IIIT-H

Web Mining
Lecture 6: Topic Models

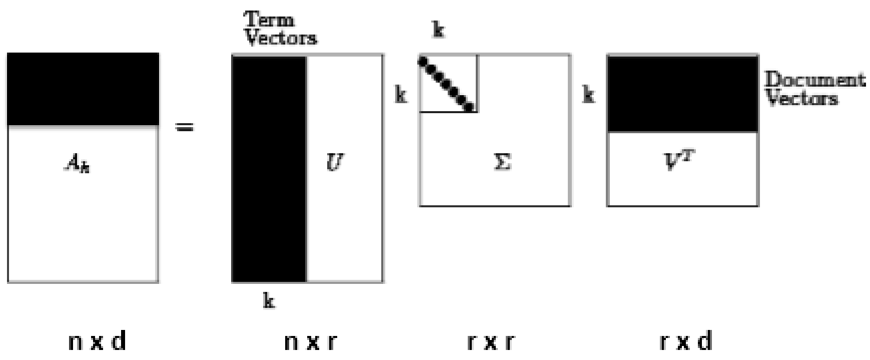
Manish Gupta

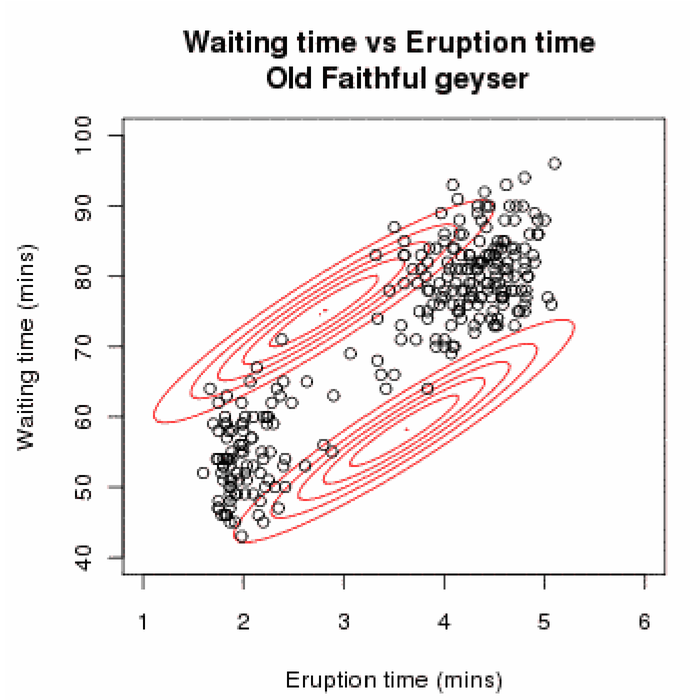
17th Aug 2013

Slides borrowed (and modified) from
http://knight.cis.temple.edu/~yates/cis8538/sp11/slides/intro-to-lsa-lda.ppt
http://home.etf.rs/~vm/tutorial/Coimbra/**slides**/3_Jelisavcic.pptx
http://mallet.cs.umass.edu/mallet-tutorial.pdf
http://www.cse.ust.hk/~lzhang/teach/6931a/slides/lda-zhou.pdf

Recap of Lecture 5: LSI and EM

- Singular Value Decomposition (SVD)
- Latent Semantic Indexing (LSI)
- K-Means
- Expectation Maximization (EM)





Announcements

- Assignment 1 submission date is Aug 22 9pm
- Rescheduling of lectures
  - Makeup class for Aug 24 lecture will be on Aug 22 6-7:30pm
  - Makeup class for Aug 28 lecture will be on Sep 2 6-7:30pm
- Tutorial 1
  - 17^{th} Aug 4:30pm at 103 Himalaya
  - Doubt clarification of Lectures and IRE concepts
  - Concepts/Tools required for Assignment 1: Hadoop, Pig, Hive

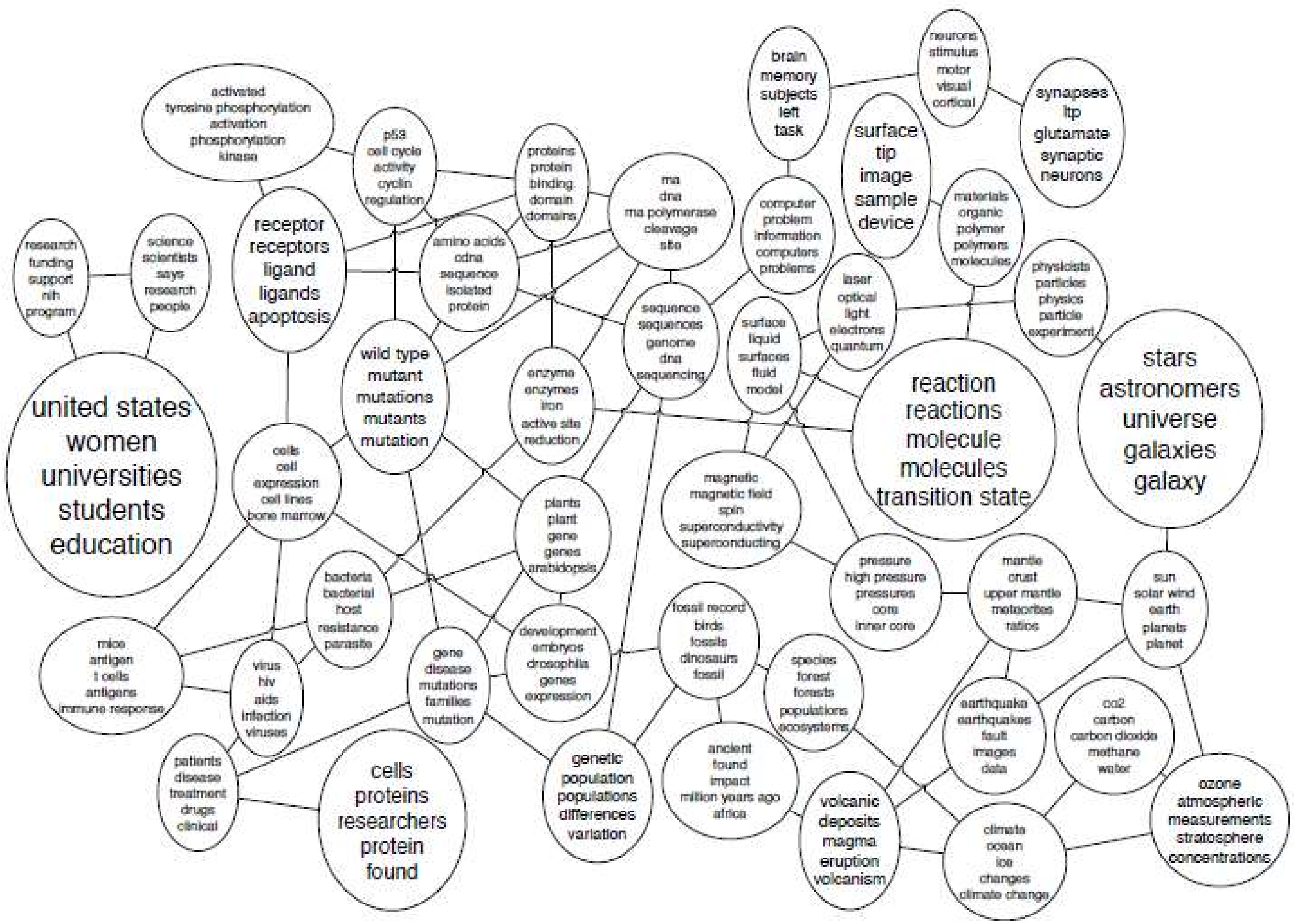
Today's Agenda

- Probabilistic Latent Semantic Analysis (PLSA)
- Latent Dirichlet Allocation (LDA)
- Other Topic Models

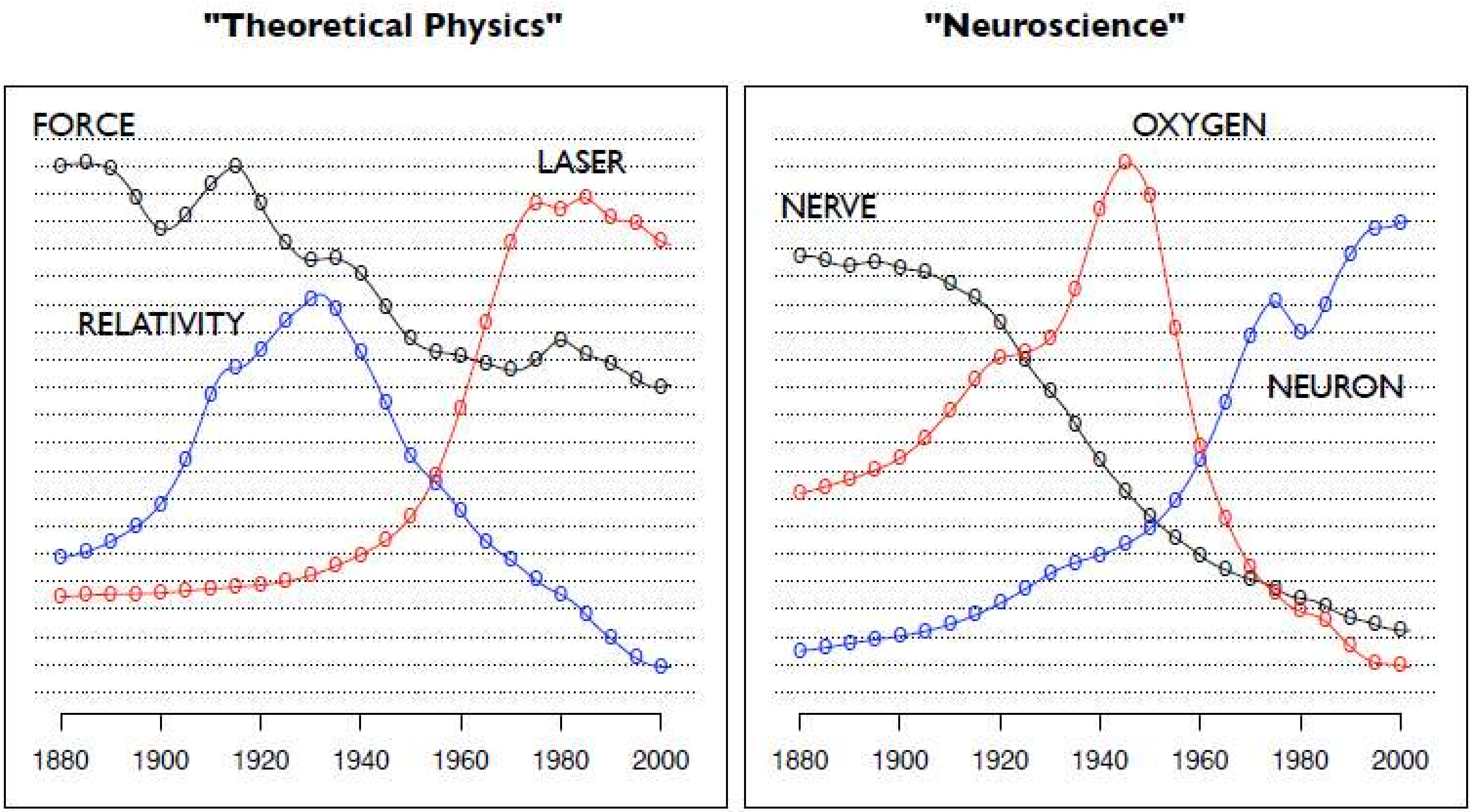
Discover Topics from a Corpus

PRINTING	PLAY	TEAM	JUDGE	HYPOTHESIS	STUDY
PAPER	PLAYS	GAME	TRIAL	EXPERIMENT	TEST
PRINT	STAGE	BASKETBALL	COURT	SCIENTIFIC	STUDYING
PRINTED	AUDIENCE	PLAYERS	CASE	OBSERVATIONS	HOMEWORK
TYPE	THEATER	PLAYER	JURY	SCIENTISTS	NEED
PROCESS	ACTORS	PLAY	ACCUSED	EXPERIMENTS	CLASS
INK	DRAMA	PLAYING	GUILTY	SCIENTIST	MATH
PRESS	SHAKESPEARE	SOCCER	DEFENDANT	EXPERIMENTAL	TRY
IMAGE	ACTOR	PLAYED	JUSTICE	TEST	TEACHER
PRINTER	THEATRE	BALL	EVIDENCE	METHOD	WRITE
PRINTS	PLAYWRIGHT	TEAMS	WITNESSES	HYPOTHESES	PLAN
PRINTERS	PERFORMANCE	BASKET	CRIME	TESTED	ARITHMETIC
COPY	DRAMATIC	FOOTBALL	LAWYER	EVIDENCE	ASSIGNMENT
COPIES	COSTUMES	SCORE	WITNESS	BASED	PLACE
FORM	COMEDY	COURT	ATTORNEY	OBSERVATION	STUDIED
OFFSET	TRAGEDY	GAMES	HEARING	SCIENCE	CAREFULLY
GRAPHIC	CHARACTERS	TRY	INNOCENT	FACTS	DECIDE
SURFACE	SCENES	COACH	DEFENSE	DATA	IMPORTANT
PRODUCED	OPERA	GYM	CHARGE	RESULTS	NOTEBOOK
CHARACTERS	PERFORMED	SHOT	CRIMINAL	EXPLANATION	REVIEW

Model Connections between Topics



Model the Evolution of Topics over Time



Annotate Documents according to these Topics

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE

SKY WATER BUILDING
PEOPLE WATER

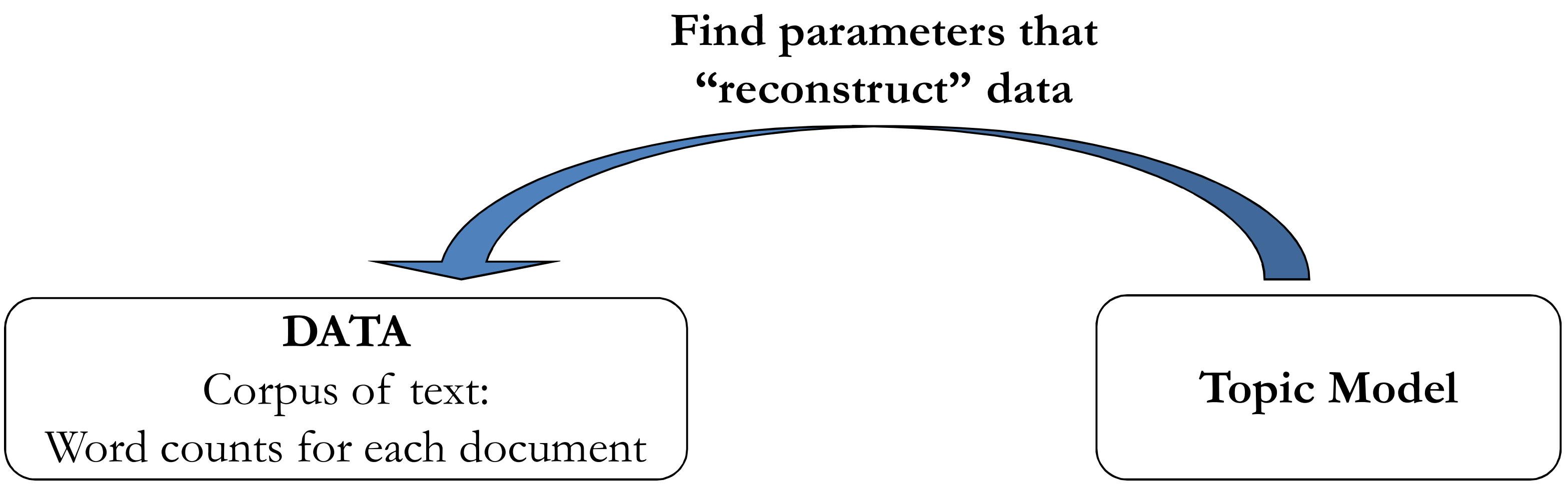
FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

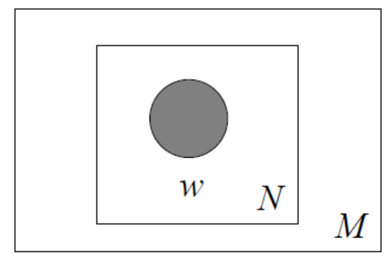
Probabilistic Topic Models

- Extract topics from large collections of text
- Topics are interpretable unlike the arbitrary dimensions of LSA
- Exchangeability assumption: Usually, it is assumed that the order of words in the document is not important. Similarly, order of documents is unimportant.
- Generative model

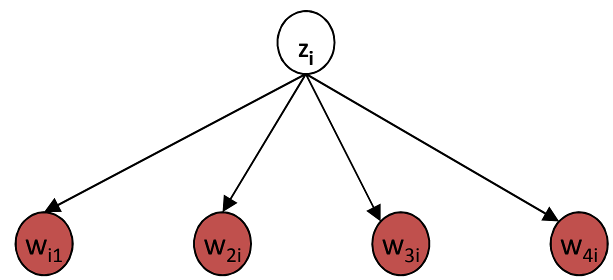


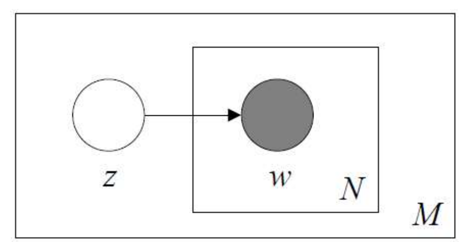
Unigram Model & Mixture of Unigrams

- Unigram model
  - Under the unigram model, the words of every document are drawn independently from a single multinomial distribution
  - $P(\boldsymbol{w}) = \prod_{n=1}^{N} P(w_n)$



- Mixture of unigrams
  - Under this mixture model, each document is generated by first choosing a topic z and then generating N words independently from the conditional multinomial
  - $P(\boldsymbol{w}) = \sum_{z} P(z) \prod_{n=1}^{N} P(w_n|z)$



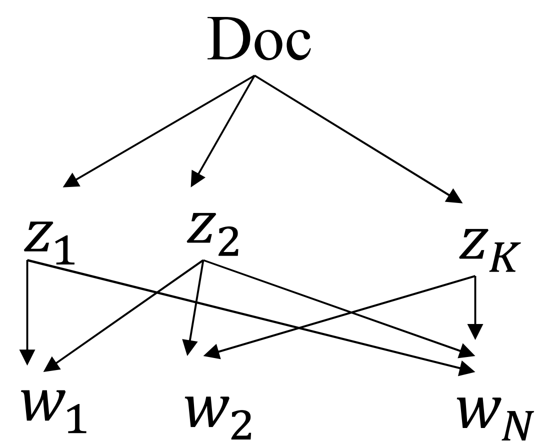


Today's Agenda

- **Probabilistic Latent Semantic Analysis (PLSA)**
- Latent Dirichlet Allocation (LDA)
- Other topic models

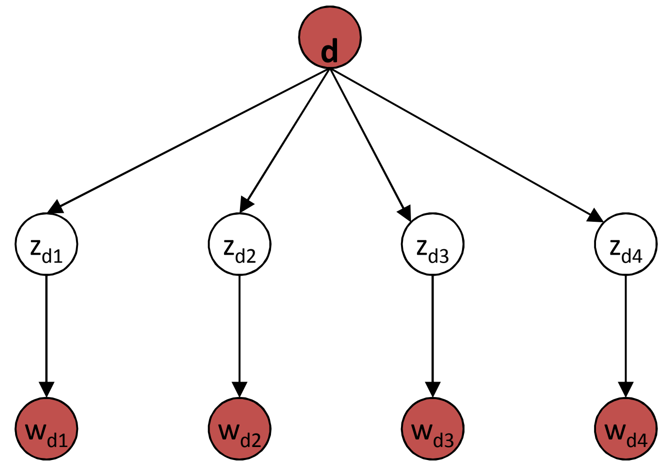
Probabilistic Latent Semantic Analysis

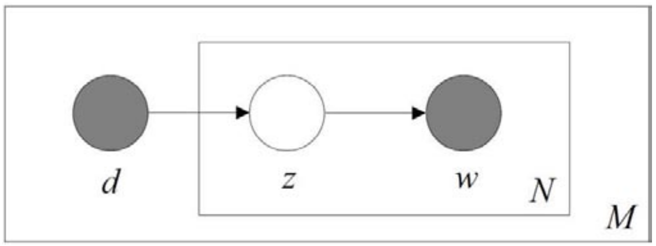
- Generative (Aspect) Model
  - Each document is a probability distribution over latent topics or aspects
  - Each topic z is a probability distribution over words $p(w|z)$
- Model fitting using tempered-EM algorithm
- Shown to solve
  - Polysemy
  - Synonymy
- Has a better statistical foundation than LSA



PLSA Aspect Model

- Generative Model
  - Select a doc with probability $P(d)$
  - Pick a latent topic z with probability $P(z|d)$
  - Generate a word w with probability $P(w|z)$
- Latent Variable model for general co-occurrence data
  - Associate each observation (w,d) with a class variable $z \in \{z_1, \ldots, z_K\}$
- $P(\boldsymbol{w}) = \prod_{n=1}^{N} (\sum_z P(w_n|z) P(z|d))$





Aspect Model

- Joint probability model
  - $P(d,w) = P(d)P(w|d)$

Multinomials

Mixture weights

Multinomial Mixtures

- Where $P(w|d) = \sum_{z \in Z} P(w|z)P(z|d)$

- Conditional Independence Assumption
  - Documents and words are independent given z
- Hence, $P(d,w) = \sum_{z \in Z} P(z)P(d,w|z)$

$$= \sum_{z \in Z} P(z)P(w|z)P(d|z)$$

Advantages of this Model over Document Clustering

- Documents are not related to a single cluster (i.e. aspect)
  - For each z, P(z|d) defines a specific mixture of factors
  - This offers more flexibility, and produces effective modeling
- Now, we have to compute P(z), P(z|d), P(w|z), given the documents(d) and words(w).

