



Natural Language Processing

Lecture 11: Topic Modeling

Qun Liu, Valentin Malykh
Huawei Noah's Ark Lab



Spring 2022
A course delivered at KFU, Kazan



Content

- 1 ML Recap
- 2 Sampling Basics
- 3 Topic Modeling



Content

1 ML Recap

2 Sampling Basics

3 Topic Modeling

Generative vs. Discriminative Models

- Recall that, in Bayesian networks, there could be many different, but equivalent models of the same joint distribution



Discriminative



Generative

- Although these two models are equivalent (in the sense that they imply the same independence relations), they can differ significantly when it comes to inference/prediction

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA



Generative vs. Discriminative Models



Discriminative



Generative

- Generative models: we can think of the observations as being generated by the latent variables
 - Start sampling at the top and work downwards
 - Examples?

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA



Generative vs. Discriminative Models



Discriminative



Generative

- Generative models: we can think of the observations as being generated by the latent variables
 - Start sampling at the top and work downwards
 - Examples: **HMMs, naïve Bayes, LDA**



Generative vs. Discriminative Models



Discriminative



Generative

- **Discriminative models:** most useful for discriminating the values of the latent variables
 - Almost always used for supervised learning
 - Examples?

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA



Generative vs. Discriminative Models



Discriminative



Generative

- Discriminative models: most useful for discriminating the values of the latent variables
 - Almost always used for supervised learning
 - Examples: CRFs

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA

Generative vs. Discriminative Models



Discriminative



Generative

- Suppose we are only interested in the prediction task (i.e., estimating $p(Y|X)$)
 - Discriminative model: $p(X, Y) = p(X)p(Y|X)$
 - Generative model: $p(X, Y) = p(Y)p(X|Y)$



Models of Text Documents

- Bag-of-words models: assume that the ordering of words in a document do not matter
 - This is typically false as certain phrases can only appear together
- Unigram model: all words in a document are drawn uniformly at random from categorical distribution
- Mixture of unigrams model: for each document, we first choose a topic z and then generate words for the document from the conditional distribution $p(w|z)$
 - Topics are just probability distributions over words

A slide from: Nicholas Ruozzi, University of Texas, Dallas CS6347 (2015), Lecture 17: Topic Models and LDA



Topic Models

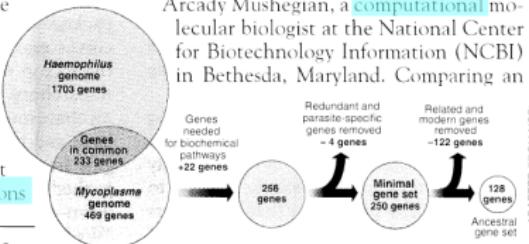
Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,¹ two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.



Content

1 ML Recap

2 Sampling Basics

3 Topic Modeling



Aside: don't always sample!

"Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse."

— Alan Sokal, 1996

Example: numerical solutions to (nice) 1D integrals are fast

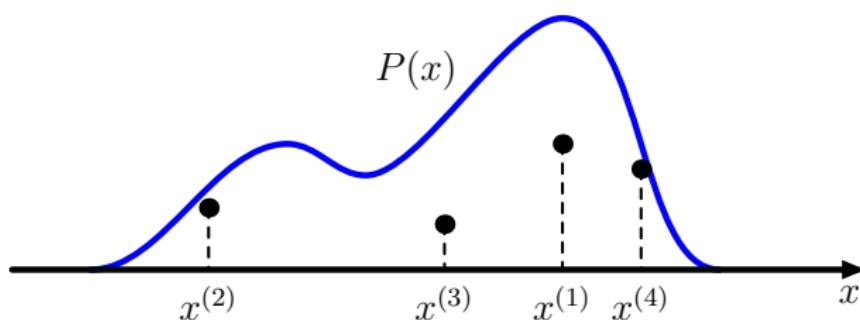
```
octave:1> 4 * quadl(@(x) sqrt(1-x.^2), 0, 1, tolerance)
```

Gives π to 6 dp's in 108 evaluations, machine precision in 2598.

