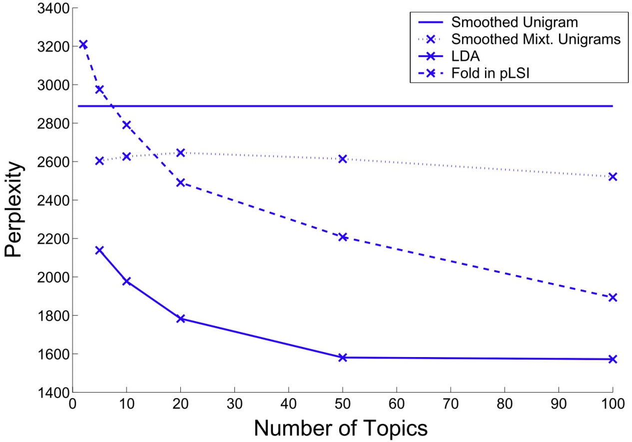
- Model only learns $p(z|d)$ for training set
- Number of parameters is linear with size of training set
- Prone to overfitting

Comparison of latent variable models

- Test generalization performance on unlabeled corpus.
- Train, then measure likelihood on held-out test set.
- Lower *perplexity* indicates better generalization

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

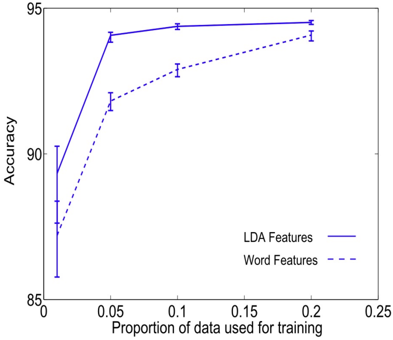
Comparison of latent variable models

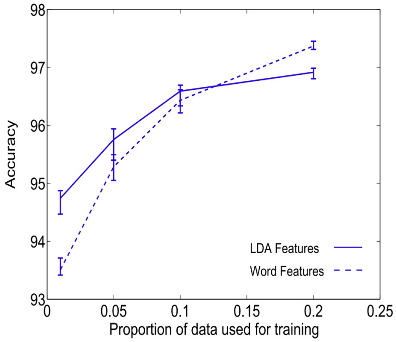


Blei et al. 2003

LDA for classification

1. Use LDA with 50 topics on all documents
2. Use posterior Dirichlet parameters $\gamma^*(\mathbf{w})$ as document features
3. Train Using SVM and compare to SVM trained on word features

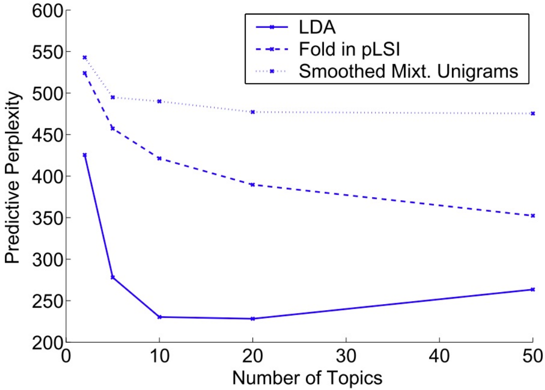




Blei et al. 2003

LDA for collaborative filtering

1. Users as documents; preferred movie choices as words.
2. Train LDA on fully observed set of users.
3. For each unobserved user, hold out one movie and evaluate probability assigned to that movie.



Blei et al. 2003

Extensions of LDA

- Correlated topic model (CTM): model correlations between occurrences of topics
- Dynamic topic model (DTM): remove exchangeability between documents to model evolution of corpus over time.