Model Fitting with Tempered EM

- We need to max Log-likelihood function from the aspect model $L = \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d, w)$
- We use Expectation Maximization (EM)
  - To avoid over-fitting, tempered EM is proposed

EM Steps

- E-Step
  - Expectation step where expectation of the likelihood function is calculated with the current parameter values
  - Posteriors for the latent variables ***z*** is calculated as
    $$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')}$$
- M-Step
  - Update the parameters with the calculated posterior probabilities
  - Find the parameters that maximizes the likelihood function

M Step

- All these equations use $p(z|d,w)$ calculated in E Step
  - $P(w|z) = \frac{\sum_d n(d,w)P(z|d,w)}{\sum_{d,w'} n(d,w')P(z|d,w')}$
  - $P(d|z) = \frac{\sum_w n(d,w)P(z|d,w)}{\sum_{d',w} n(d',w)P(z|d',w)}$
  - $P(z) = \frac{1}{R}\sum_{d,w} n(d,w)P(z|d,w)$ where $R \equiv \sum_{d,w} n(d,w)$
- Converges to local maximum of the likelihood function

Tempered-EM to avoid Over-Fitting

- Trade off between Predictive performance on the training data and Unseen new data
- Must prevent the model to over fit the training data
- Propose a change to the E-Step
- Reduce the effect of fitting as we do more steps
- Introduce control parameter β
- $P_\beta(z|d,w) = \frac{P(z)[P(d|z)P(w|z)]^\beta}{\sum_{z'} P(z')[P(d|z')P(w|z')]^\beta}$
- β (temperature variable) starts from the value of 1, and decreases

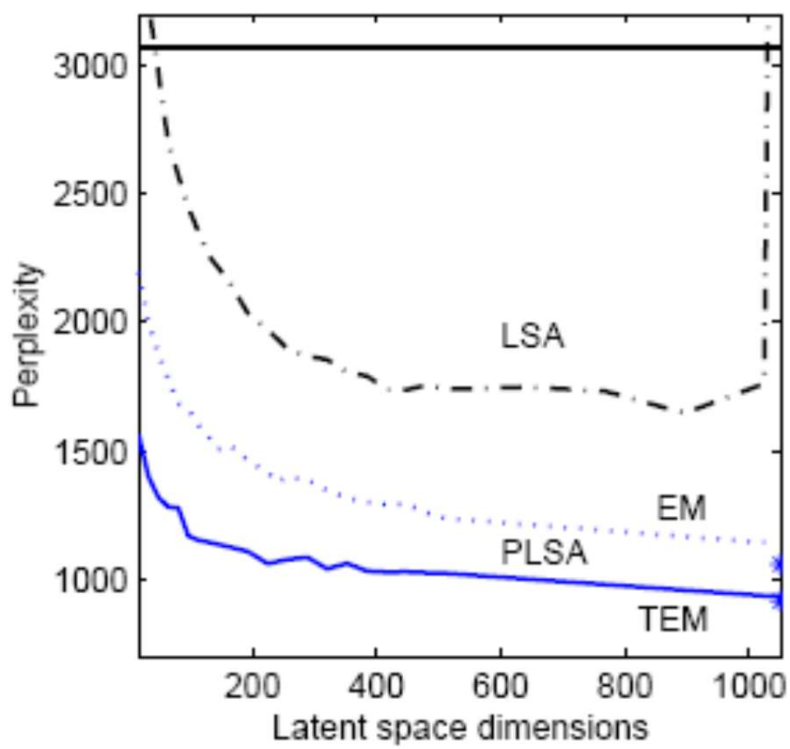
Choosing β

- How to choose a proper β?
- It defines
  - Underfit Vs Overfit
- Simple solution using held-out data (part of training data)
  - Using the training data for β starting from 1
  - Test the model with held-out data
  - If improvement, continue with the same β
  - If no improvement, $\beta \leftarrow n\beta$ where n<1

Perplexity Comparison

- Perplexity – Log-averaged inverse probability on unseen data
- High probability will give lower perplexity, thus good predictions
- MED data

(a)



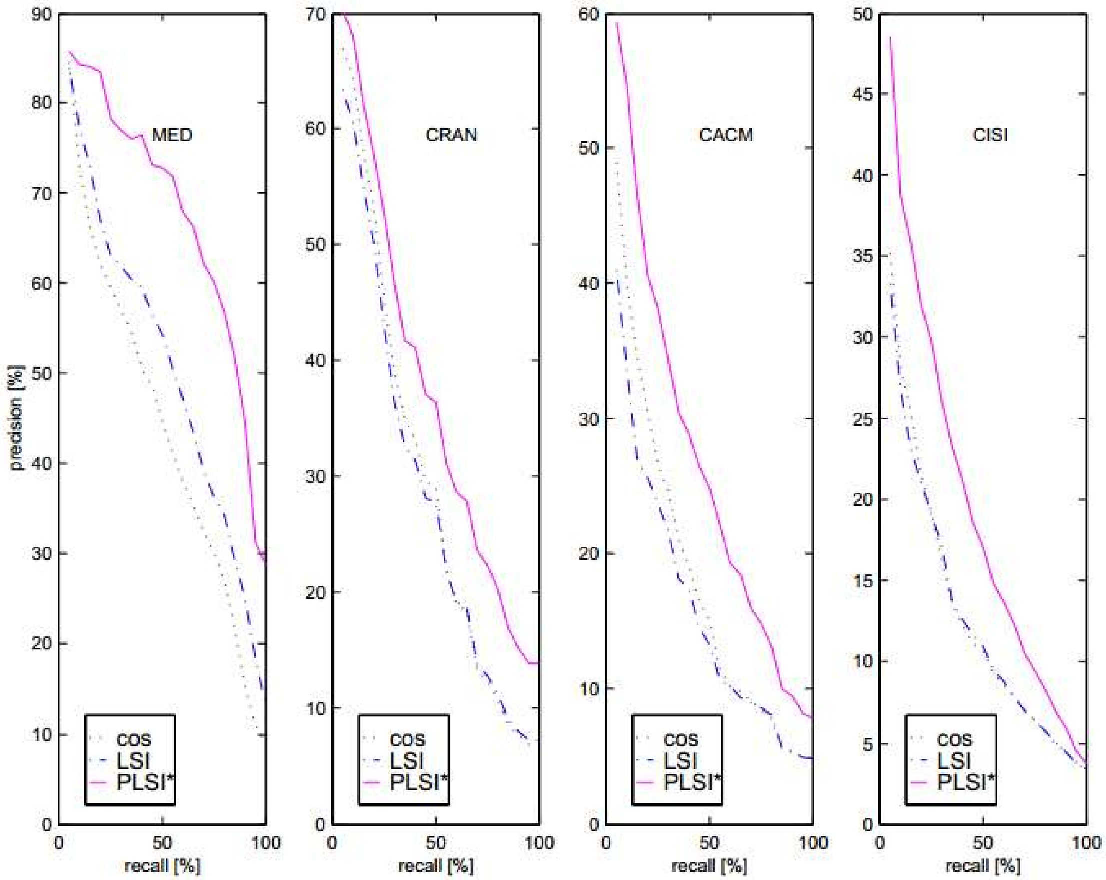
PLSA handles Polysemy

"segment 1"	"segment 2"	"matrix 1"	"matrix 2"	"line 1"	"line 2"	"power 1"	power 2"
imag	speaker	robust	manufactur	constraint	alpha	POWER	load
SEGMENT	speech	MATRIX	cell	LINE	redshift	spectrum	memori
texture	recogni	eigenvalu	part	match	LINE	omega	vlsi
color	signal	uncertainti	MATRIX	locat	galaxi	mpc	POWER
tissue	train	plane	cellular	imag	quasar	hsup	systolic
brain	hmm	linear	famili	geometr	absorp	larg	input
slice	source	condition	design	impos	high	redshift	complex
cluster	speakerind.	perturb	machinepart	segment	ssup	galaxi	arrai
mri	SEGMENT	root	format	fundament	densiti	standard	present
volume	sound	suffici	group	recogn	veloc	model	implement

- Segment: Image region vs. phonetic segment
- Matrix: Rectangular array of numbers vs. material in which something is embedded
- Line: Line in an image vs. line in a spectrum
- Power: Power of radiating objects vs. electric power

PLSA vs. LSI

	MED		CRAN		CACM		CISI	
	prec.	impr.	prec.	impr.	prec.	impr.	prec.	impr.
cos+tf	44.3	-	29.9	-	17.9	-	12.7	-
LSI	51.7	+16.7	*28.7	-4.0	*16.0	-11.6	12.8	+0.8
PLSI	63.9	+44.2	35.1	+17.4	22.9	+27.9	18.8	+48.0
PLSI*	66.3	+49.7	37.5	+25.4	26.8	+49.7	20.1	+58.3



Comparing PLSA and LSA

- LSA and PLSA perform dimensionality reduction
  - In LSA, by keeping only K singular values
  - In PLSA, by having K aspects
- Comparison to SVD
  - U Matrix related to P(d|z) (doc to aspect)
  - V Matrix related to P(z|w) (aspect to term)
  - Σ Matrix related to P(z) (aspect strength)
- The main difference is the way the approximation is done
  - PLSA generates a model (aspect model) and maximizes its predictive power
  - Selecting the proper value of K is heuristic in LSA
  - Model selection in statistics can determine optimal K in PLSA
- The computational cost of LSI is $O(W^2Z^3)$, while computational complexity of PLSA is $O(WDZ^2)$

Today's Agenda

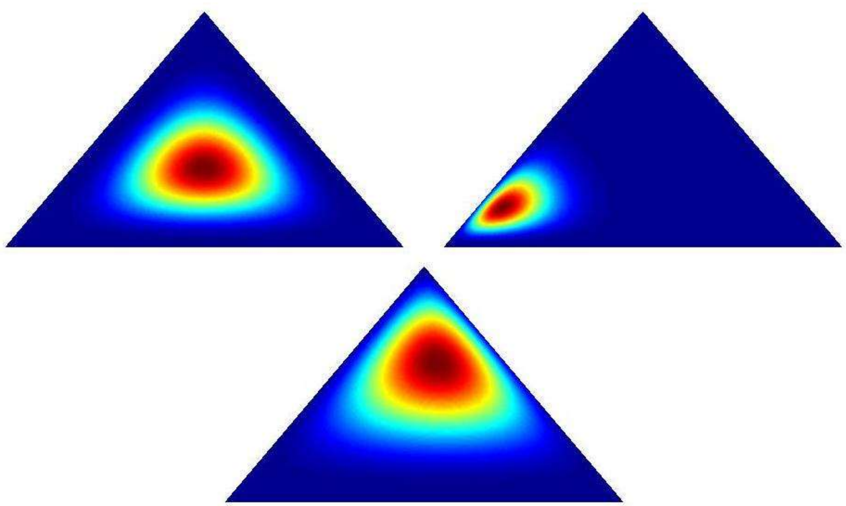
- Probabilistic Latent Semantic Analysis (PLSA)
- **Latent Dirichlet Allocation (LDA)**
- Other topic models

Motivations for LDA

- Problem of PLSI
  - There is no natural way to use it to assign probability to a previously unseen document
  - The linear growth in parameters suggests that the model is prone to overfitting and empirically, overfitting is indeed a serious problem
- We would like to be Bayesian about our topic mixture proportions, rather than fix a single one.

Dirichlet Distributions

- In the LDA model, we would like to say that the *topic mixture proportions* for each document are drawn from some distribution.
- So, we want to put a distribution on multinomials. That is, k-tuples of non-negative numbers that sum to one.
- The space of all of these multinomials has a nice geometric interpretation as a (k-1)-*simplex*, which is just a generalization of a triangle to (k-1) dimensions.

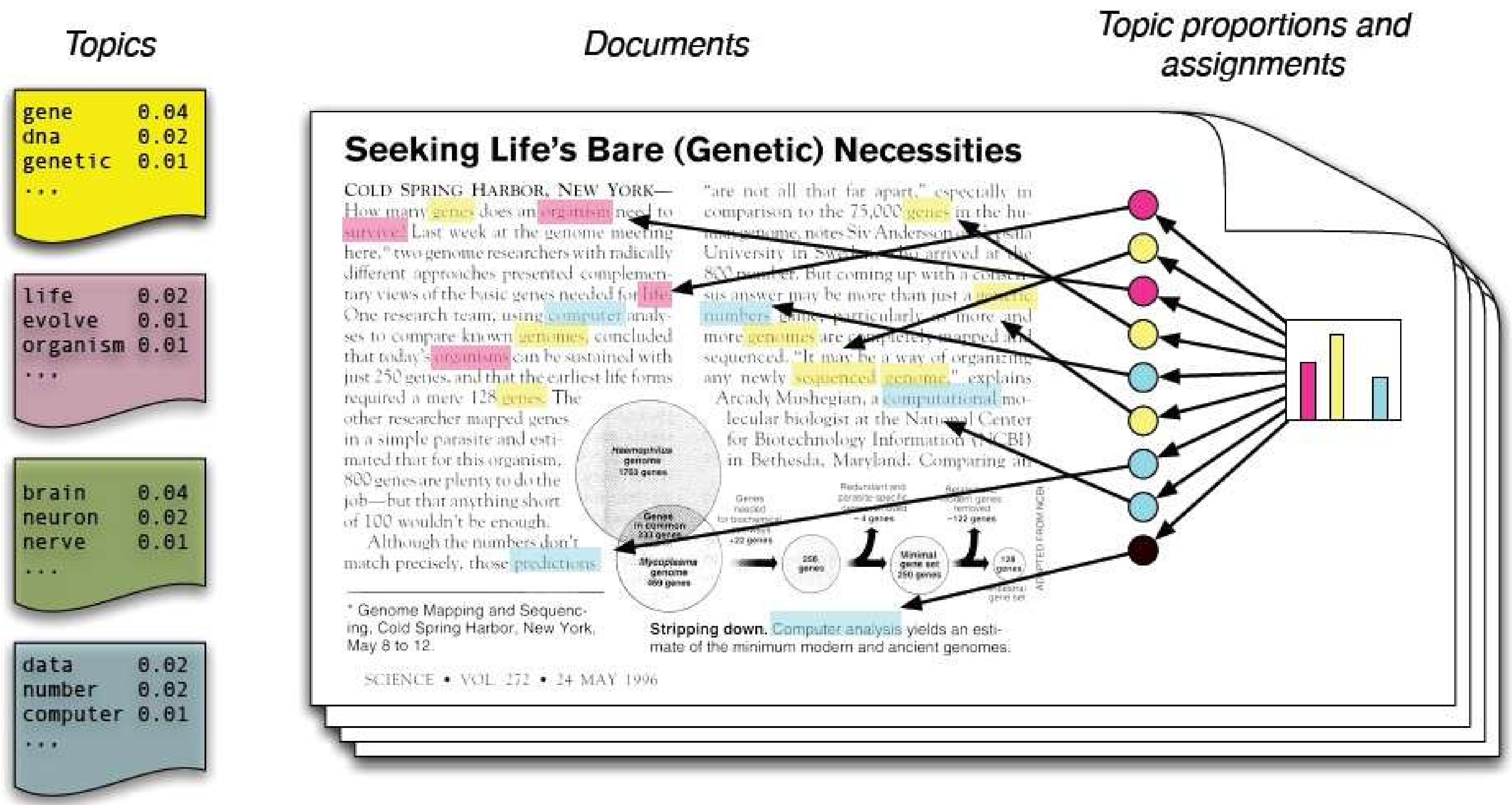


Dirichlet Distributions

- Useful Facts:
  - This distribution is defined over a (k-1)-simplex. That is, it takes k non-negative arguments which sum to one. Consequently it is a natural distribution to use over multinomial distributions.
  - In fact, the Dirichlet distribution is the conjugate prior to the multinomial distribution. (This means that if our likelihood is multinomial with a Dirichlet prior, then the posterior is also Dirichlet!)
  - The Dirichlet parameter α_i can be thought of as a prior count of the i^{th} topic.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