(NB Matlab's quadl fails at zero tolerance)

Sampling from distributions

Draw points uniformly under the curve:

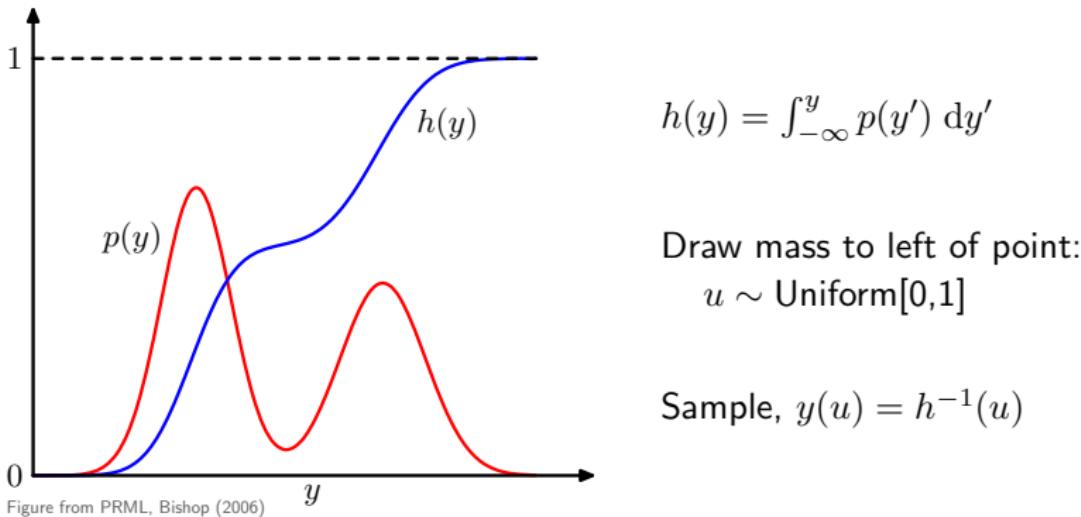


Probability mass to left of point $\sim \text{Uniform}[0,1]$

7

Sampling from distributions

How to convert samples from a Uniform[0,1] generator:



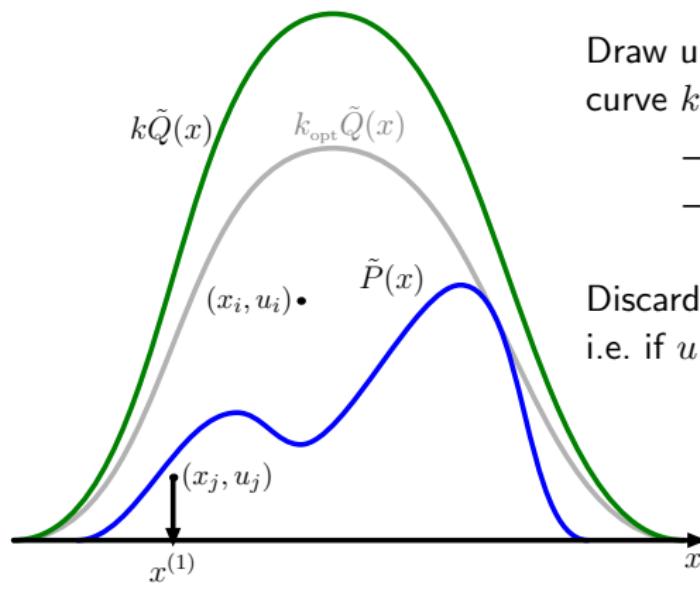
Although we can't always compute and invert $h(y)$

8

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Rejection sampling

Sampling underneath a $\tilde{P}(x) \propto P(x)$ curve is also valid



Draw underneath a simple curve $k\tilde{Q}(x) \geq \tilde{P}(x)$:

- Draw $x \sim Q(x)$
- height $u \sim \text{Uniform}[0, k\tilde{Q}(x)]$

Discard the point if above \tilde{P} , i.e. if $u > \tilde{P}(x)$

9

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Content

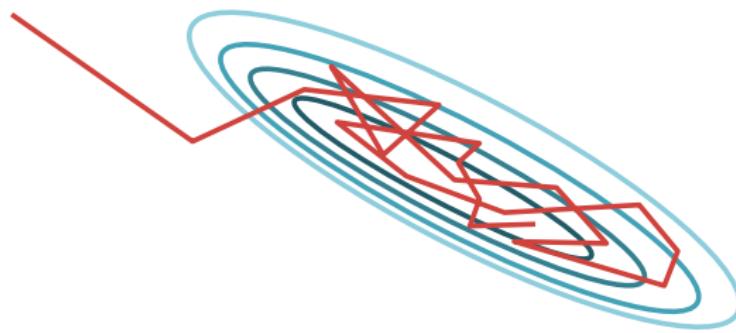
2

- Sampling Basics
 - Gibbs Sampling



MCMC

- **Goal:** Draw approximate, correlated samples from a target distribution $p(x)$
- **MCMC:** Performs a biased random walk to explore the distribution



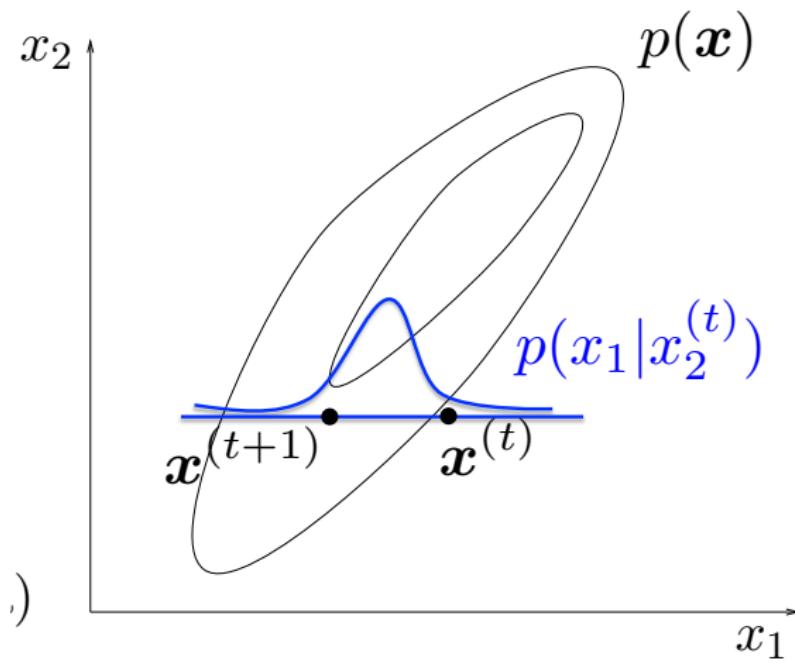
11

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling





Gibbs Sampling

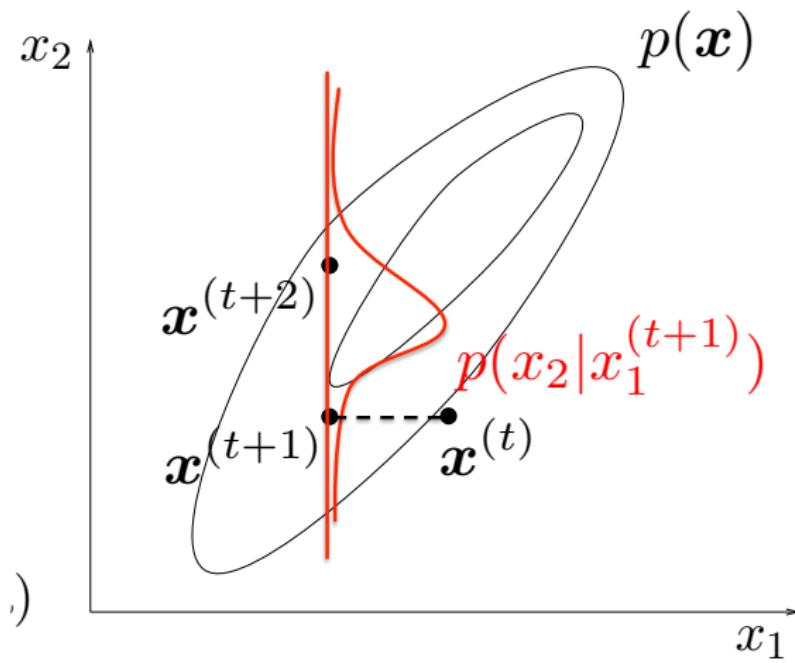


14

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Gibbs Sampling