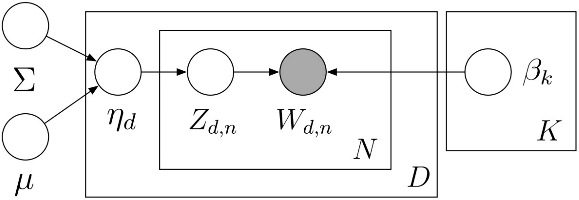
Correlated topic model

- Dirichlet assumption in LDA implies that components of θ are nearly independent.
- Solution: use logistic normal distribution

New generative process for documents:

1. Sample $\eta \sim \mathcal{N}(\mu, \Sigma)$
2. For each of the N words:
   2.1 Choose a topic $z_n \sim multinomial(f(\eta))$, $f(\eta_i) = \frac{e^{\eta_i}}{\sum_j e^{\eta_j}}$
   2.2 Choose a word $w_n \sim multinomial(\beta_n)$

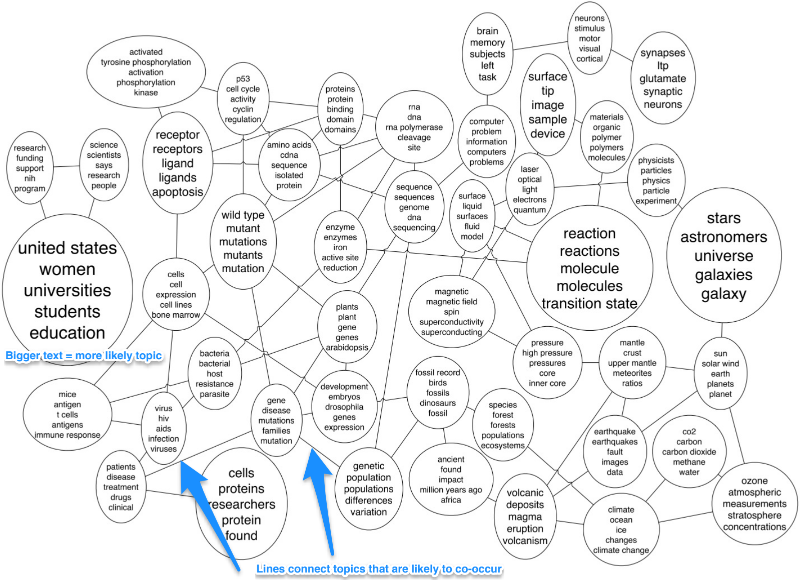
Correlated topic model



Blei et al. 2009

$$p(\eta, \mathbf{z}, \mathbf{w} | \mu, \Sigma, \beta) = p(\eta | \mu, \Sigma) \prod_{n=1}^{N} p(z_n | f(\eta)) p(w_n | z_n, \beta) \qquad (20)$$

Correlated topic model



Blei et al. 2009

Dynamic topic model

- Many corpora, such as journals, email, news, and search logs reflect evolving content.
- For example, a neuroscience article written in 1903 would be very different from one written in 2017.
- A DTM captures the evolution of topics in a sequentially-organized corpus.
- Documents are divided into time-slices (e.g. by year).
- The topics associated with slice t evolve from those associated with slice $t - 1$.

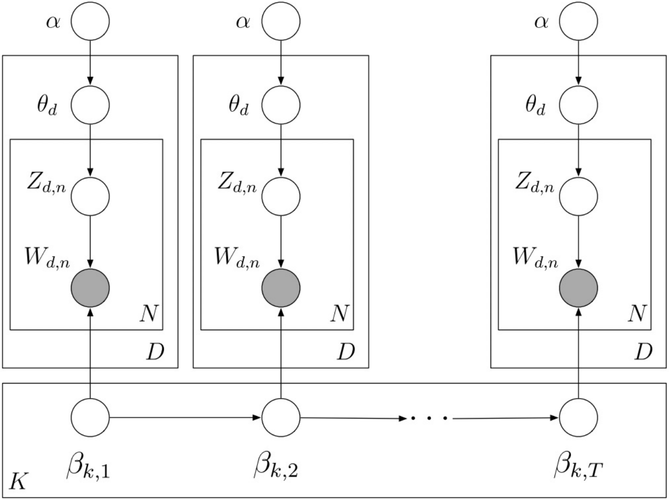
Dynamic topic model

New generative process for documents:

1. Sample topic distributions $\pi_t \sim \mathcal{N}(\pi_{t-1}, \sigma^2 I)$
2. For each document:
   2.1 Sample $\theta_d \sim Dir(\alpha)$
   2.2 For each word:
      2.2.1 Choose a topic $z_n \sim multinomial(\theta_d))$
      2.2.2 Draw $w_n \sim multinomial(f(\pi_{t,z}))$

$$p(\theta_d, \pi_t, \mathbf{z}, \mathbf{w} | \sigma, \alpha, \pi_{t-1}) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(\pi_t|\pi_{t-1}) p(w_n|z_n, f(\pi_t)) \tag{21}$$

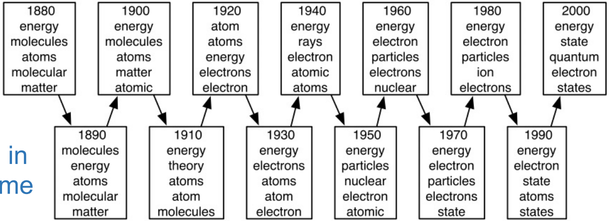
Dynamic topic model



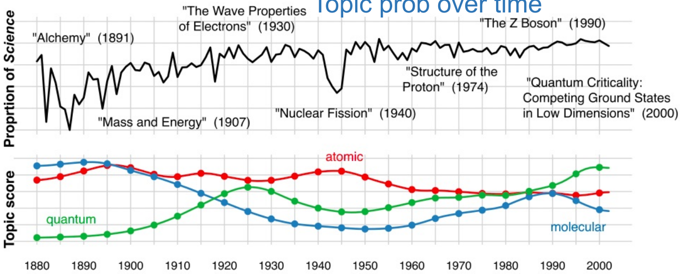
Blei et al. 2009

Dynamic topic model

Top 5 terms in topic over time



Topic prob over time



Term prob in topic over time

Blei et al. 2009

Dynamic topic model

Query	Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts (1994)
1	Global Text Matching for Information Retrieval (1991)
2	Automatic Text Analysis (1970)
3	Language-Independent Categorization of Text (1995)
4	Developments in Automatic Text Retrieval (1991)
5	Simple and Rapid Method for the Coding of Punched Cards (1962)
6	Data Processing by Optical Coincidence (1961)
7	Pattern-Analyzing Memory (1976)
8	The Storing of Pamphlets (1899)
9	A Punched-Card Technique for Computing Means (1946)
10	Database Systems (1982)

Blei et al. 2009

Dynamic topic model

Query	Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts (1994)
1	Global Text Matching for Information Retrieval (1991)
2	Automatic Text Analysis (1970)
3	Language-Independent Categorization of Text (1995)
4	Developments in Automatic Text Retrieval (1991)
5	Simple and Rapid Method for the Coding of Punched Cards (1962)
6	Data Processing by Optical Coincidence (1961)
7	Pattern-Analyzing Memory (1976)
8	The Storing of Pamphlets (1899)
9	A Punched-Card Technique for Computing Means (1946)
10	Database Systems (1982)

Blei et al. 2009

Most documents are relatively recent papers about organizing documents on computer systems.

Dynamic topic model

Query	Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts (1994)
1	Global Text Matching for Information Retrieval (1991)
2	Automatic Text Analysis (1970)
3	Language-Independent Categorization of Text (1995)
4	Developments in Automatic Text Retrieval (1991)
5	Simple and Rapid Method for the Coding of Punched Cards (1962)
6	Data Processing by Optical Coincidence (1961)
7	Pattern-Analyzing Memory (1976)
8	The Storing of Pamphlets (1899)
9	A Punched-Card Technique for Computing Means (1946)
10	Database Systems (1982)

Blei et al. 2009

Some older documents are about organizing information on punch cards.