LDA Intuitive Representation

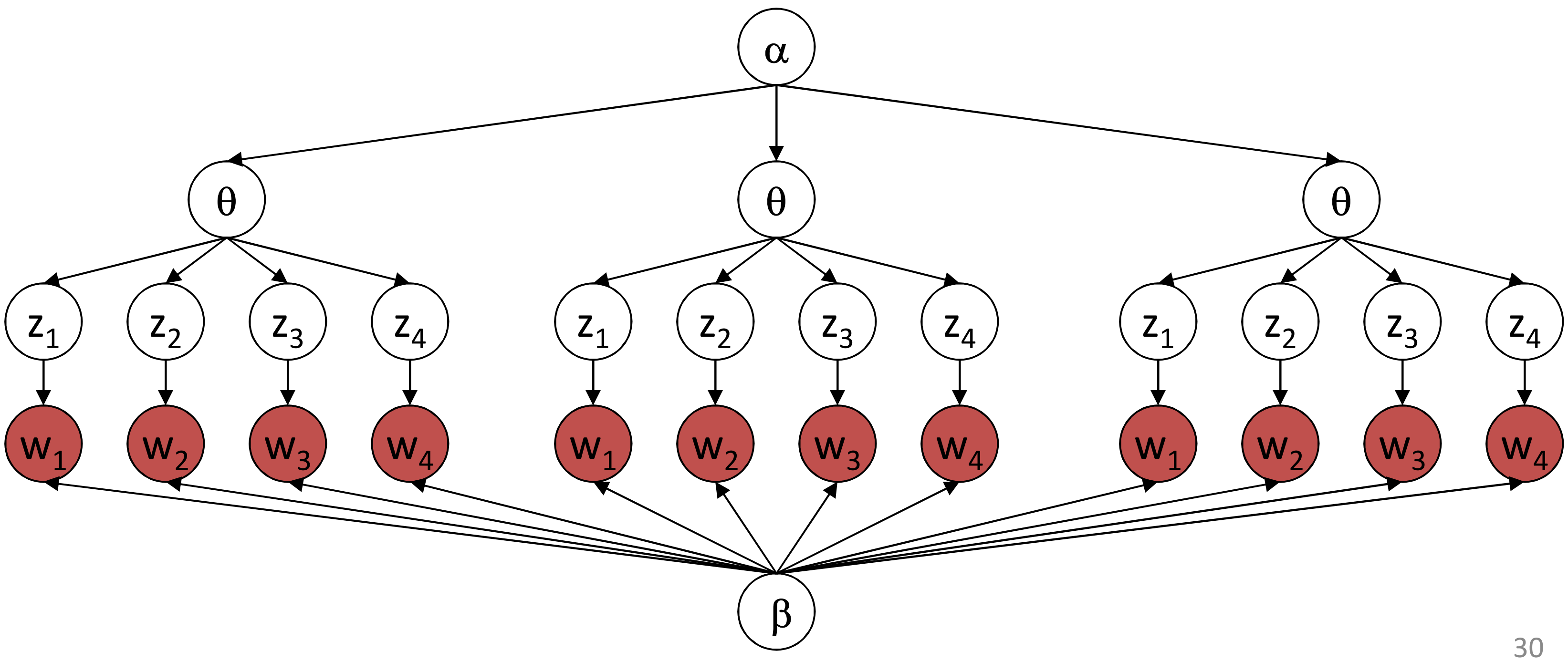


The LDA Model

For each document,

- Choose $N \sim Poisson(\xi)$
- Choose $\theta \sim \text{Dirichlet}(\alpha)$
- For each of the N words w_n:
  - Choose a topic $z_n \sim \text{Multinomial}(\theta)$
  - Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

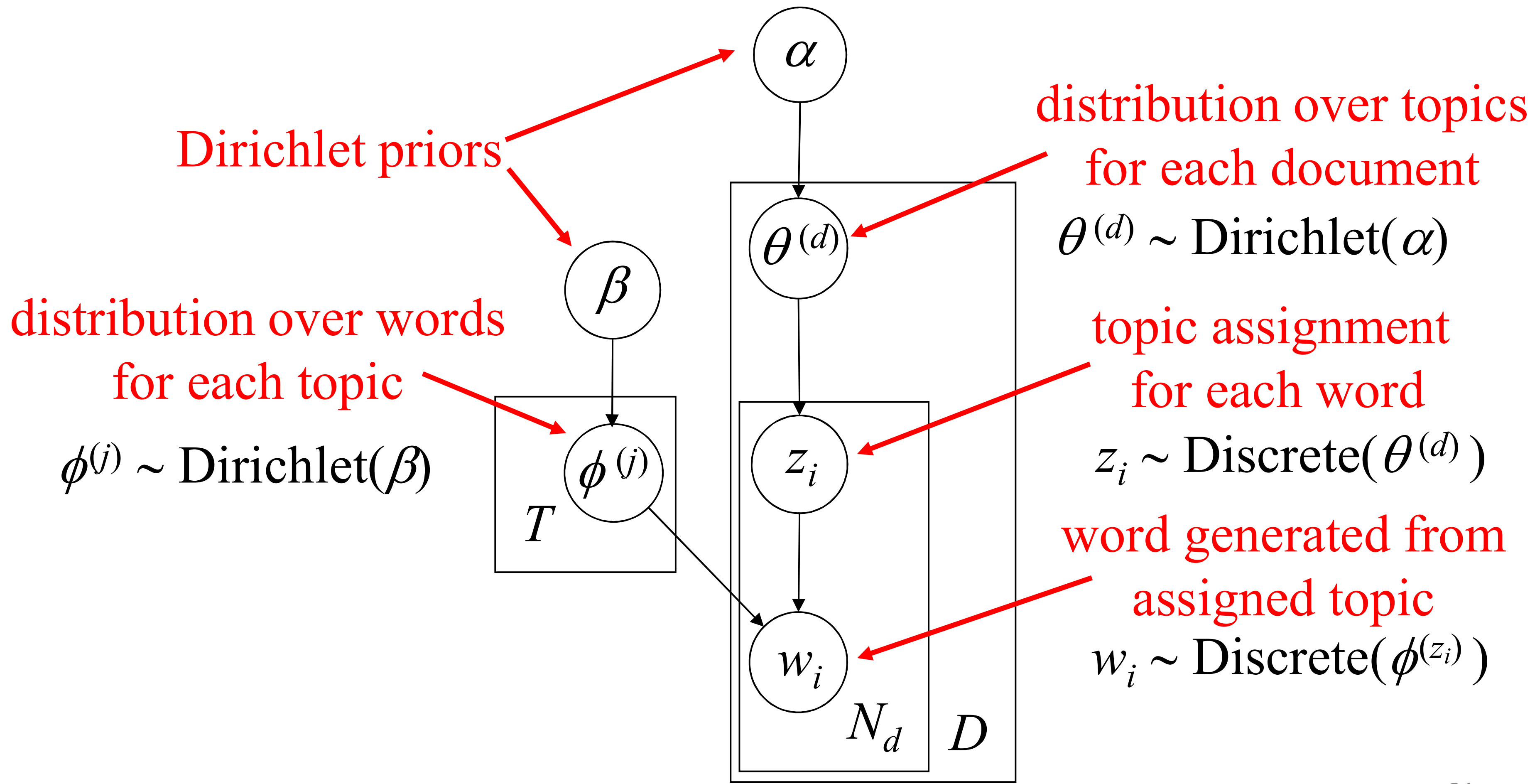
Note Multinomial is also called as Discrete



Latent Dirichlet Allocation

Let K=#topics, D=#documents, N_d=#words in doc d, V=vocabulary size

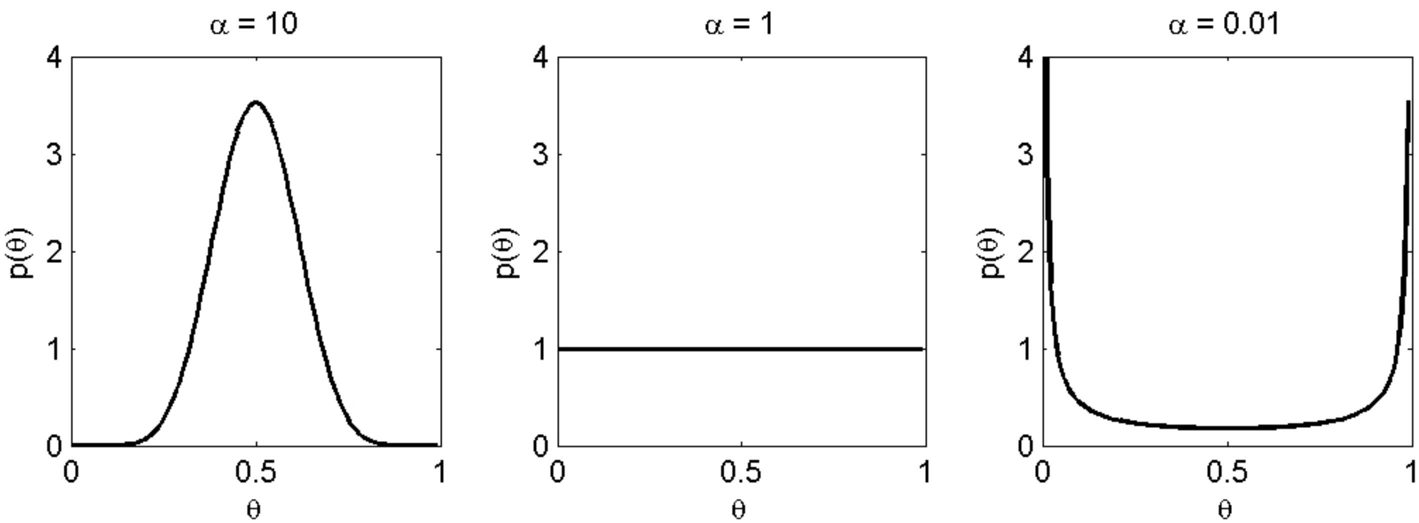
$$\alpha^{K\times 1}, \beta^{V\times 1}, \phi^{K\times V}, \theta^{K\times D}$$



Dirichlet Priors

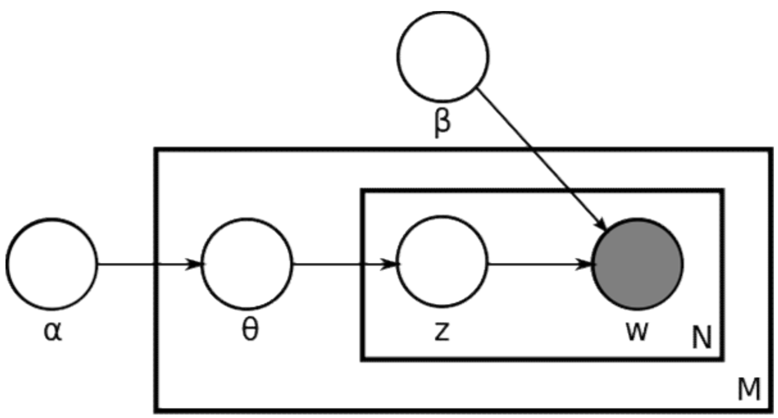
pdf	$\frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$ where $\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$

- Hyperparameters α determine form of the prior

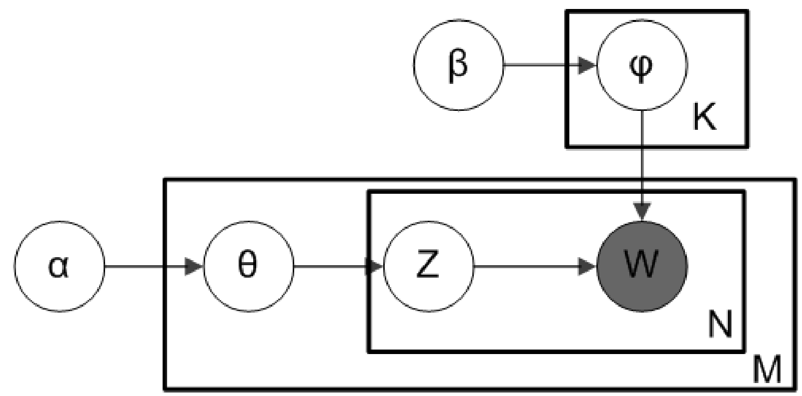


Parameters and Variables in Latent Dirichlet Allocation

- There are three levels to LDA representation
  - α, β are corpus-level parameters
  - θ_d are document-level variables
  - z_{dn}, w_{dn} are word-level variables



LDA



Smoothed LDA

Number of Parameters

- Unigram model
  - No parameters
- Mixture of unigrams
  - K-1 parameters; $p(z)$
- PLSA/PLSI
  - KV+KM parameters; $p(w_n|z)$ and $p(z|d)$
- LDA
  - K+KV (K+V for smoothed LDA); α and β

Relationship with Other Latent Variable Models

- The unigram model find a single point on the word simplex and posits that all words in the corpus come from the corresponding distribution.
- The mixture of unigram models posits that for each documents, one of the k points on the word simplex is chosen randomly and all the words of the document are drawn from the distribution
- The pLSI model posits that each word of a training documents comes from a randomly chosen topic. The topics are themselves drawn from a document-specific distribution over topics.
- LDA posits that each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter

Recall EM

- Situation: You have some data X which you believe was obtained from some model parameterized with parameters θ and has some latent variables Z.
- Aim is to find MLE of θ such that log data likelihood is maximized across all possible values for the latent variables.
- E.g., Gaussian mixture models where X=data, $Z = C_j$, $\theta = \{\pi_k, \mu_k, \sigma_k\}$
- Finding values of both latent variables and parameters together may not be possible in a closed form
  - This needs taking derivative of log data likelihood wrt each latent variable and each parameter
  - Results into interlocking equations which may not be solvable
- So an iterative solution is used
  - Find posterior probability values of latent variables assuming that parameters are known. (E step)
  - Find parameter values assuming latent variable values are known. (M step)

EM Refresher

The General EM Algorithm Summary:
Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variable $\mathbf{X}$ and latent variables $\mathbf{Z}$, with parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Initialize the parameters $\boldsymbol{\theta}^{old}$
2. **E Step** Construct $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$

   $$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

   which is the conditional *expectation* of the complete-data log-likelihood.
3. **M Step** Evaluate $\boldsymbol{\theta}^{new}$ via

   $$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{12}$$

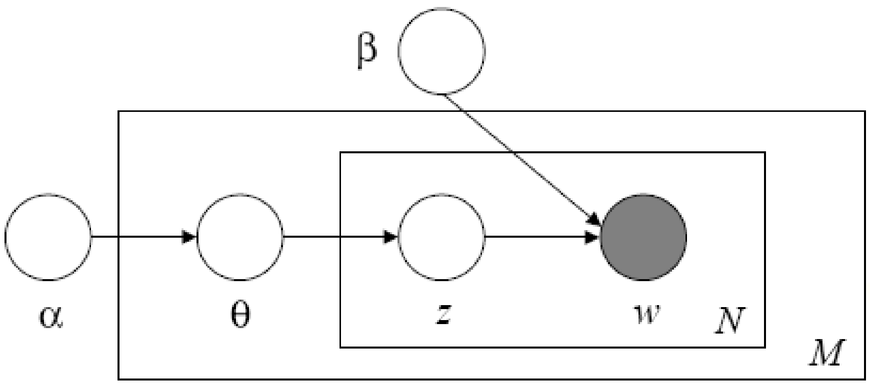
4. Check log likelihood and parameter values for convergence, if not converged let $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$ and return to step 2.

EM for LDA

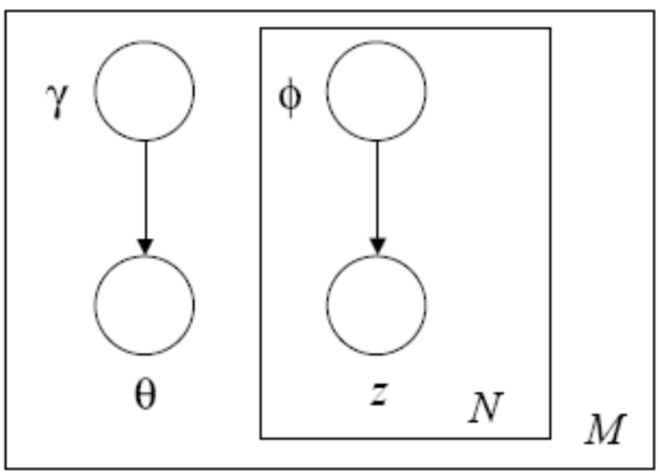
- For LDA, latent variables are θ and z. Parameters are α and β. Data is words in a document.
- EM
  - Initialize α and β
  - Iterate
    - Find posterior θ and z using current α and β
    - Find new α and β
- The inference problem in LDA (E step of EM) is to compute the posterior of the hidden variables given a document and corpus parameters α and β. That is, compute $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$.
  - $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)}$
- But $p(w|\alpha, \beta)$ is difficult to compute (intractable)
- So we turn to alternatives
  - Variational Inference
  - Markov Chain Monte Carlo (Gibbs sampling)
  - Kalman Filtering

Variational EM for LDA

- The idea is to obtain a tractable lower bound on log likelihood.



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

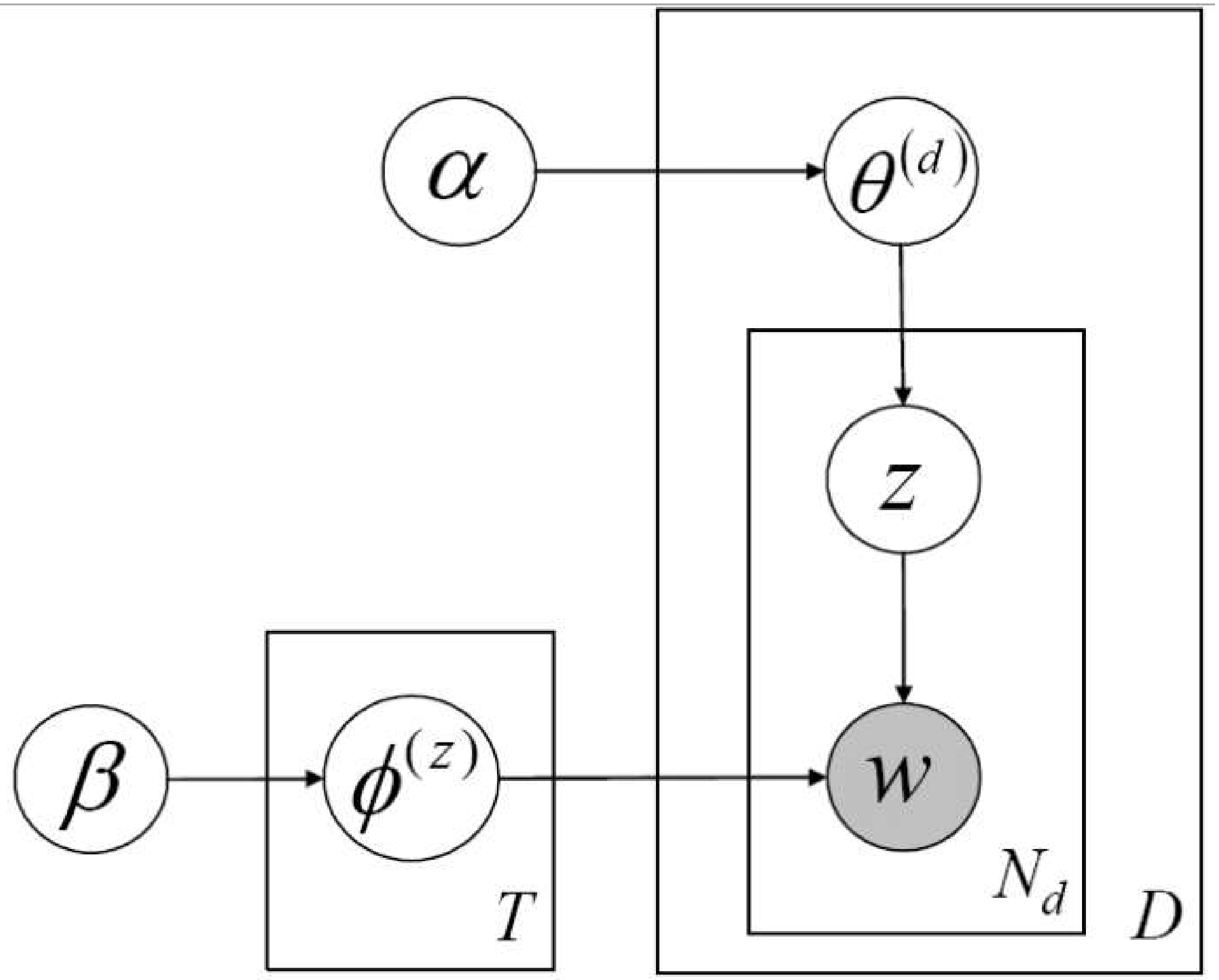


$$q(\theta, z \mid \gamma, \phi) = p(\theta \mid \gamma)\prod_{n=1}^{N} p(z_n \mid \phi_n)$$

Variational EM LDA Algorithm

- Input: Number of topics K, Corpus with M documents and N_d words in document d
- Output: Model parameters: β, θ, z
- Initialize $\phi, \gamma, \alpha, \beta$
- Iterate until convergence of data log likelihood
  - E step: Estimate γ, ϕ using α and β from previous iteration
  - M step: Estimate α and β using γ and ϕ from previous iteration
- Return parameters

Collapsed Gibbs Sampling Method



Gibbs Sampling

- The point of Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution.
- To obtain k samples of $X = \{x_1, \dots, x_n\}$ from joint distribution $p(x_1, \dots, x_n)$
  - Start with initial value X^0 for each variable
  - For each sample, $i = \{1, \dots, k\}$, sample each variable x_j^i from conditional distribution $p(x_j | x_i^i, \dots, x_{j-1}^i, x_{j+1}^i, \dots, x_n^i)$
- Burn in period
- Thinning: Considering every n^{th} sample.