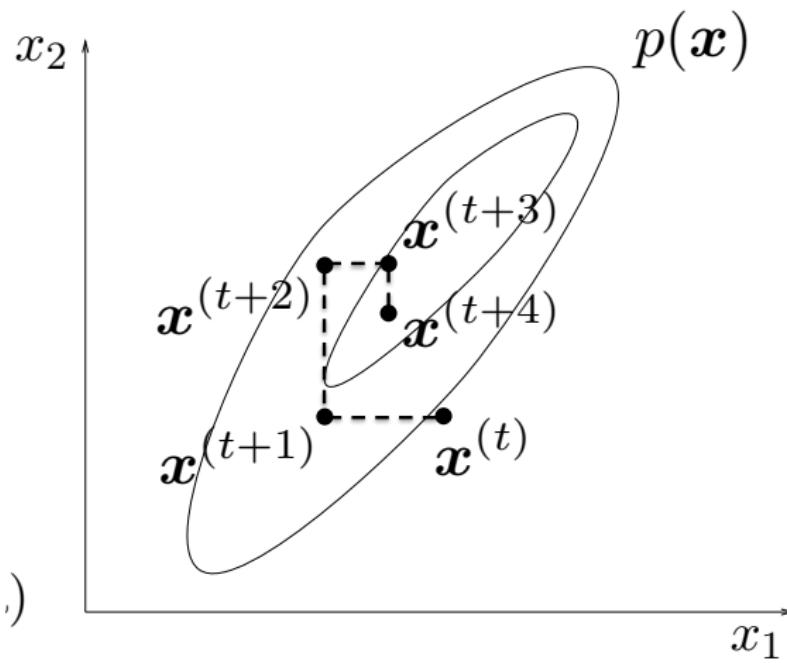


15

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Gibbs Sampling



16

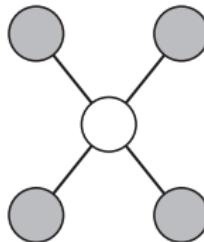
A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



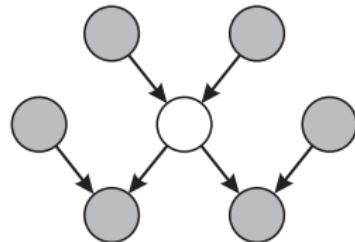
Gibbs Sampling

Full conditionals only need to condition on the Markov Blanket

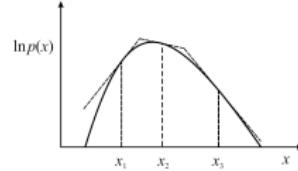
MRF



Bayes Net



- Must be “easy” to sample from conditionals
- Many conditionals are log-concave and are amenable to adaptive rejection sampling



17

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



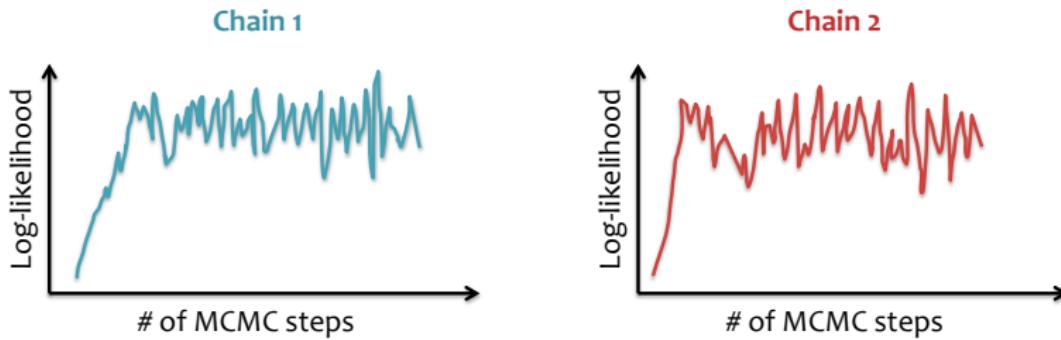
Why does Gibbs sampling work?