Dynamic topic model

Query	Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts (1994)
1	Global Text Matching for Information Retrieval (1991)
2	Automatic Text Analysis (1970)
3	Language-Independent Categorization of Text (1995)
4	Developments in Automatic Text Retrieval (1991)
5	Simple and Rapid Method for the Coding of Punched Cards (1962)
6	Data Processing by Optical Coincidence (1961)
7	Pattern-Analyzing Memory (1976)
8	The Storing of Pamphlets (1899)
9	A Punched-Card Technique for Computing Means (1946)
10	Database Systems (1982)

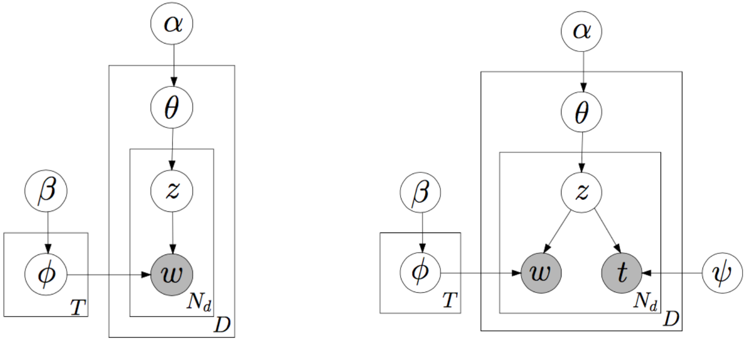
Blei et al. 2009

These authors thought *they* had too much information to easily store!

Topics over Time (TOT)

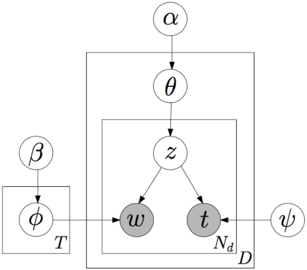
Goal: Continuous distribution over timestamps of topics

- Meaning of a topic is constant
- Topic occurrence and correlations change over time



Wang and McCallum 2005

TOT Strategy: Gibbs Sampling

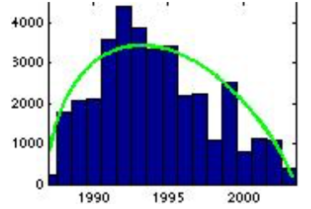


1. Draw T multinomials ϕ_z from a Dirichlet prior β, one for each topic z
2. For each document d, draw a multinomial θ_d from a Dirichlet prior α; then for each word w_{di} in doc d:
   - Draw a topic z_{di} from θ_d
   - Draw a word w_{di} from $\phi_{z_{di}}$
   - Draw a timestamp t_{di} from $\psi_{z_{di}}$

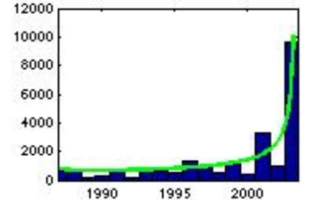
$$P(z_{di}|\mathbf{w}, \mathbf{t}, \mathbf{z}_{-di}, \alpha, \beta, \Psi) \propto (m_{dz_{di}} + \alpha_{z_{di}} - 1)$$
$$\times \quad \frac{n_{z_{di}w_{di}} + \beta_{w_{di}} - 1}{\sum_{v=1}^{V}(n_{z_{di}v} + \beta_v) - 1} \frac{(1 - t_{di})^{\psi_{z_{di}1} - 1} t_{di}^{\psi_{z_{di}2} - 1}}{B(\psi_{z_{di}1}, \psi_{z_{di}2})},$$

TOT Results

Recurrent NN



Game Theory



Recurrent NN		Game Theory	
state	0.05963	game	0.02850
recurrent	0.03765	strategy	0.02378
sequence	0.03616	play	0.01490
sequences	0.02462	games	0.01473
time	0.02402	player	0.01451
states	0.02057	agents	0.01346
transition	0.01300	expert	0.01281
finite	0.01242	strategies	0.01123
length	0.01154	opponent...	0.01088
strings	0.01013	nash	0.00848

Wang and McCallum 2005