Gibbs Sampling for LDA

- Gibbs sampling procedure is to estimate

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w})$$

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d)} + \alpha}{n_{-i,\cdot}^{(d)} + K\alpha}$$

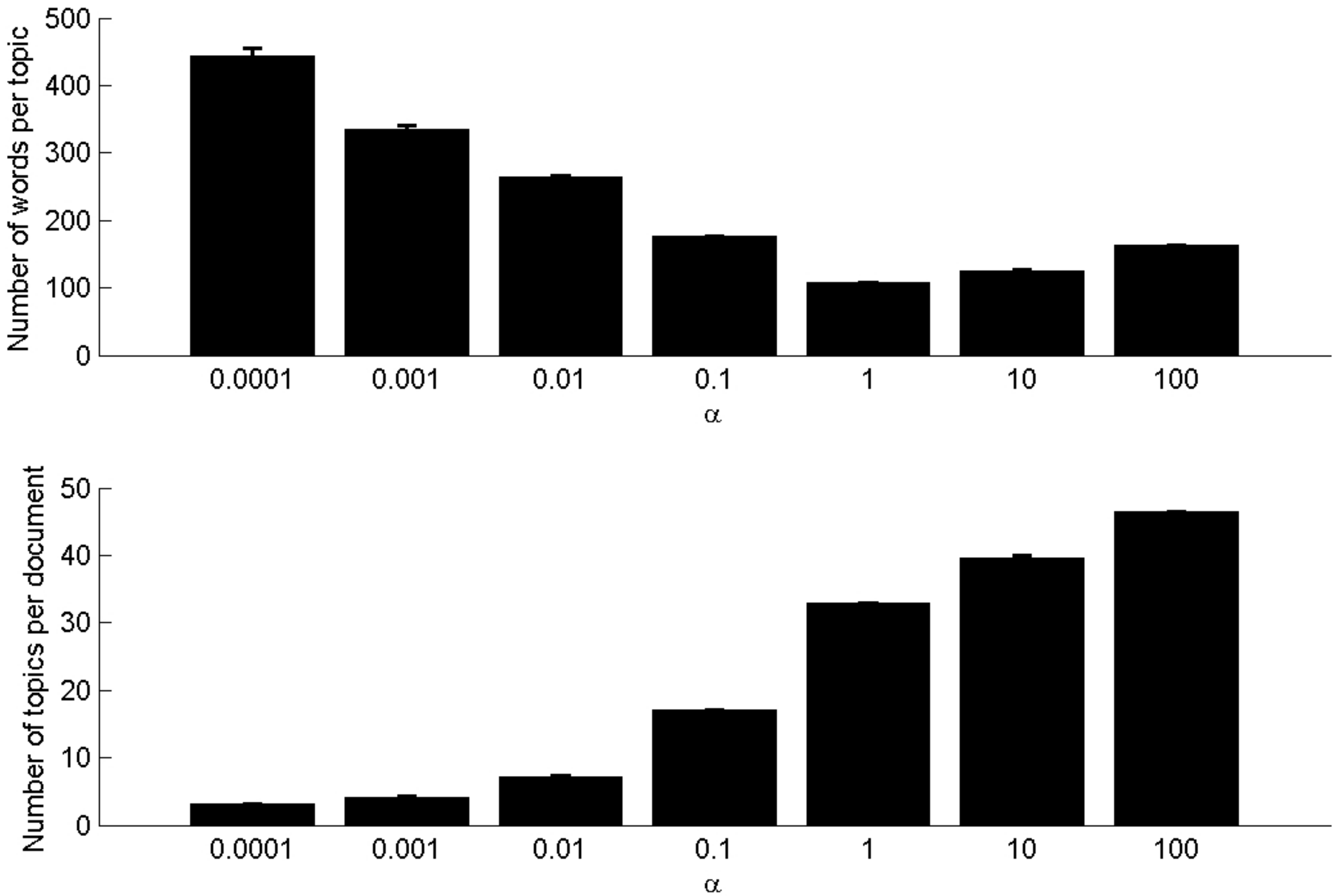
Need to record four count variables:

- document-topic count $n_{-i,j}^{(d)}$
- document-topic sum $n_{-i,\cdot}^{(d)}$ (actually a constant)
- topic-term count $n_{-i,j}^{(w_i)}$
- topic-term sum $n_{-i,j}^{(\cdot)}$

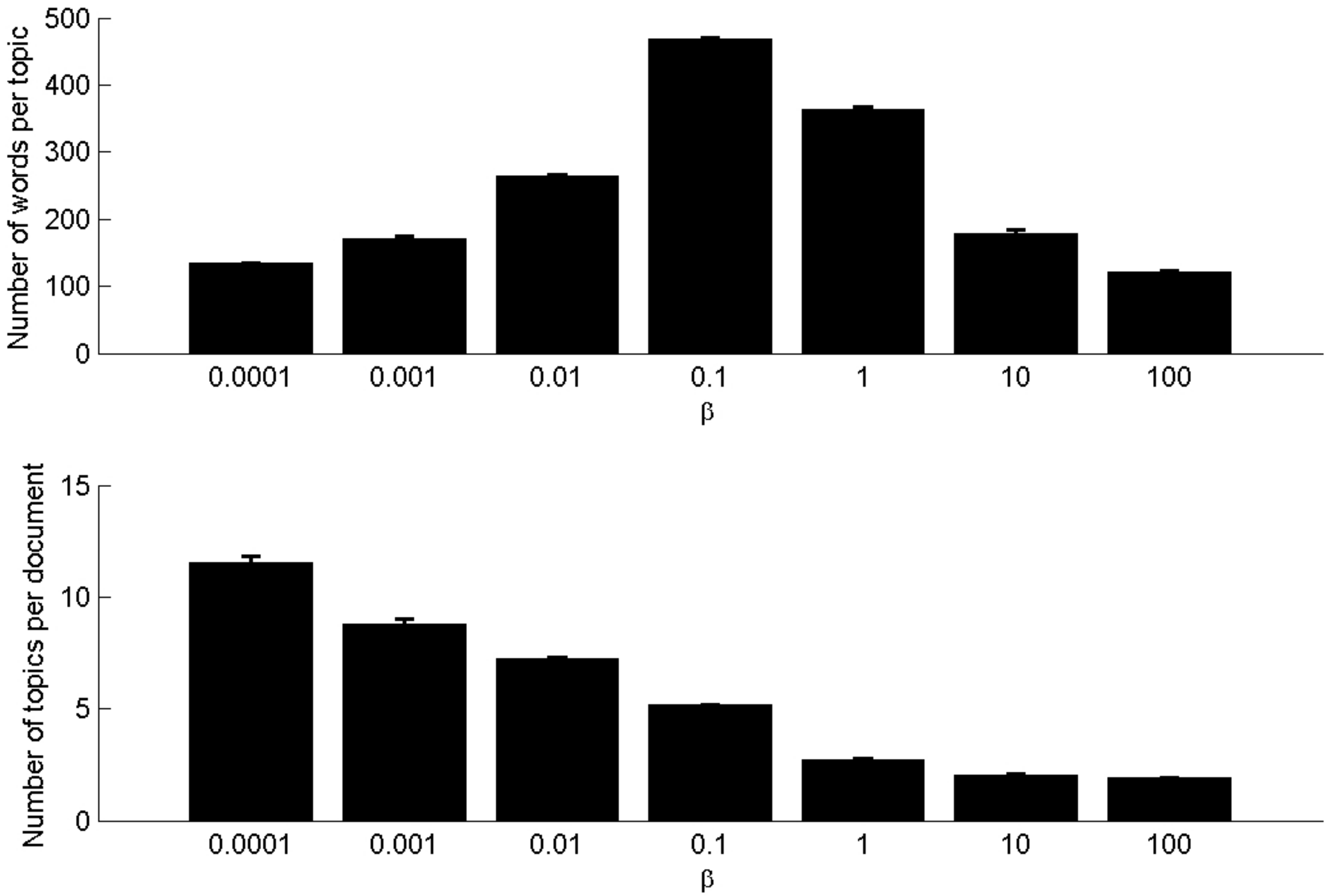
Effects of Hyperparameters

- α and β control the relative sparsity of ϕ and θ
  - smaller α, fewer topics per document
  - smaller β, fewer words per topic
- Good assignments **z** compromise in sparsity

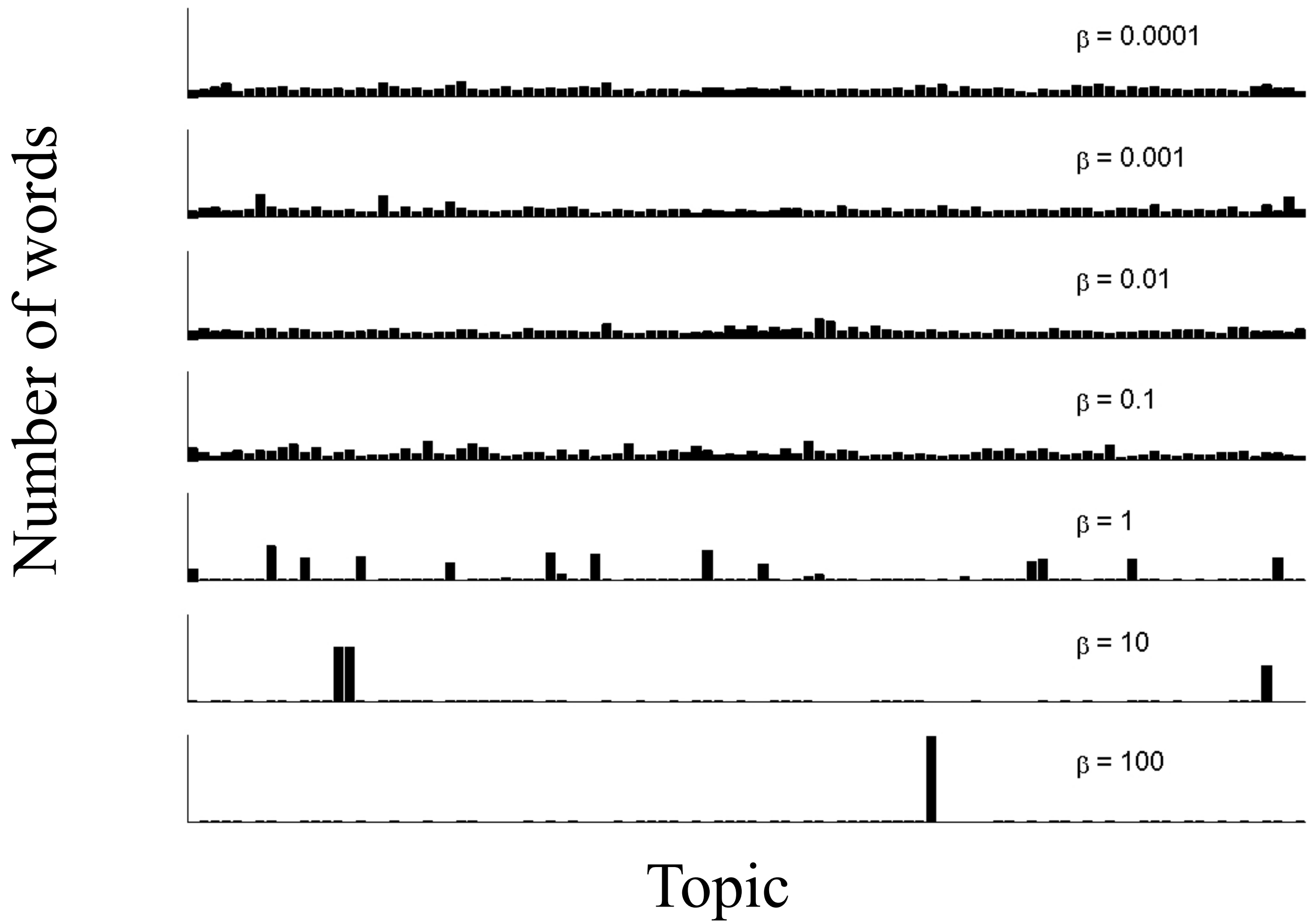
Varying α



Varying β



Number of Words per Topic

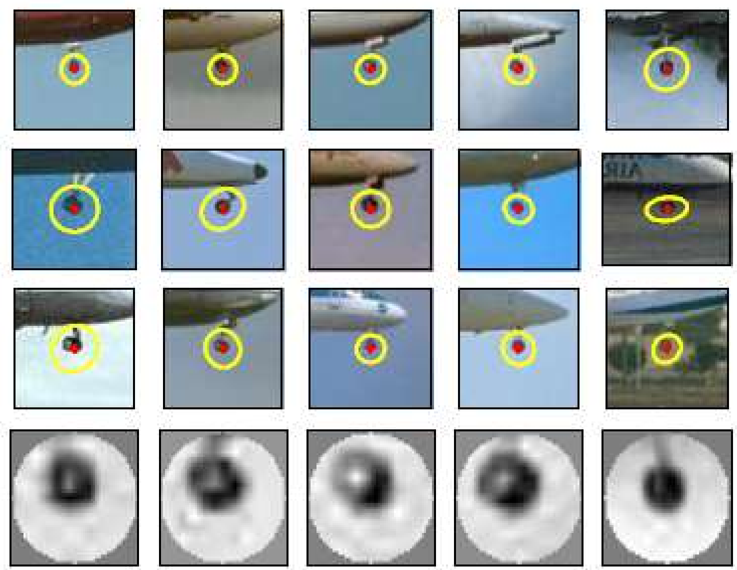


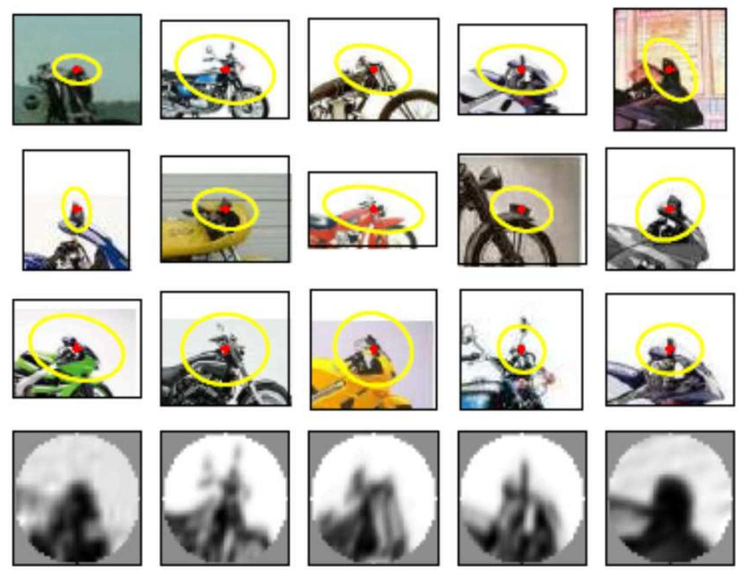
Today’s Agenda

- Probabilistic Latent Semantic Analysis (PLSA)
- Latent Dirichlet Allocation (LDA)
- **Other topic models**
  - Topic modeling in a nutshell:
    Text + (Probabilistic Graphical Model + Inference algorithm) -> Topics

Visual Words

- Idea: Given a collection of images,
  - Think of each image as a document.
  - Think of feature patches of each image as words.
  - Apply the LDA model to extract topics.



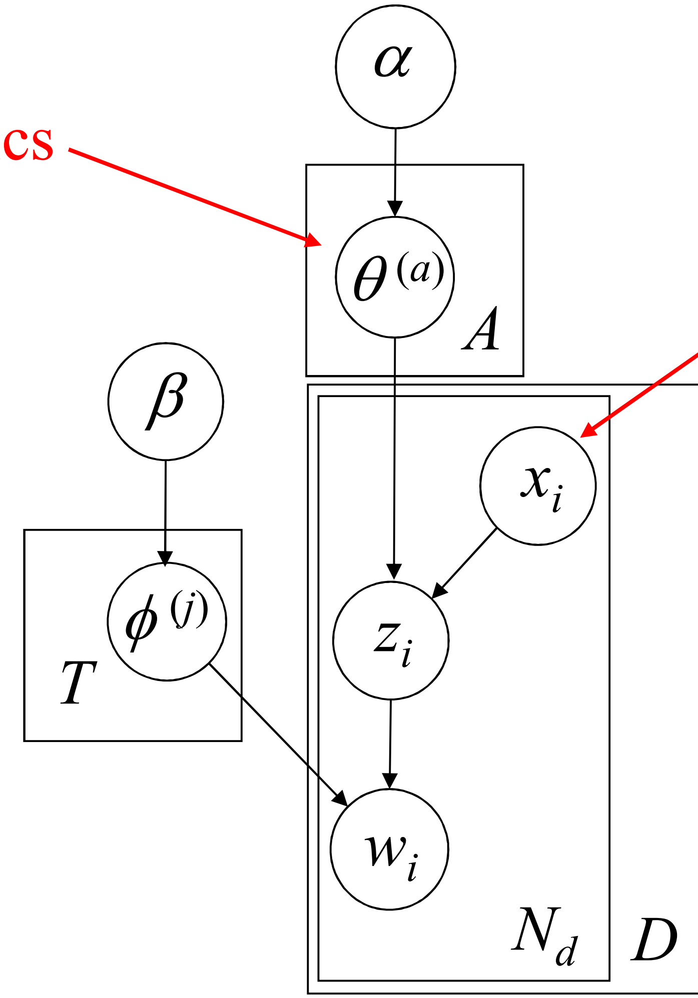


The Author-Topic Model

each author has a distribution over topics

$$\theta^{(a)} \sim \text{Dirichlet}(\alpha)$$

$$\phi^{(j)} \sim \text{Dirichlet}(\beta)$$



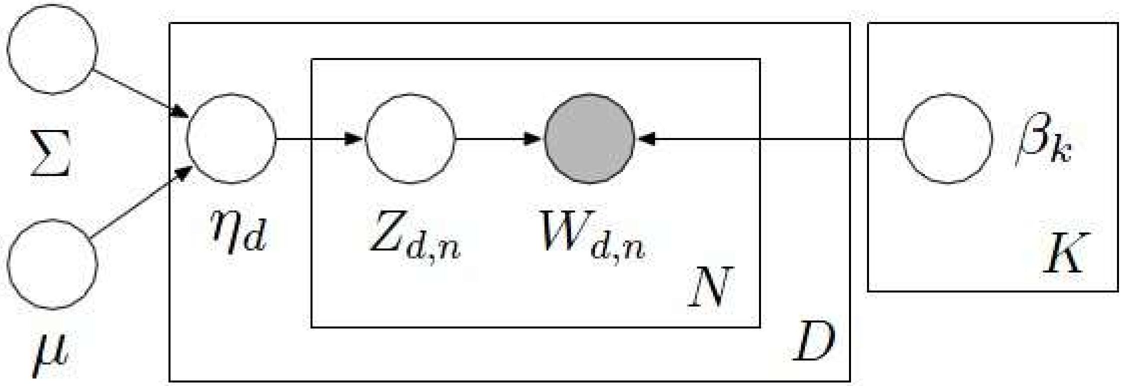
the author of each word is chosen uniformly at random

$$x_i \sim \text{Uniform}(A^{(d)})$$

$$z_i \sim \text{Discrete}(\theta^{(x_i)})$$

$$w_i \sim \text{Discrete}(\phi^{(z_i)})$$

Correlated Topic Model



(1) Draw $\eta \mid \{\mu, \Sigma\} \sim N(\mu, \Sigma)$.
(2) For $n \in \{1, \dots, N\}$:
  (a) Draw topic assignment $Z_n \mid \eta$ from Mult($f(\eta)$).
  (b) Draw word $W_n \mid \{z_n, \beta_{1:K}\}$ from Mult(β_{z_n}).