- Metropolis-Hastings
 - Markov chains
 - Stationary distribution
 - MH Algorithm
 - Constructs a Markov chain whose stationary distribution is the desired distribution
 - Proof that samples will be from desired distribution:
 - Sufficient conditions for constructing a markov chain with desired stationary distribution:
 - ergodicity
 - detailed balance (stronger, than what we need, but easier for the proof)
- Gibbs Sampling is a special case of Metropolis-Hastings
 - a special proposal distribution, which ensures the hastings ratio is always 1.0



Practical Issues

- **Question:** How do we assess convergence of the Markov chain?
- **Answer:** It's not easy!
 - Compare statistics of multiple independent chains
 - Ex: Compare log-likelihoods

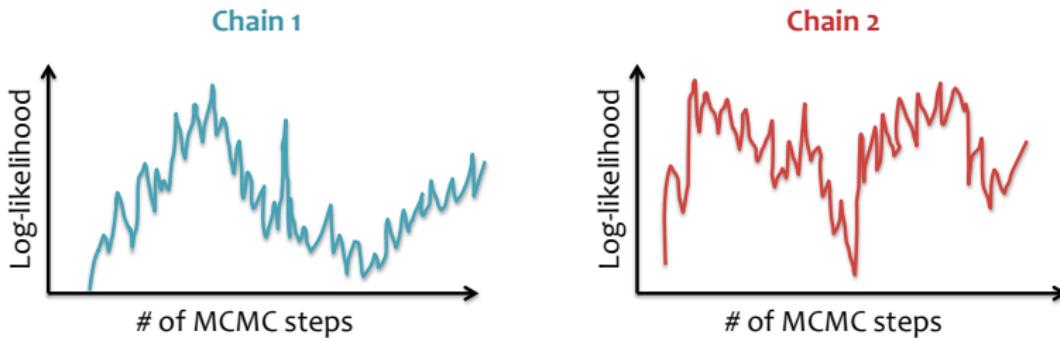


19



Practical Issues

- **Question:** How do we assess convergence of the Markov chain?
- **Answer:** It's not easy!
 - Compare statistics of multiple independent chains
 - Ex: Compare log-likelihoods

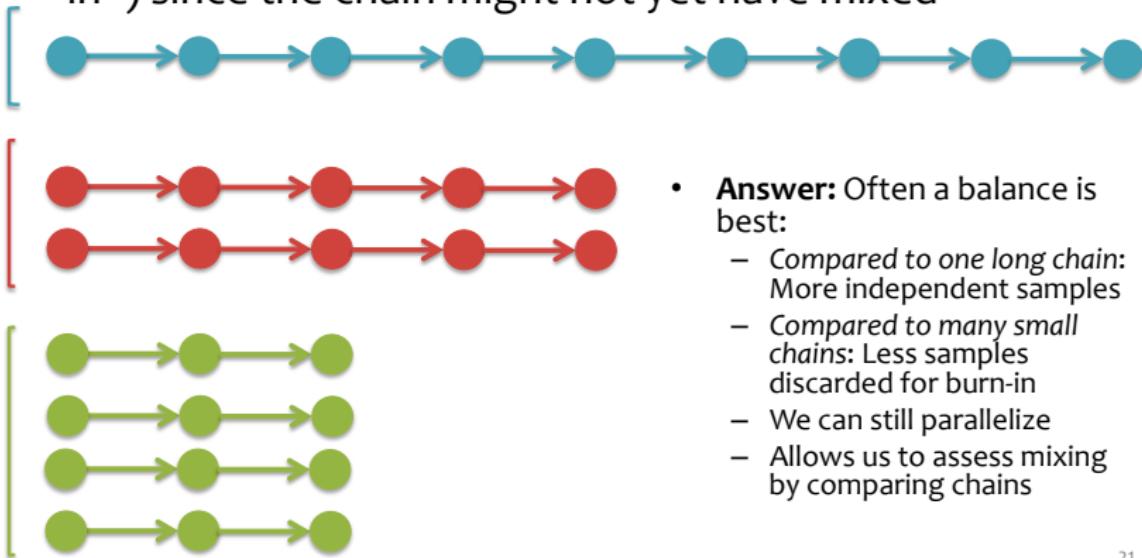


20



Practical Issues

- **Question:** Is one long Markov chain better than many short ones?
- **Note:** typical to discard initial samples (aka. “burn-in”) since the chain might not yet have mixed