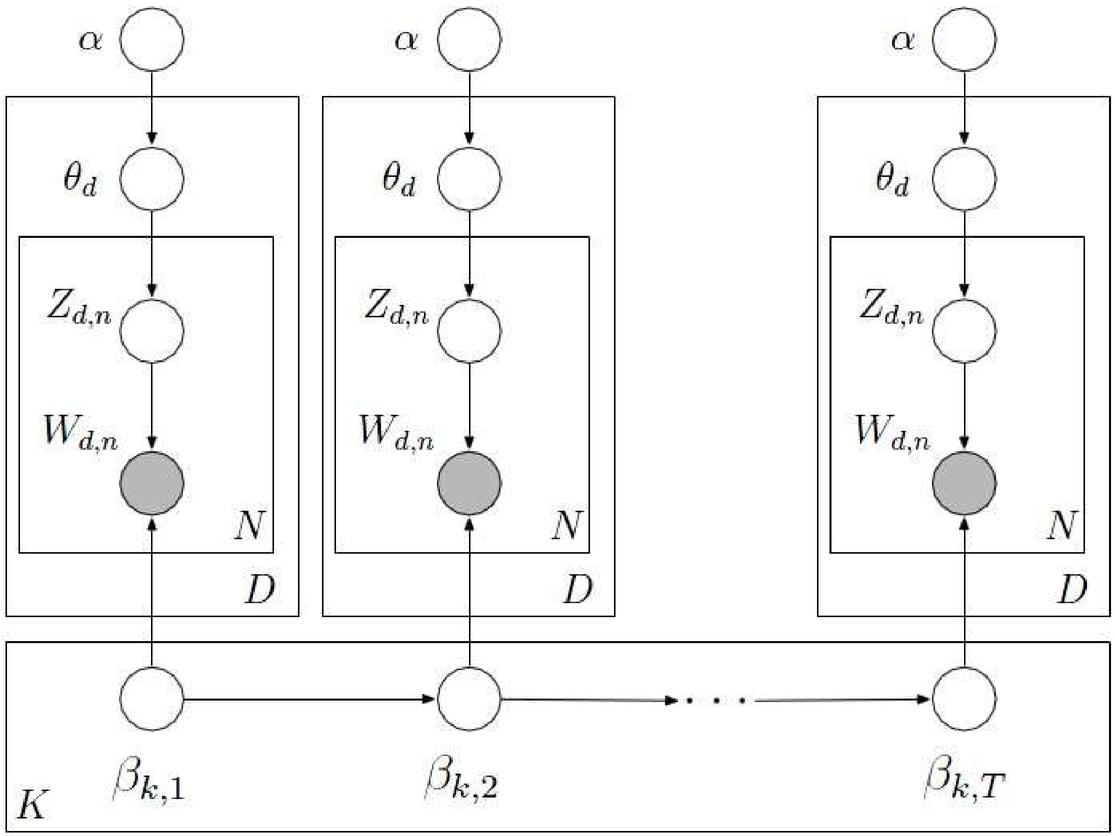
The function that maps the real-vector η to the simplex is

$$f(\eta_i) = \frac{\exp\{\eta_i\}}{\sum_j \exp\{\eta_j\}}. \tag{15}$$

Correlation between topics: E.g., an article about genetics may be likely to also be about health and disease, but unlikely to also be about x-ray astronomy

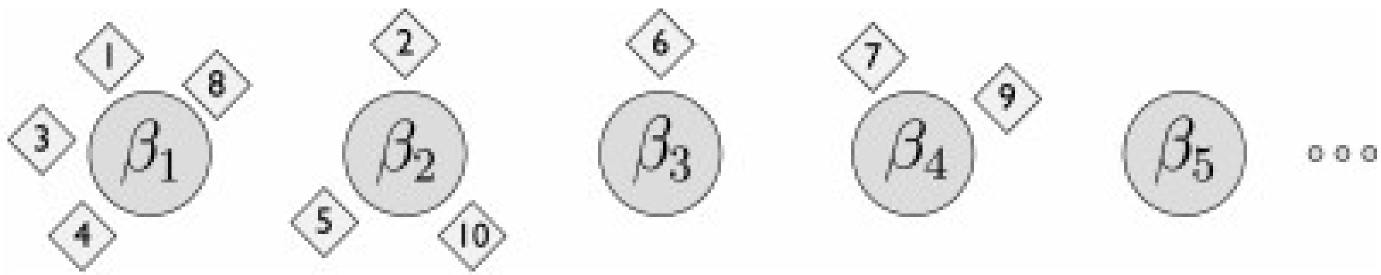
Dynamic Topic Model

- Documents order
- Documents are exchangable in LDA
- (DTM) captures the evolution of topics in a sequentially organized corpus of documents (Ordered by year)



Non-parametric Topic Models

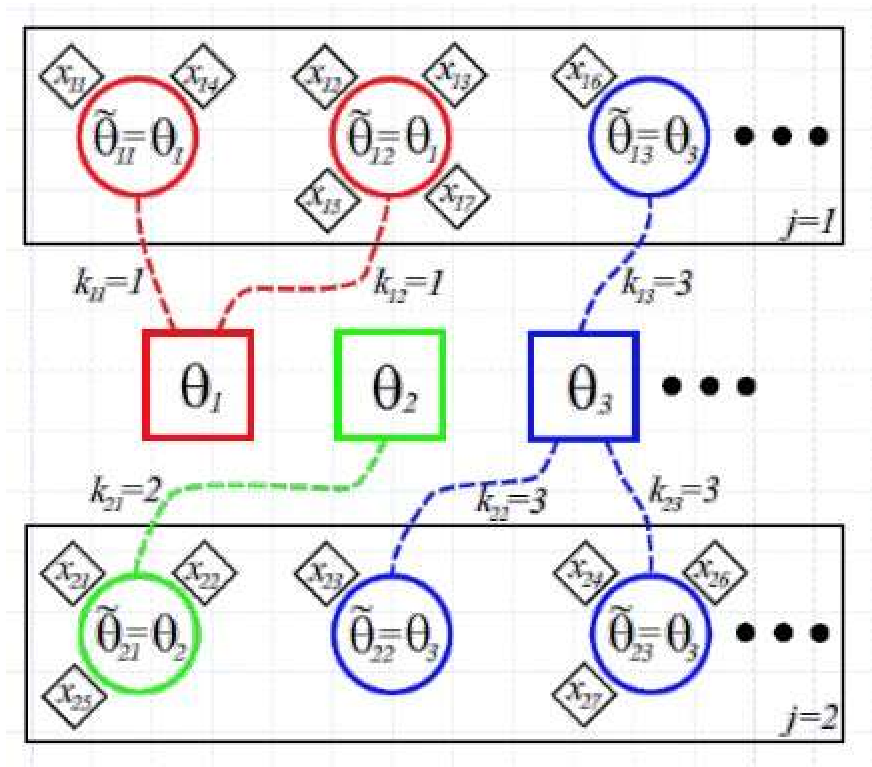
- Dirichlet process
- Can be seen as a infinite dimension Dirichlet distribution
- Chinese restaurant process



Customer 1 is seated at an unoccupied table with probability 1. At time *n* + 1, a new customer chooses uniformly at random to sit at one of the following *n* + 1 places: directly to the left of one of the *n* customers already sitting at an occupied table, or at a new, unoccupied circular table. Each table thus corresponds to a block of a random partition.

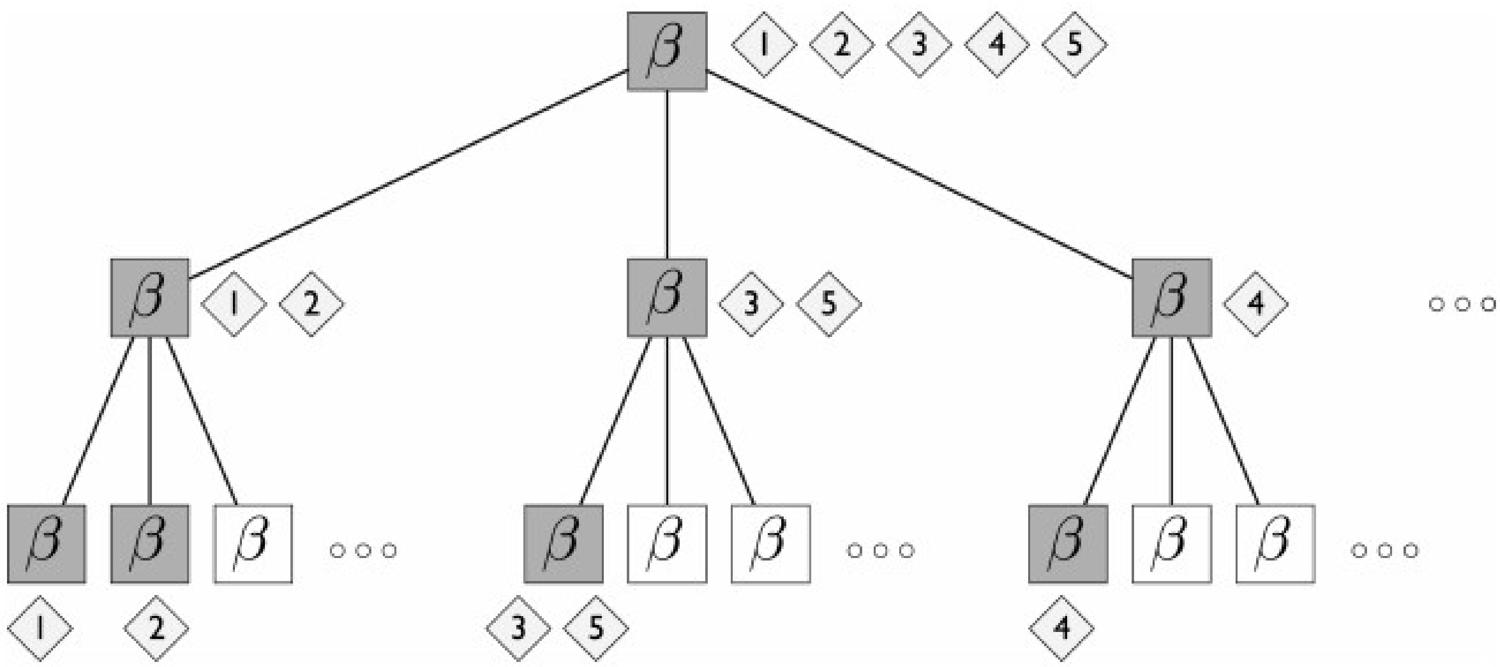
Non-parametric Topic Models

- Hierarchical Dirichlet Process



Hierarchical Topic Model

The nested Chinese Restaurant Process



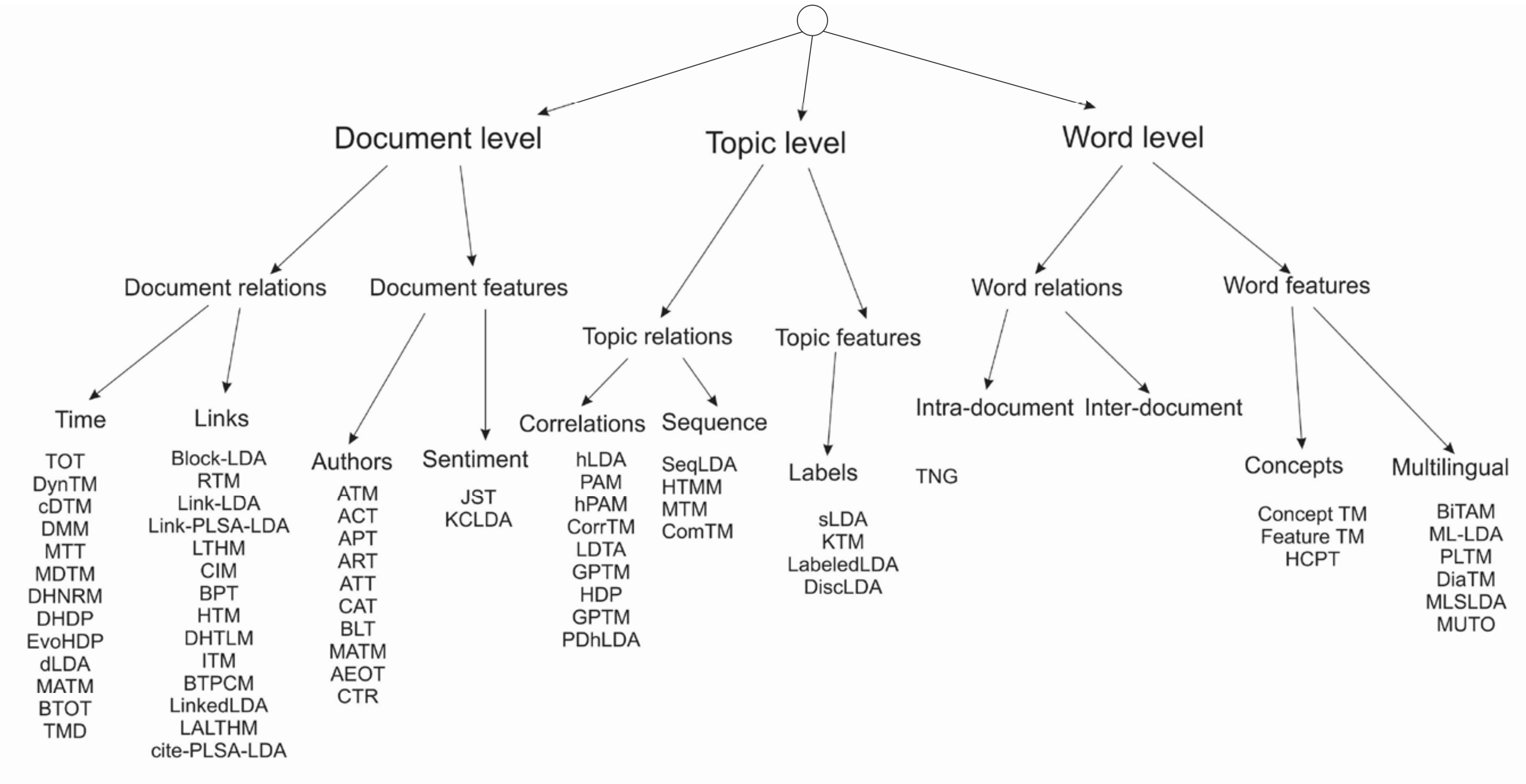
Classification of Other Topic Models

- Relaxing the exchangeability assumption
  - Document relations
    - Time
    - Links
  - Topic relations
    - Correlations
    - Sequence
  - Word relations
    - Intra-document (Sequentiality)
    - Inter-document (Entity recognition)

Classification of Other Topic Models

- Modeling with additional data
  - Document features
    - Sentiment
    - Authors
  - Topic features
    - Labels
  - Word features
    - Concepts

Classification of Other Topic Models



Take-away Messages

- Last lecture we studied about LSA as way to cluster words into concepts
- PLSA provides interpretable word clusters (topics/aspects/classes)
- LDA generalizes to unseen documents with limited number of parameters
- Many topic models have been proposed since then to capture large variety of intuitions to group words into topics

Further Reading

- Thomas Hofmann, Probabilistic Latent Semantic Analysis. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99) http://www.cs.brown.edu/~th/papers/Hofmann-UAI99.pdf
- D. M. Blei et al., “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, 3, pp. 993–1022, January 2003.
- D. Blei and J. Lafferty, “Topic models,” in A. Srivastava and M. Sahami, (eds.), *Text Mining: Theory and Applications*. Taylor and Francis, 2009.
- D. Blei. "Introduction to Probabilistic Topic Models," http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf

Preview of Lecture 7: Recommender Systems (1)

- Introduction
- Formal Model
- Offline Components: Collaborative Filtering in Cold-start Situations

Disclaimers

- This course represents opinions of the instructor only. It does not reflect views of Microsoft or any other entity (except of authors from whom the slides have been borrowed).
- Algorithms, techniques, features, etc mentioned here might or might not be in use by Microsoft or any other company.
- Lot of material covered in this course is borrowed from slides across many universities and conference tutorials. These are gratefully acknowledged.

Thanks!

Variational Inference for LDA Derivation

Inference (E step of EM)

- The inference problem in LDA (E step of EM) is to compute the posterior of the hidden variables given a document and corpus parameters α and β. That is, compute $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$.
  - $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)}$

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

$$p(\mathbf{w}|\mathbf{z}, \beta) = \prod^{N} \beta_{z_n, w_n}$$

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

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

Marginalize over θ and z

Inference (E step of EM)

- $p(\boldsymbol{w}|\alpha,\beta) =$
$\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\int\left(\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{k}\prod_{j=1}^{V}(\theta_i\beta_{ij})^{w_n^j}\right)d\theta$
- Unfortunately, exact inference is intractable due to the coupling between θ and β in the summation over latent topics

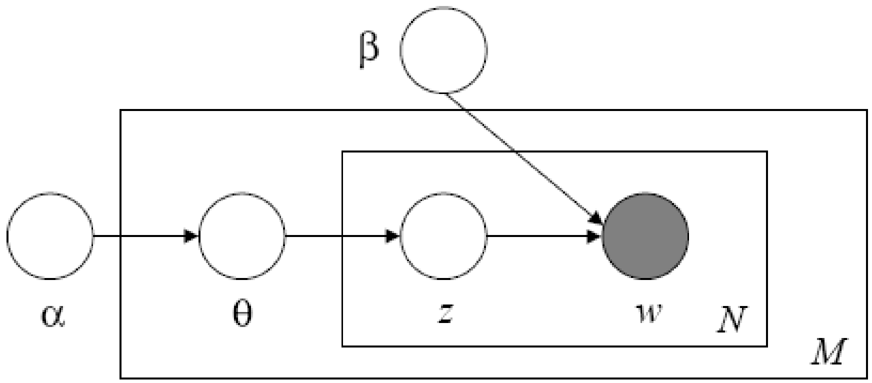
- So we turn to alternatives
  - Variational Inference
  - Markov Chain Monte Carlo (Gibbs sampling)
  - Kalman Filtering

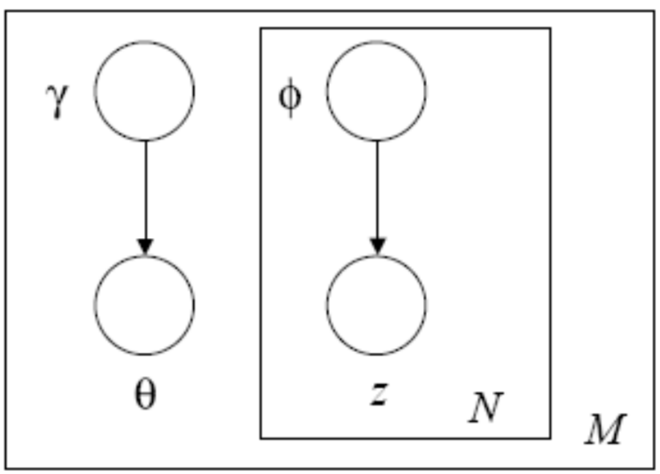
Variational Inference

- In variational inference, P(Z|X) is estimated using $P(Z|X) \approx Q(Z)$
- $Q(Z)$ is restricted to belong to a family of distributions of simpler form than $P(Z|X)$ with the intention of having Q(Z) similar to P(Z|X)
- The idea is to obtain a tractable lower bound on log likelihood.
- A simple way to obtain a tractable family of lower bound is to consider simple modifications of the original graph model in which some of the edges and nodes are removed.

Inference and parameter estimation

- Drop some edges and the **w** nodes





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

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

- Distance between Q and P should be as small as possible.

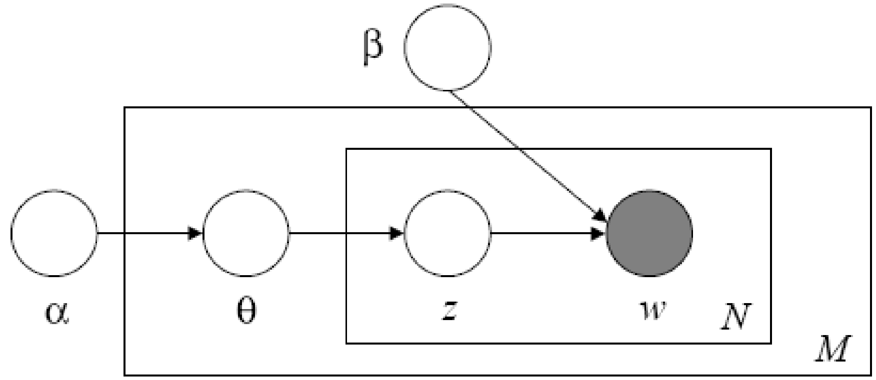
$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} KL(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta))$$

- KL divergence between Q and P is
  - $D_{KL}(Q||P) = \sum_Z Q(Z) log \frac{Q(Z)}{P(Z|X)} = \sum_Z Q(Z) log \frac{Q(Z)}{P(Z,X)} + \log P(X)$
- In other words, $\log p(\boldsymbol{w}|\alpha, \beta) = E_q[\log p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)] - E_q[\log q(\theta, z|\gamma, \phi)] + D(q(\theta, \boldsymbol{z}|\gamma, \phi)||p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta))$
- Now, with fixed α and β from the previous EM iteration, log $p(\boldsymbol{w}|\alpha, \beta)$ is constant.
- Let $\mathrm{L}(\gamma, \phi; \alpha, \beta) = E_q[\log p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)] - E_q[\log q(\theta, z|\gamma, \phi)]$
- Then, to make minimize the KL divergence, one needs to maximize $\mathrm{L}(\gamma, \phi; \alpha, \beta)$

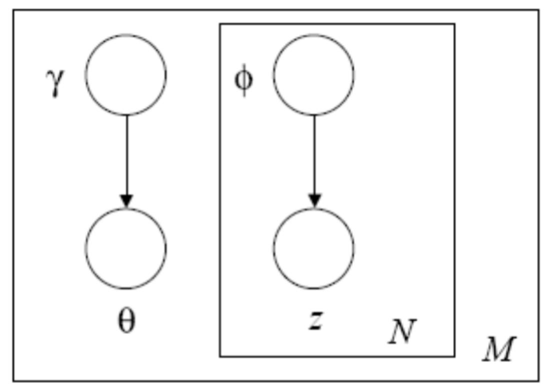
- Thus, we need to find variational parameters γ and ϕ such that $\mathrm{L}(\gamma, \phi;$ is maximized.
- $\mathrm{L}(\gamma, \phi; \alpha, \beta) = E_q[\log p(\theta, \mathbf{z}, \boldsymbol{w} | \alpha, \beta)]$
  $E_q[\log q(\theta, z | \gamma, \phi)]$

$$
\begin{aligned}
&= E_q[\log p(\theta|\alpha)] \\
&+ E_q[\log p(z|\theta)] \\
&+ E_q[\log p(w|z, \beta)] \\
&- E_q[\log q(\theta|\gamma)] \\
&- E_q[\log q(z|\phi)]
\end{aligned}
$$

- Next we will compute each of the above 5 terms
- But before that let us study computation of first moments of exponentially family of distributions



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



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

Exponential Family of Distributions

Exponential family

- An exponential family distribution has the form

$$p(x|\eta) = h(x)\exp\{\eta^T t(x) - a(\eta)\}$$

- The different parts of this equation are
  - The natural parameter η
  - The sufficient statistic $t(x)$
  - The underlying measure $h(x)$
  - The log normalizer $a(\eta)$

$$a(\eta) = \log \int h(x)\exp\{\eta^T t(x)\}$$

First Moment

First Moment

- The derivatives of the log normalizer gives the moments of the sufficient statistics

$$
\begin{aligned}
\frac{d}{d\eta}a(\eta) &= \frac{d}{d\eta}(\log \int \exp\{\eta^T t(x)\}h(x)dx) \\
&= \frac{\int t(x)\exp\{\eta^T t(x)\}h(x)dx}{\int \exp\{\eta^T t(x)\}h(x)dx} \\
&= \int t(x)\exp\{\eta^T t(x) - a(\eta)\}h(x)dx \\
&= E[t(X)]
\end{aligned}
$$

How to compute expectation of conditional logs?

Computing $E[\log(\theta|\alpha)]$

- The Dirichlet distribution $p(\theta|\alpha)$:

$$
\begin{aligned}
p(\theta|\alpha) &= \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\sum_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \\
&= \exp\{(\sum_{i=1}^{K} (\alpha_i - 1) \log \theta_i) + \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i)\}
\end{aligned}
$$

- Sufficient statistics: $\log \theta_i$.
- Log normalizer: $\sum_{i=1}^{K} \log \Gamma(\alpha_i) - \log \Gamma(\sum_{i=1}^{K} \alpha_i)$

How to compute expectation of conditional logs?

- The expectation $E[\log(\theta|\alpha)]$ is:

$$
\begin{aligned}
E[\log \theta_i|\alpha] &= a(\alpha)' = (\sum_{i=1}^{K} \log \Gamma(\alpha_i) - \log \Gamma(\sum_{i=1}^{K} \alpha_i))' \\
&= \psi(\alpha_i) - \psi(\sum_{j=1}^{K} \alpha_j).
\end{aligned}
$$

where ψ is the digamma function, the first derivative of the log Gamma function.

Computing the Five Terms

Computing $E_q[\log p(\theta|\alpha)]$

- $E_q[\log p(\theta|\alpha)]$ is given by

$$\begin{aligned} E_q[\log p(\theta|\alpha)] &= \sum_{i=1}^{K}(\alpha_i - 1)E_q[\log \theta_i] \\ &+ \log \Gamma(\sum_{i=1}^{K}\alpha_i) - \sum_{i=1}^{K}\log \Gamma(\alpha_i). \end{aligned}$$

- θ is generated by $Dir(\theta|\gamma)$: $E_q[\log \theta_i] = \psi(\gamma_i) - \psi(\sum_{j=1}^{K}\gamma_j)$.
- Then we have:

$$\begin{aligned} E_q[\log p(\theta|\alpha)] &= \sum_{i=1}^{K}(\alpha_i - 1)\psi(\gamma_i) - \psi(\sum_{j=1}^{K}\gamma_j) \\ &+ \log \Gamma(\sum_{i=1}^{K}\alpha_i) - \sum_{i=1}^{K}\log \Gamma(\alpha_i). \end{aligned}$$

Computing the Five Terms

Computing $E_q[\log p(z|\theta)]$

$E_q[\log p(z|\theta)]$ is given by

$$
\begin{aligned}
E_q[\log p(z|\theta)] &= E_q[\sum_{n=1}^{N}\sum_{i=1}^{K} z_{ni}\log\theta_i] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{K} E_q[z_{ni}]E_q[\log\theta_i] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k} \phi_{ni}(\psi(\gamma_i) - \psi(\sum_{j=1}^{K}\gamma_j))
\end{aligned}
$$

where z is generated from $Mult(z|\phi)$ and θ is generated from $Dir(\theta|\gamma)$.

Computing the Five Terms

Computing $E_q[\log p(w|z,\beta)]$

$E_q[\log p(w|z,\beta)]$ is given by

$$
\begin{aligned}
E_q[\log p(w|z,\beta)] &= E_q[\sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} z_{ni} w_n^j \log \beta_{ij}] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} E_q[z_{ni}] w_n^j \log \beta_{ij} \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V} \phi_{ni} w_n^j \log \beta_{ij}
\end{aligned}
$$

Computing the Five Terms

Computing $E_q[\log q(\theta|\gamma)]$

$E_q[\log q(\theta|\gamma)]$ is given by

$$E_q[\log p(\theta|\gamma)] = \sum_{i=1}^{k}(\gamma_i - 1)E_q[\log \theta_i] + \log \Gamma(\sum_{i=1}^{k} \gamma_i) - \sum_{i=1}^{k} \log \Gamma(\gamma_i)$$

Then, we have

$$\begin{aligned} E_q[\log p(\theta|\gamma)] &= \log \Gamma(\sum_{i=1}^{k} \gamma_i) - \sum_{i=1}^{k} \log \Gamma(\gamma_i) \\ &+ \sum_{i=1}^{k}(\gamma_i - 1)(\psi(\gamma_i) - \psi(\sum_{j=1}^{k} \gamma_j)) \end{aligned}$$

Computing the Five Terms

Computing $E_q[\log q(z|\phi)]$

$E_q[\log q(z|\phi)]$ is given by

$$
\begin{aligned}
E_q[\log q(z|\phi)] &= E_q[\sum_{n=1}^{N}\sum_{i=1}^{k} z_{ni} \log \phi_{ni}] \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k} E_q[z_{ni}] \log \phi_{ni} \\
&= \sum_{n=1}^{N}\sum_{i=1}^{k} \phi_{ni} \log \phi_{ni}
\end{aligned}
$$

Variational Inference

Finally, $L(\gamma, \phi; \alpha, \beta)$ is

$$
\begin{aligned}
(\gamma, \phi; \alpha, \beta) \;&=\; \log \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) \\
&+ \sum_{i=1}^{K} (\alpha_i - 1)(\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j)) \\
&+ \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{ni}(\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j)) \\
&+ \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{ni} w_n^j \log \beta_{ij} \\
&- (\log \Gamma(\sum_{i=1}^{K} \gamma_i) - \sum_{i=1}^{K} \log \Gamma(\gamma_i) + \sum_{i=1}^{K} (\gamma_i - 1)(\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j)) \\
&- \sum^{N} \sum^{K} \phi_{ni} \log \phi_{ni}.
\end{aligned}
$$

Variational Multinomial

- Maximize $L(\gamma, \phi; \alpha, \beta)$ with respect to ϕ_{ni}:

$$\begin{aligned} L_{\phi_{ni}} &= \phi_{ni}(\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j)) + \phi_{ni} \log \beta_{iv} \\ &- \phi_{ni} \log \phi_{ni} + \lambda(\sum_{j=1}^{K} \phi_{ni} - 1). \end{aligned}$$

- Taking derivatives with respect to ϕ_{ni}:

$$\frac{\partial L}{\partial \phi_{ni}} = (\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j)) + \log \beta_{iv} - \log \phi_{ni} - 1 + \lambda.$$

- Setting this derivative to zero yields

$$\phi_{ni} \propto \beta_{iv} \exp(\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j)).$$

Variational Dirichlet

Maximize $L(\gamma, \phi; \alpha, \beta)$ with respect to γ_i:

$$
\begin{aligned}
L_{\gamma} &= \sum_{i=1}^{K} (\psi(\gamma_i) - \psi(\sum_{j=1}^{K} \gamma_j))(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) \\
&- \log \Gamma(\sum_{j=1}^{K} \gamma_j) + \sum_{i=1}^{K} \log \Gamma(\gamma_i)
\end{aligned}
$$

- Taking the derivative with respect to γ_i

$$
\frac{\partial L}{\partial \gamma_i} = \psi'(\gamma_i)(\alpha_i + \sum_{n=1}^{N} \phi_{ni} - \gamma_i) - \psi''(\sum_{j=1}^{K} \gamma_j) \sum_{j=1}^{K} (\alpha_j + \sum_{n=1}^{N} \phi_{nj} - \gamma_j)
$$

- Setting this equation to zero yields:

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

Variational Inference Algorithm

1. initialize $\phi^0_{ni} = \frac{1}{K}$ for all i and n.
2. initialize $\gamma_i = \alpha_i + \frac{N}{K}$ for all i
3. **repeat**
4. **for** $n = 1$ to N
5. **for** $i = 1$ to K
    1. $\phi^{t+1}_{ni} = \beta_{iw_n} \exp(\psi(\gamma^t_i))$.
    2. normalize ϕ^{t+1}_n to sum 1.
6. $\gamma^{t+1} = \alpha + \sum_{n=1}^{N} \phi^{t+1}_n$
7. **until** convergence

Parameter Estimation

- In the variational E-step, maximize the lower bound $L(\gamma, \phi; \alpha, \beta)$ with respect to the variational parameters γ and ϕ.
- In the M-step, maximize the bound with respect to the model parameters α and β.

M-Step: Maximize the lower bound on the log likelihood of

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

Conditional Multinomials

- Maximize $L(\gamma, \phi; \alpha, \beta)$ with respect to β:

$$L_\beta = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{dni} w_{dn}^{j} \log \beta_{ij} + \sum_{i=1}^{K} \lambda_i \Big(\sum_{j=1}^{V} \beta_{ij} - 1\Big).$$

- Taking the derivative with respect to β_{ij} and setting it to zero:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}.$$

Dirichlet

- Maximize $L(\gamma, \phi; \alpha, \beta)$ with respect to α:

$$
\begin{aligned}
L_\alpha &= \sum_{d=1}^{M} (\log \Gamma(\sum_{j=1}^{K} \alpha_j) - \sum_{i=1}^{K} \log \Gamma(\alpha_i)) \\
&+ \sum_{i=1}^{K} ((\alpha_i - 1)(\psi(\gamma_{di}) - \psi(\sum_{j=1}^{K} \gamma_{dj})))
\end{aligned}
$$

- Taking the derivative with respect to α_i

$$
\frac{\partial L}{\partial \alpha_i} = M(\psi(\sum_{j=1}^{K} \alpha_j) - \psi(\alpha_i)) + \sum_{d=1}^{M} (\psi(\gamma_{di}) - \psi(\sum_{j=1}^{K} \gamma_{dj})).
$$

- It is difficult to compute α_i by setting the derivative to zero.

This derivative depends on α_j, where $j \neq i$, and we therefore must use an iterative method

Newton Raphson Method

- Compute the Hessian Matrix by

$$\frac{\partial^2 L}{\partial \alpha_i \partial \alpha_j} = M(\psi'(\sum_{j=1}^{K} \alpha_j) - \delta(i,j)\psi'(\alpha_i)).$$

- Input this Hessian Matrix and the derivative to Newton Method.

The Newton-Raphson optimization technique finds a stationary point of a function by iterating:

$$\alpha_{\text{new}} = \alpha_{\text{old}} - H(\alpha_{\text{old}})^{-1} g(\alpha_{\text{old}})$$

where $H(\alpha)$ and $g(\alpha)$ are the Hessian matrix and gradient respectively at the point α. In general, this algorithm scales as $O(N^3)$ due to the matrix inversion.

Efficient Newton Raphson

If the Hessian matrix is of the form:

$$H = \operatorname{diag}(h) + \mathbf{1}z\mathbf{1}^{\mathrm{T}}, \tag{10}$$

where $\operatorname{diag}(h)$ is defined to be a diagonal matrix with the elements of the vector h along the diagonal, then we can apply the matrix inversion lemma and obtain:

$$H^{-1} = \operatorname{diag}(h)^{-1} - \frac{\operatorname{diag}(h)^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\operatorname{diag}(h)^{-1}}{z^{-1} + \sum_{j=1}^{k} h_j^{-1}}$$

Multiplying by the gradient, we obtain the ith component:

$$(H^{-1}g)_i = \frac{g_i - c}{h_i}$$

where

$$c = \frac{\sum_{j=1}^{k} g_j/h_j}{z^{-1} + \sum_{j=1}^{k} h_j^{-1}}.$$

Observe that this expression depends only on the $2k$ values h_i and g_i and thus yields a Newton-Raphson algorithm that has linear time complexity.

Variational EM LDA Algorithm

Algorithm 1: Variational Expectation-Maximization LDA

Input: Number of topics K

Corpus with M documents and N_d words in document d

Output: Model parameters: β, θ, z

initialize $\phi^0_{ni} := 1/k$ for all i in k and n in N_d

initialize $\gamma_i := \alpha_i + N/k$ for all i in k

initialize $\alpha := 50/k$

initialize $\beta_{ij} := 0$ for all i in k and j in V

- E step
- M step
- if loglikelihood converged then
  - return parameters
- else
  - go back to E-step
- endif

Variational EM LDA Algorithm

//E-Step (determine ϕ and γ and compute expected likelihood)
loglikelihood := 0
for $d = 1$ **to** M
  repeat
    for $n = 1$ **to** N_d
      for $i = 1$ **to** K
        $\phi_{dni}^{t+1} := \beta_{iw_n} \exp\left(\Psi(\gamma_{di}^{t})\right)$
      endfor
      normalize ϕ_{dni}^{t+1} to sum to 1
    endfor
    $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_{dn}^{t+1}$
  until convergence of ϕ_d and γ_d
  loglikelihood := loglikelihood + $L(\gamma, \phi; \alpha, \beta)$ // See equation BLAH
endfor

Variational EM LDA Algorithm

//M-Step (maximize the log likelihood of the variational distribution)
for $d = 1$ **to M**
  for $i = 1$ **to K**
    for $j = 1$ **to V**
      $\beta_{ij} := \phi_{dni} w_{dnj}$
    endfor
    normalize β_i to sum to 1
  endfor
endfor
estimate α via Eq. (8)

Gibbs Sampling for LDA Derivation

Gibbs Sampling for LDA

- Gibbs sampling procedure is to estimate

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w})$$

$$\begin{aligned}
P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) &\propto P(z_i = j, \mathbf{z}_{-i}, \mathbf{w}) \\
&= P(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}_{-i}) \\
&= P(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) P(z_i = j | \mathbf{z}_{-i})
\end{aligned}$$

$$\begin{aligned}
& P(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \\
= & \int P(w_i | z_i = j, \phi^{(j)}) P(\phi^{(j)} | \mathbf{z}_{-i}, \mathbf{w}_{-i}) d\phi^{(j)} \\
= & \int \phi_{w_i}^{(j)} P(\phi^{(j)} | \mathbf{z}_{-i}, \mathbf{w}_{-i}) d\phi^{(j)}
\end{aligned}$$

$$\begin{aligned}
P(\phi^{(j)} | \mathbf{z}_{-i}, \mathbf{w}_{-i}) &\propto P(\mathbf{w}_{-i} | \phi^{(j)}, z_{-i}) P(\phi^j) \\
&\sim Dirichlet(\beta + n_{-i,j}^{(w)})
\end{aligned}$$

Gibbs Sampling for LDA

- By the property of the expectation of Dirichlet

$n_{-i,j}^{(w)}$ is the number of instances of word w assigned to topic j.