- **Answer:** Often a balance is best:
 - Compared to one long chain: More independent samples
 - Compared to many small chains: Less samples discarded for burn-in
 - We can still parallelize
 - Allows us to assess mixing by comparing chains

21



Content

1 ML Recap

2 Sampling Basics

3 Topic Modeling



Topic Modeling

Motivation:

Suppose you're given a massive corpora and asked to carry out the following tasks

- **Organize** the documents into **thematic categories**
- **Describe** the evolution of those categories **over time**
- Enable a domain expert to **analyze and understand** the content
- Find **relationships** between the categories
- Understand how **authorship** influences the content



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Topic Modeling

Motivation:

Suppose you're given a massive corpora and asked to carry out the following tasks

- **Organize** the documents into **thematic categories**
- **Describe** the evolution of those categories **over time**
- Enable a domain expert to **analyze and understand** the content
- Find **relationships** between the categories
- Understand how **authorship** influences the content

Topic Modeling:

A method of (usually unsupervised) discovery of latent or hidden structure in a corpus

- Applied primarily to text corpora, but **techniques are more general**
- Provides a **modeling toolbox**
- Has prompted the exploration of a variety of new **inference methods** to accommodate **large-scale datasets**



Topic Modeling

Dirichlet-multinomial regression (DMR) topic model on ICML
(Mimno & McCallum, 2008)

Topic 0 [0.152]



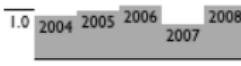
problem, optimization, problems, convex, convex optimization, linear, semidefinite programming, formulation, sets, constraints, proposed, margin, maximum margin, optimization problem, linear programming, programming, procedure, method, cutting plane, solutions

Topic 54 [0.051]



decision trees, trees, tree, decision tree, decision, tree ensemble, junction tree, decision tree learners, leaf nodes, arithmetic circuits, ensembles modts, skewing, ensembles, anytime induction decision trees, trees trees, random forests, objective decision trees, tree learners, trees grove, candidate split

Topic 99 [0.066]



inference, approximate inference, exact inference, markov chain, models, approximate, gibbs sampling, variational, bayesian, variational inference, variational bayesian, approximation, sampling, methods, exact, bayesian inference, dynamic bayesian, process, mcmc, efficient

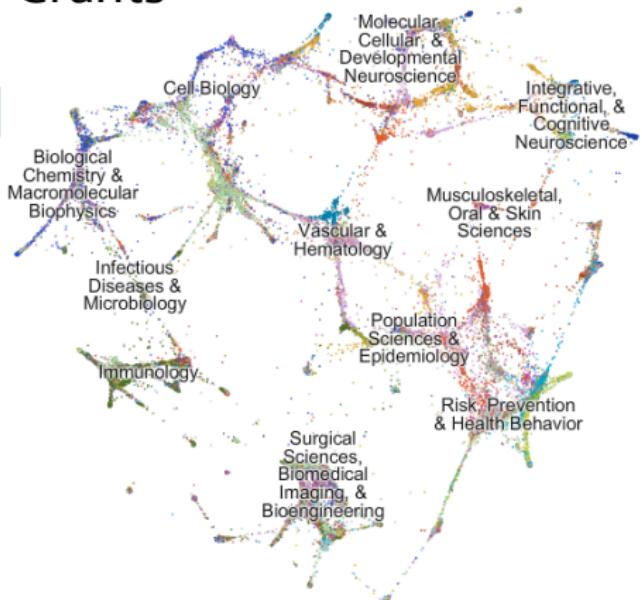
<http://www.cs.umass.edu/~mimno/icml100.html>



Topic Modeling

- Map of NIH Grants

(Talley et al., 2011)