$n_{-i,j}$ total number of words assigned to topic j.

$$P(w_i | z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta}$$

Gibbs Sampling for LDA

Similarly, for the 2nd term, we have

$$\begin{aligned} P(z_i = j|\mathbf{z}_{-i}) &= \int P(z_i = j|\theta^{(d)})P(\theta^{(d)}|\mathbf{z}_{-i})d\theta^{(d)} \\ P(\theta^{(d)}|\mathbf{z}_{-i}) &\propto P(\mathbf{z}_{-i}|\theta^{(d)})P(\theta^{(d)}) \\ &\sim Dirichlet(n^{(d)}_{-i,j} + \alpha) \end{aligned}$$

where $n^{(d)}_{-i,j}$ is the number of words assigned to topic j excluding current one.

$$P(z_i = j|z_{-i}) = \frac{n^{(d)}_{-i,j} + \alpha}{n^{(d)}_{-i,\cdot} + K\alpha}$$

where $n^{(d)}_{-i,\cdot}$ is the total number of topics assigned to document d excluding current one.

Gibbs Sampling for LDA

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + W\beta} \frac{n^{(d)}_{-i,j} + \alpha}{n^{(d)}_{-i,\cdot} + K\alpha}$$

Need to record four count variables:

- document-topic count $n^{(d)}_{-i,j}$
- document-topic sum $n^{(d)}_{-i,\cdot}$ (actually a constant)
- topic-term count $n^{(w_i)}_{-i,j}$
- topic-term sum $n^{(\cdot)}_{-i,j}$

Parameter Estimation

To obtain ϕ, and θ, two ways, (draw one sample of z or draw multiple samples of z to calculate the average)

$$\phi_{j,w} = \frac{n_w^{(j)} + \beta}{\sum_{w=1}^{V} n_w^{(j)} + V\beta}$$

$$\theta_j^{(d)} = \frac{n_j^{(d)} + \alpha}{\sum_{z=1}^{K} n_z^{(d)} + K\alpha}$$

where $n_w^{(j)}$ is the freqency of word assigned to topic j, and $n_z^{(d)}$ is the number of words assigned to topic z.

Compared with VB, Gibbs Sampling is easy to implement.
Easy to extend.
More efficient. Faster to obtain good approximation.

Other Topic Models: Document Relations

- In base model (LDA) documents are exchangeable (document exchangeability assumption)
- By removing this assumption, we can build more complex model
- More complex model -> New (more specific) applications
- Two types of document relations:
  a) Sequential (time)
  b) Networked (links, citations, references…)

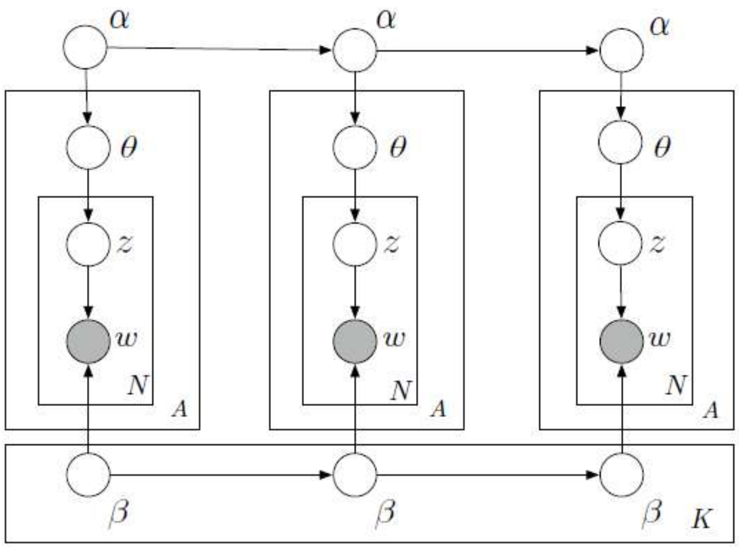
Other Topic Models

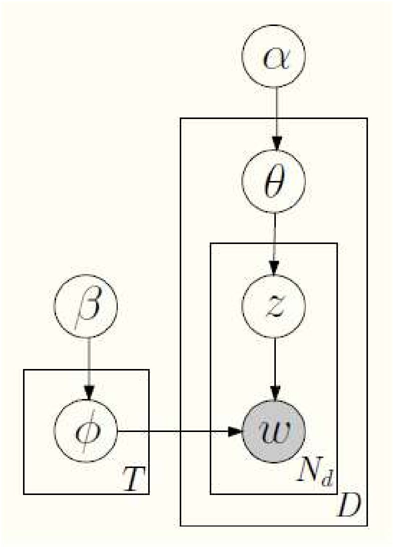
- Modeling time: topic detection and tracking
  - Trend detection:
    What was popular? What will be popular?
  - Event detection:
    Something important has happened
  - Topic tracking:
    Evolution of a specific topic

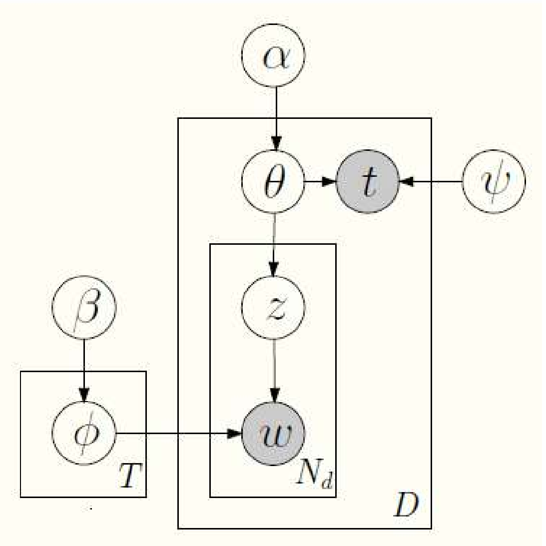
Other Topic Models

- Modeling time: two approaches
  - Markov dependency
    - Short-distance
    - Dynamic Topic Model
  - Time as additional feature
    - Long-distance
    - Topics-Over-Time

Other Topic Models







Other Topic Models

- Modeling document networks
  - Web (documents with hyperlinks)
  - Messages (documents with senders and recipients)
  - Scientific papers (documents and citations)

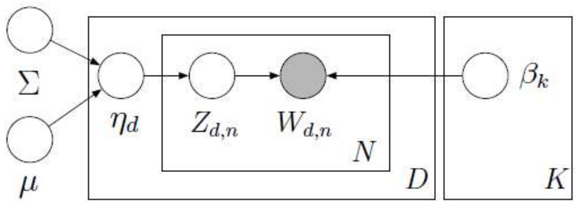
Other Topic Models: Topic Relations

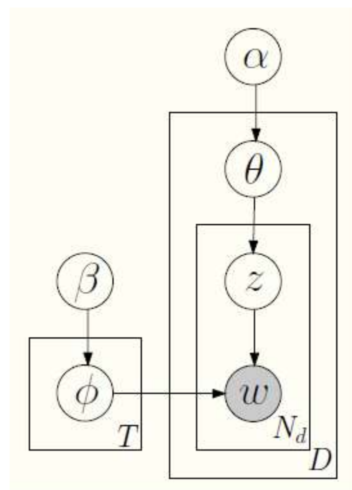
- In base model (LDA) topics are "exchangeable" (topic exchangeability assumption)
- By removing this assumption, we can build more complex model
- More complex model -> New (more specific) applications
- Two types of topic relations:
  a) Correlations (topic hierarchy, similarity,...)
  b) Sequence (linear structure of text)

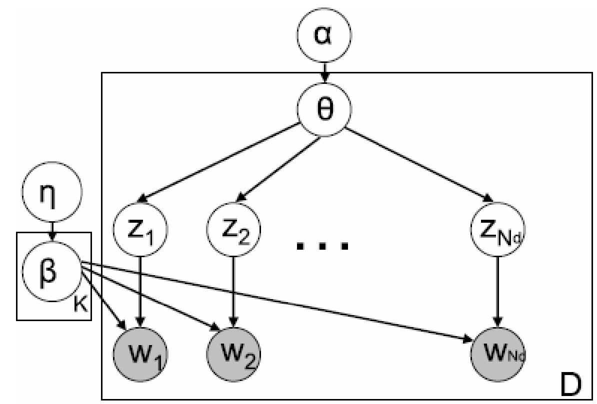
Other Topic Models

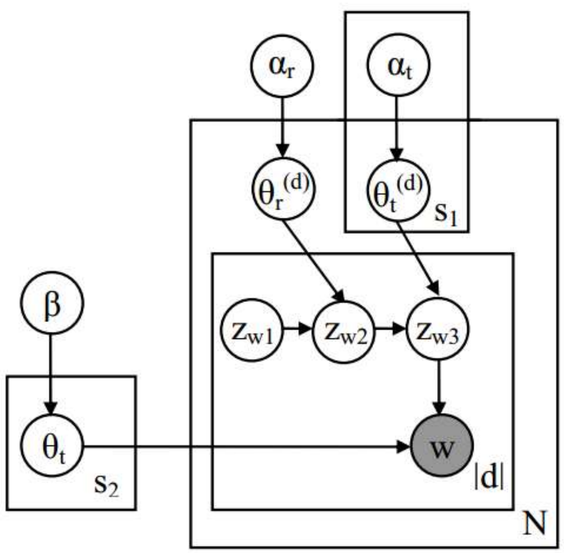
- Topic correlations:
  – Instead of finding “flat” topic structure:
    - Topic hierarchy: super-topics and sub-topics
    - Topic correlation matrix
    - Arbitrary DAG structure
- Topic sequence:
  – Sequential nature of the human language:
    - Text is written from beginning to the end
    - Topics in latter chapters of the text tend to depend on previous
    - Markov property

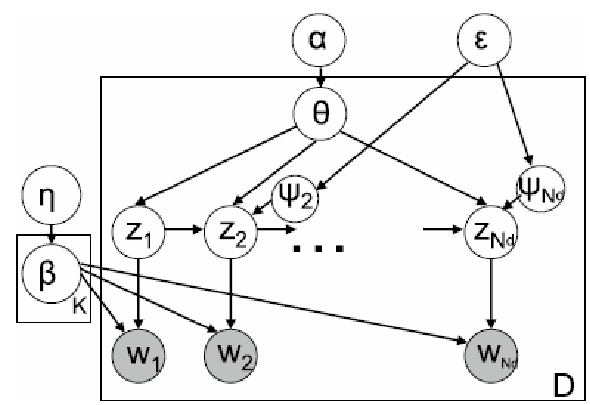
Other Topic Models











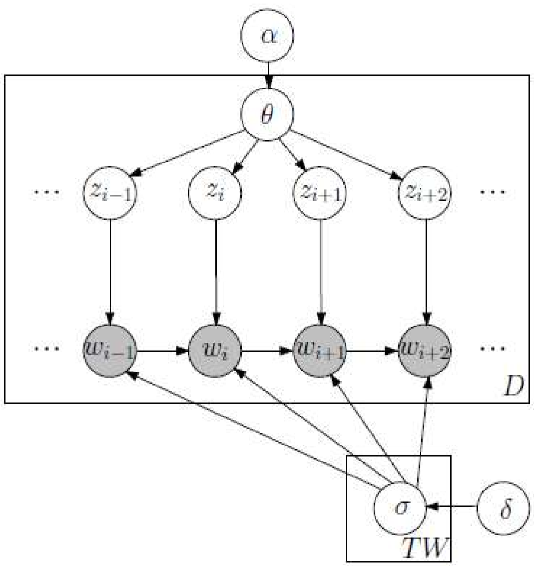
Other Topic Models: Word Relations

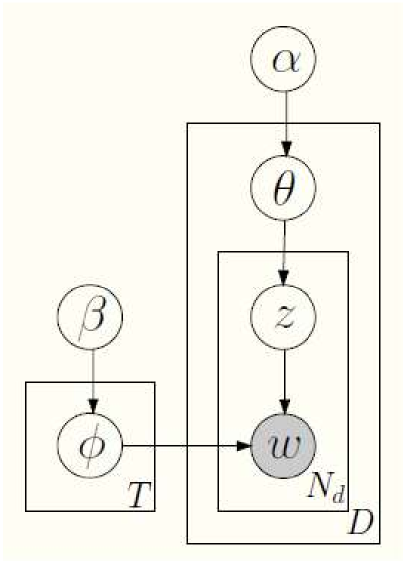
- In base model (LDA) words are “exchangeable” (word exchangeability assumption)
- By removing this assumption,
  we can build more complex model
- More complex model -> New (more specific) applications
- Two types of word relations:
  a) Intra-document (word sequence)
  b) Inter-document (entity recognition, multilinguality…)

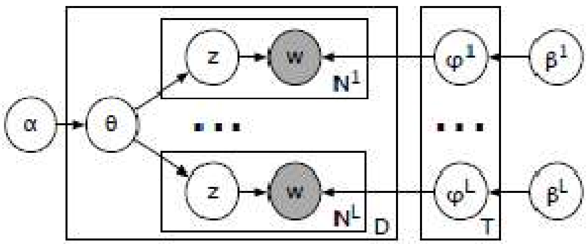
Other Topic Models

- Intra-document word relations:
  - Sequential nature of text:
    - Modeling phrases and n-grams
    - Markov property
- Inter-document word relations:
  - Some words can be treated as special entities
    - Not sufficiently investigated
  - Multilingual models
    - Harnessing multiple languages
    - Bridging the language gap

Other Topic Models







Other Topic Models

- Relieving the aforementioned exchangeability assumptions is not the only way to extend the LDA model to new problems and more complex domains

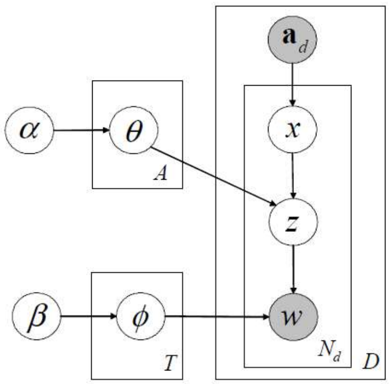
- Extension can be made by utilizing additional features on any of the three levels (document, topic, word)

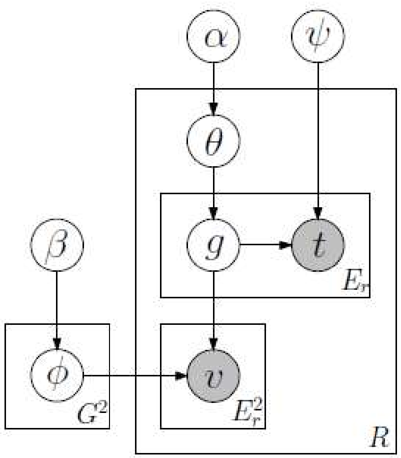
- Combining different features from different domains can solve new compound problems (eg. time-evolution of topic hierarchies)

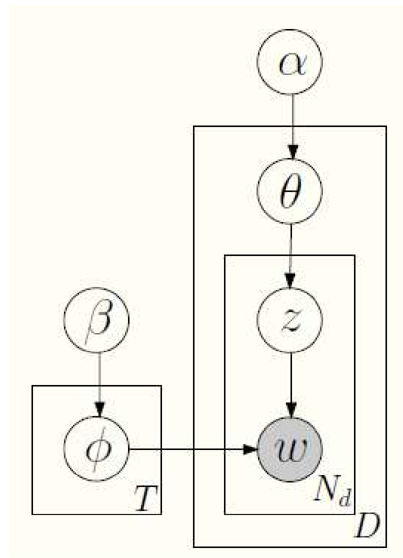
Other Topic Models

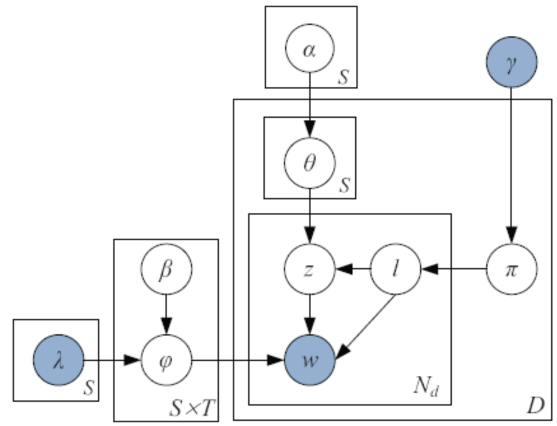
- Examples of models
  with additional features on document level:
  - Author topic models
  - Group topic models
  - Sentiment topic models
  - Opinion topic models

Other Topic Models



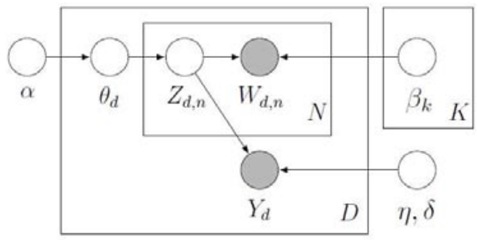


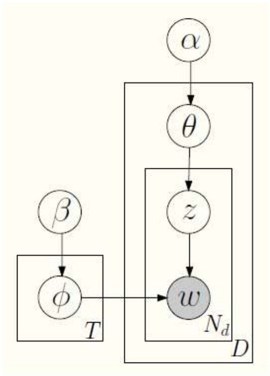


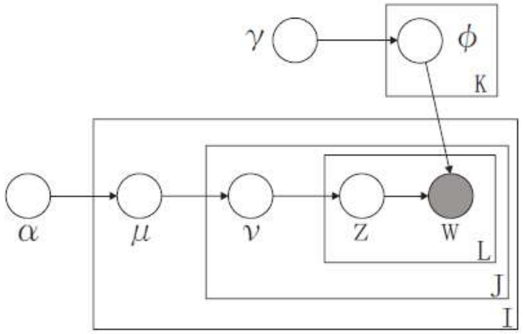


Other Topic Models

- Examples of models
  with additional features on topic level:
  - Supervised topic models
  - Segmentation topic models

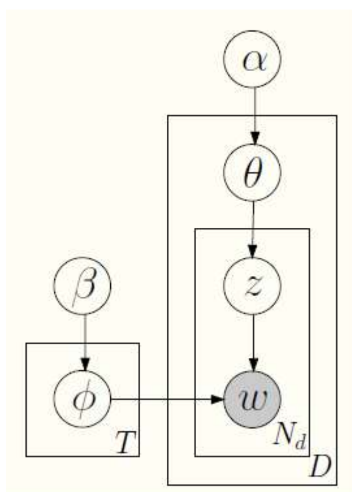


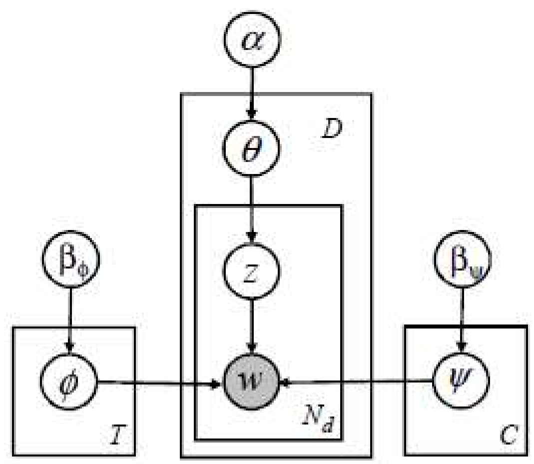




Other Topic Models

- Examples of models
  with additional features on word level:
  - Concept topic models
  - Entity disambiguation topic models

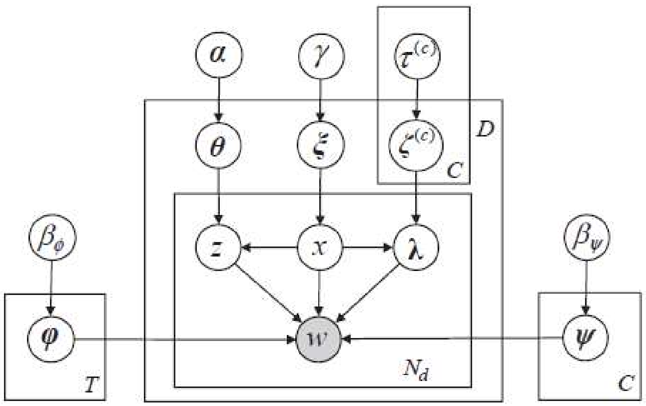


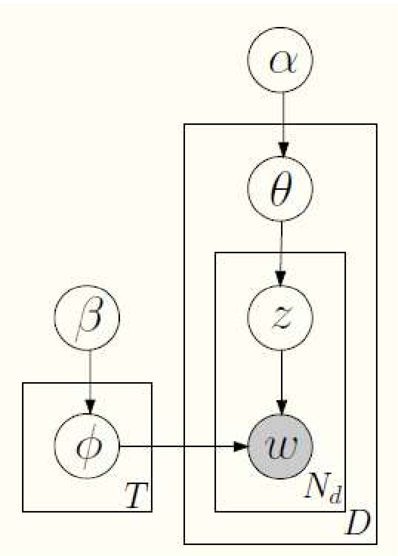


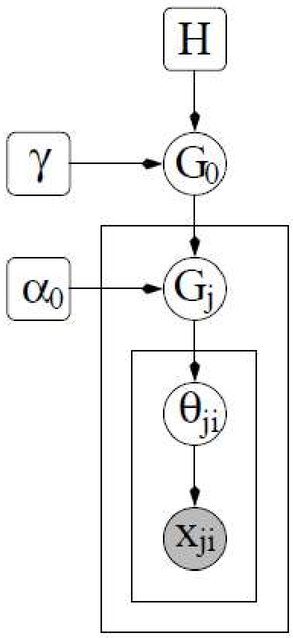
Other Topic Models

- Using simple additional features sometimes is not enough:
  - How to implement knowledge?
    - Complex set of features (with their dependencies)
    - Markov logic networks?
    - Incorporate knowledge through priors
    - Room for improvement!
- Number of parameters is often not known in advance:
  - How many topics are there in a corpus?
    - Solution: non-parametric distributions
    - Dirichlet process (Chinese restaurant process, Stick-breaking process, Pitman-Yor process, Indian buffet process….)

Other Topic Models







General Bibliography

- http://www.cs.princeton.edu/~mimno/topics.html
- Andrew M. Dai, Amos J. Storkey. The Grouped Author-Topic Model for Unsupervised Entity Resolution . ICANN (2011).
- Chaitanya Chemudugunta, Padhraic Smyth, Mark Steyvers. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. NIPS (2006).
- Chang Wang, James Fan, Aditya Kalyanpur, David Gondek. Relation Extraction with Relation Topics. EMNLP (2011).
- Chang Wang, Sridhar Mahadevan. Multiscale Analysis of Document Corpora Based on Diffusion Models. IJCAI (2009).
- Claudio Taranto, Nicola Di Mauro, Floriana Esposito. rsLDA: a Bayesian Hierarchical Model for Relational Learning. ICDKE (2011).
- D. Newman, S. Block. Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper. JASIST () 2006 pp. .
- Daniel Ramage, David Hall, Ramesh Nallapati, Christopher D. Manning. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora. EMNLP (2009).
- Daniel Ramage, Susan Dumais, Dan Liebling. Characterizing Microblogs with Topic Models. ICWSM (2010).
- David Andrzejewski, Anne Mulhern, Ben Liblit, Xiaojin Zhu. Statistical Debugging using Latent Topic Models. ECML (2007).
- David Andrzejewski, Xiaojin Zhu, Mark Craven, Ben Recht. A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation using First-Order Logic. IJCAI (2011).
- David Andrzejewski, Xiaojin Zhu, Mark Craven. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. ICML (2009).

General Bibliography

- David Blei, Michael Jordan. Modeling Annotated Data. SIGIR (2003).
- David M. Blei, Andrew Ng, Michael Jordan. Latent Dirichlet allocation. JMLR (3) 2003 pp. 993-1022.
- David M. Blei, John D. Lafferty. A Correlated Topic model of Science. AAS (1) 2007 pp. 17-35.
- David M. Blei, Jon D. McAuliffe. Supervised Topic Models. NIPS (2007).
- David Mimno, Andrew McCallum. Expertise Modeling for Matching Papers with Reviewers. KDD (2007).
- David Mimno, Andrew McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. UAI (2008).
- David Mimno, Wei Li, Andrew McCallum. Mixtures of Hierarchical Topics with Pachinko Allocation. ICML (2007).
- David Newman, Chaitanya Chemudugunta, Padhraic Smyth. Statistical entity-topic models. KDD (2006).
- Feng Yan, Ningyi Xu, Yuan Qi. Parallel Inference for Latent Dirichlet Allocation on Graphics Processing Units. NIPS (2009).
- Gabriel Doyle, Charles Elkan. Accounting for Burstiness in Topic Models. ICML (2009).
- Hal Daumé III. Markov Random Topic Fields. (2009).
- Hanna M. Wallach. Topic modeling: beyond bag-of-words. ICML (2006).

General Bibliography

- Jacob Eisenstein, Amr Ahmed, Eric P. Xing. Sparse Additive Generative Models of Text. ICML (2011).
- Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, Eric P. Xing. A Latent Variable Model for Geographic Lexical Variation. EMNLP (2010).
- Joseph Reisinger, Austin Waters, Brian Silverthorn, Raymond J. Mooney. Spherical Topic Models. ICML (2010).
- Jukka Perkiö, Wray L. Buntine, Sami Perttu. Exploring Independent Trends in a Topic-Based Search Engine. Web Intelligence (2004).
- Jun Zhu, Amr Ahmed, Eric P. Xing. MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification. ICML (2009).
- Jun Zhu, Eric P. Xing. Conditional Topic Random Fields. ICML (2010).
- Matthew Hoffman, David M. Blei, Francis Bach. Online Learning for Latent Dirichlet Allocation. NIPS (2010).
- Matthew Purver, Konrad Körding, Thomas L. Griffiths, Joshua Tenenbaum. Unsupervised Topic Modelling for Multi-Party Spoken Discourse. ACL (2006).
- Michal Rosen-Zvi, Tom Griffiths, Mark Steyvers, Padhraic Smyth. The Author-Topic Model for Authors and Documents. UAI (2004).
- Pradipto Das, Rohini Srihari, Yun Fu. Simultaneous Joint and Conditional Modeling of Documents Tagged from Two Perspectives. (2011).
- Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. WWW (2007).
- Ruslan Salakhutdinov, Geoffrey Hinton. Replicated Softmax: an Undirected Topic Model. NIPS (2009).

General Bibliography

- Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. Indexing by Latent Semantic Analysis. JASIS (41) 1990 pp. 391-407.
- Shuang-Hong Yang, Steven P. Crain, Hongyuan Zha. Bridging the language gap: topic adaptation for documents with different technicality. AIStats (2011).
- Simon Lacoste-Julien, Fei Sha, Michael I. Jordan. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. NIPS (2008).
- Thomas Hofmann. Probilistic latent semantic analysis. UAI (1999).
- Thomas K. Landauer, Susan T. Dumais. Solutions to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review. 1997
- Thomas L. Griffiths, Mark Steyvers. Finding Scientific Topics. PNAS (101) 2004 pp. 5228-5235.
- Wei Li, Andrew McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. ICML (2006).
- Wei-Hao Lin, Eric P. Xing, Alexander Hauptmann. A Joint Topic and Perspective Model for Ideological Discourse. ECML PKDD (2008).
- Wray L. Buntine, Aleks Jakulin. Discrete Component Analysis. SLSFS (2005).
- Xiaojin Zhu, David M. Blei, John Lafferty. TagLDA: Bringing document structure knowledge into topic models. (2006).
- Xing Wei, Bruce Croft. LDA-based document models for ad-hoc retrieval. SIGIR (2006).

Topic Models: Bibliometrics and Cross-Language

- Bibliometrics
  - David Hall, Daniel Jurafsky, Christopher D. Manning. Studying the History of Ideas Using Topic Models. EMNLP (2008).
  - David Mimno, Andrew McCallum. Mining a digital library for influential authors. JCDL (2007).
  - Elena Erosheva, Stephen Fienberg, John Lafferty. Mixed Membership Models of Scientific Publications. PNAS (101) 2004 pp. 5220-5227.
  - Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences. ICML (2007).
  - Sean Gerrish, David M. Blei. A language-based approach to measuring scholarly impact. ICML (2010).
- Cross-language
  - Bin Zhao, Eric P. Xing. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation. NIPS (2007).
  - Bing Zhao, Eric P. Xing. BiTAM: Bilingual Topic AdMixture Models for Word Alignment. ACL (2006).
  - David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, Andrew McCallum. Polylingual Topic Models. EMNLP (2009).
  - David Mimno. Reconstructing Pompeian Households. UAI (2011).
  - Jagadeesh Jagarlamudi, Hal Daumé III. Extracting Multilingual Topics from Unaligned Comparable Corpora. (2010).
  - Jordan Boyd-Graber, David M. Blei. Multilingual Topic Models for Unaligned Text. UAI (2009).
  - Xiaochuan Ni, Jian-Tao Sun, Jian Hu, Zheng Chen. Mining Multilingual Topics from Wikipedia. WWW (2009).

Topic Models: Evaluation

- Evaluation
  - Claudiu Musat, Julien Velcin, Stefan Trausan-Matu, Marian-Andrei Rizoiu. Improving Topic Evaluation Using Conceptual Knowledge. IJCAI (2011).
  - David Mimno, David Blei. Bayesian Checking for Topic Models. EMNLP (2011).
  - David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, Andrew McCallum. Optimizing Semantic Coherence in Topic Models. EMNLP (2011).
  - David Newman, Jey Han Lau, Karl Grieser, Timothy Baldwin. Automatic Evaluation of Topic Coherence. NAACL (2010).
  - Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, David Mimno. Evaluation Methods for Topic Models. ICML (2009).
  - Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. NIPS (2009).
  - Loulwah AlSumait, Daniel Barbará, James Gentle, Carlotta Domeniconi. Topic Significance Ranking of LDA Generative Models. ECML (2009).
  - Wray L. Buntine. Estimating Likelihoods for Topic Models. Asian Conference on Machine Learning (2009).

Topic Models: Implementations

- Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit. (2002).
- Brad Block. Collapsed variational HDP. (2011).
- Daniel Ramage, Evan Rosen. Stanford Topic Modeling Toolbox. (2009).
- David M. Blei. lda-c. (2003).
- Gregor Heinrich. Infinite LDA. (2011).
- John Langford. Vowpal Wabbit. (2011).
- Jonathan Chang. R package 'lda'. (2011).
- Mark Steyvers, Tom Griffiths. Matlab Topic Modeling Toolbox. (2005).
- Radim Rehurek. gensim. (2009).
- Ramesh Nallapati. multithreaded lda-c. (2010).
- Shravan Narayanamurthy. Yahoo! LDA. (2011).
- Wray L. Buntine. Discrete Component Analysis. (2009).
- Xuan-Hieu Phan, Cam-Tu Nguyen. GibbsLDA++. (2007).

Topic Models: Inference, Vision, User Interface, Introductory Tutorials

- Inference
  - Arthur Asuncion, Max Welling, Padhraic Smyth, Yee-Whye Teh. On Smoothing and Inference for Topic Models. UAI (2009).
  - Gregor Heinrich. Parameter Estimation for Text Analysis. (2004).
  - Indraneel Mukherjee, David Blei. Relative Performance Guarantees for Approximate Inference in Latent Dirichlet Allocation. NIPS (2008).
  - Indraneel Mukherjee, David Blei. Relative Performance Guarantees for Approximate Inference in Latent Dirichlet Allocation. NIPS (2008).
  - Yee-Whye Teh, David Newman, Max Welling. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. NIPS (2006).
- Introductory
  - David M. Blei. Introduction to Probabilistic Topic Models. Communications of the ACM 2011.
  - Mark Steyvers, Tom Griffiths. Probabilistic Topic Models. In Landauer, T., Mcnamara, D., Dennis, S., Kintsch, W., Latent Semantic Analysis: A Road to Meaning. (2006).
  - Yi Wang. Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. (2011).
- Vision
  - Chong Wang, David Blei, Fei-Fei Li. Simultaneous Image Classification and Annotation. CVPR (2009).
  - Jyri J. Kivinen, Erik B. Sudderth, Michael I. Jordan. Learning Multiscale Representations of Natural Scenes Using Dirichlet Processes. ICCV (2007).
- User interface
  - Qiaozhu Mei, Xuehua Shen, ChengXiang Zhai. Automatic labeling of multinomial topic models. KDD (2007).

Topic Models: Networks

- Andrew McCallum, Andrés Corrada-Emmanuel, Xuerui Wang. Topic and Role Discovery in Social Networks. IJCAI (2005).
- David A. Broniatowski, Christopher L. Magee. Analysis of Social Dynamics on FDA Panels Using Social Networks Extracted From Meeting Transcripts. SocCom (2010).
- David A. Broniatowski, Christopher L. Magee. Towards A Computational Analysis of Status and Leadership Styles on FDA Panels. SBP (2011).
- David Mimno, Hanna Wallach, Andrew McCallum. Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors. NIPS Workshop on Analyzing Graphs (2008).
- Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, Eric P. Xing. Mixed Membership Stochastic Blockmodels. JMLR (9) 2008 pp. 1981-2014.
- Jonathan Chang, David Blei. Relational Topic Models for Document Networks. AIStats (2009).
- Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences. ICML (2007).
- Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai. Topic modeling with network regularization. WWW (2008).
- Ramesh Nallapati, Amr Ahmed, Eric P. Xing, William Cohen. Joint Latent Topic Models for Text and Citations. KDD (2008).
- Xuerui Wang, Natasha Mohanty, Andrew McCallum. Group and Topic Discovery from Relations and Their Attributes. NIPS (2005).

Topic Models: NLP

- Jordan Boyd-Graber, David M. Blei, Xiaojin Zhu. A Topic Model for Word Sense Disambiguation. EMNLP (2007).
- Jordan Boyd-Graber, David M. Blei. PUTOP: Turning Predominant Senses into a Topic Model for WSD. SEMEVAL (2007).
- Jordan Boyd-Graber, David M. Blei. Syntactic Topic Models. NIPS (2008).
- Jun Fu Cai, Wee Sun Lee, Yee Whye Teh. NUS-ML: Improving Word Sense Disambiguation Using Topic Features. SEMEVAL (2007).
- Kristina Toutanova, Mark Johnson. A Bayesian LDA-based model for semi-supervised part-of-speech tagging. NIPS (2007).
- Mark Johnson. PCFGs, Topic Models, Adaptor Grammars, and Learning Topical Collocations and the Structure of Proper Names. (2010).
- Thomas L. Griffiths, Mark Steyvers, David M. Blei, Joshua B. Tenenbaum. Integrating Topics and Syntax. In , NIPS (2004).

Non-parametric Topic Models

- Changyou Chen, Lan Du, Wray Buntine. Sampling Table Configurations for the Hierarchical Poisson-Dirichlet Process. ECML-PKDD (2011).
- David M. Blei, Thomas Griffiths, Michael Jordan, Joshua Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. NIPS (2003).
- David M. Blei, Thomas L. Griffiths, Michael I. Jordan. The nested Chinese restaurant process and hierarchical topic models. (2007).
- Gregor Heinrich. Infinite LDA. (2011).
- Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu. Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora. KDD (2010).
- Jyri J. Kivinen, Erik B. Sudderth, Michael I. Jordan. Learning Multiscale Representations of Natural Scenes Using Dirichlet Processes. ICCV (2007).
- Wei Li, David Blei, Andrew McCallum. Nonparametric Bayes Pachinko Allocation. (2007).
- Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, David M. Blei. Hierarchical Dirichlet Processes. JASA (101) 2006 pp. .

Topic Models: Scalability and Theory

- Scalability
  - Alexander Smola, Shravan Narayanamurthy. An Architecture for Parallel Topic Models. VLDB (2010).
  - Arthur Asuncion, Padhraic Smyth, Max Welling. Asynchronous Distributed Learning of Topic Models. NIPS (2008).
  - Gregor Heinrich. A generic approach to topic models. ECML/PKDD (2009).
  - Limin Yao, David Mimno, Andrew McCallum. Efficient Methods for Topic Model Inference on Streaming Document Collections. KDD (2009).
  - Ramesh Nallapati, William Cohen, John Lafferty. Parallelized Variational EM for Latent Dirichlet Allocation: An experimental evaluation of speed and scalability. ICDM workshop on high performance data mining (2007).
- Theory
  - Alexander Hinneburg, Hans-Henning Gabriel, Andre Gohr. Bayesian Folding-In with Dirichlet Kernels for PLSI. ICDM (2007).
  - Chris Ding, Tao Li, Wei Peng. On the Equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing. Computational Statistics and Data Analysis (52) 2008 pp. 3913-3927.
  - Hanna Wallach, David Mimno, Andrew McCallum. Rethinking LDA: Why priors matter. NIPS (2009).
  - Mark Girolami, Ata Kabán. On an equivalence between pLSI and LDA. SIGIR (2003).

Temporal Topic Models

- Andre Gohr, Alexander Hinneburg, Rene Schult, Myra Spiliopoulou. Topic Evolution in a Stream of Documents. SDM (2009).
- Andre Gohr, Myra Spiliopoulou, Alexander Hinneburg. Visually Summarizing the Evolution of Documents under a Social Tag. KDIR (2010).
- Chong Wang, David M. Blei, David Heckerman. Continuous Time Dynamic Topic Models. UAI (2008).
- David A. Broniatowski, Christopher L. Magee. Towards A Computational Analysis of Status and Leadership Styles on FDA Panels. SBP (2011).
- David M. Blei, John D. Lafferty. Dynamic Topic Models. ICML (2006).
- Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu. Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora. KDD (2010).
- Ramesh Nallapati, William Cohen, Susan Ditmore, John Lafferty, Kin Ung. Multi-scale Topic Tomography. KDD (2007).
- Xuerui Wang, Andrew McCallum. Topics Over Time: a non-Markov continuous-time model of topical trends. KDD (2006).