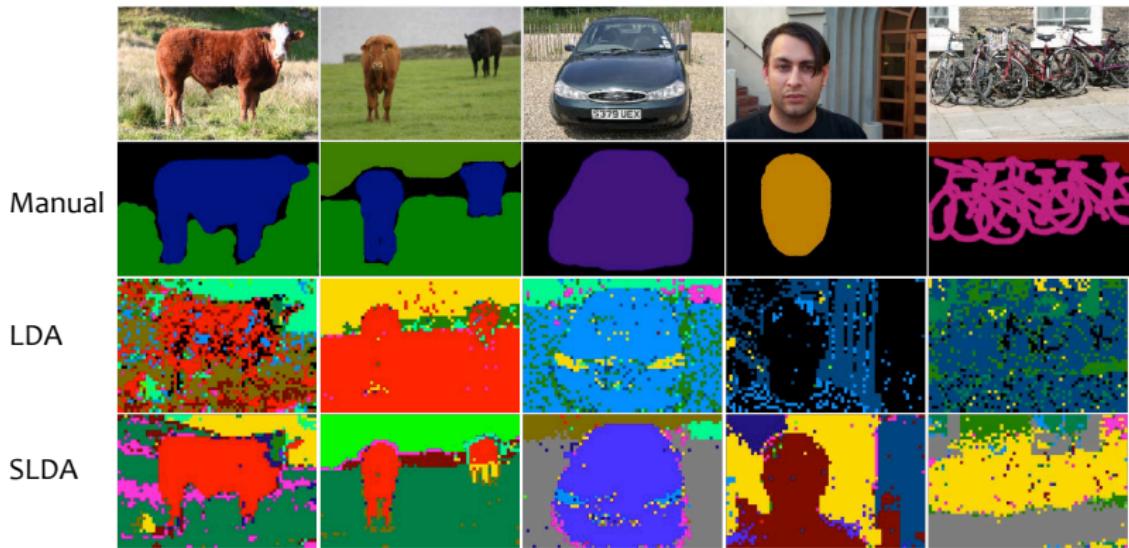
<https://app.nihmaps.org/>

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Other Applications of Topic Models

- Spacial LDA

(Wang & Grimson, 2007)



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Other Applications of Topic Models

- Word Sense Induction

(Brody & Lapata, 2009)

Senses of drug (WSJ)
1. U.S., administration, federal, against, war, dealer
2. patient, people, problem, doctor, company, abuse
3. company, million, sale, maker, stock, inc.
4. administration, food, company, approval, FDA

Senses of drug (BNC)
1. patient, treatment, effect, anti-inflammatory
2. alcohol, treatment, patient, therapy, addiction
3. patient, new, find, effect, choice, study
4. test, alcohol, patient, abuse, people, crime
5. trafficking, trafficker, charge, use, problem
6. abuse, against, problem, treatment, alcohol
7. people, wonder, find, prescription, drink, addict
8. company, dealer, police, enforcement, patient

- Selectional Preference

(Ritter et al., 2010)

Topic t	Arg1	Relations which assign highest probability to t	Arg2
18	The residue - The mixture - The reaction mixture - The solution - the mixture - the reaction mixture - the residue - The reaction - the solution - The filtrate - the reaction - The product - The crude product - The pellet - The organic layer - Thereto - This solution - The resulting solution - Next - The organic phase - The resulting mixture - C.)	was treated with, is treated with, was poured into, was extracted with, was purified by, was diluted with, was filtered through, is dissolved in, is washed with	EtOAc - CH ₂ Cl ₂ - H ₂ O - CH ₃ .sub.2Cl.sub.2 - H ₂ .sub.2O - water - MeOH - NaHCO ₃ - Et ₂ O - NHCl - CHCl ₃ - NHCl - drop-wise - CH ₂ Cl ₂ - Celite - Et ₂ .sub.2O - Cl ₃ .sub.2 - NaOH - AcOEt - CH ₂ Cl ₂ - the mixture - saturated NaHCO ₃ - SiO ₂ - H ₂ O - N hydrochloric acid - NHCl - preparative HPLC - to 0 C



Content

3

Topic Modeling

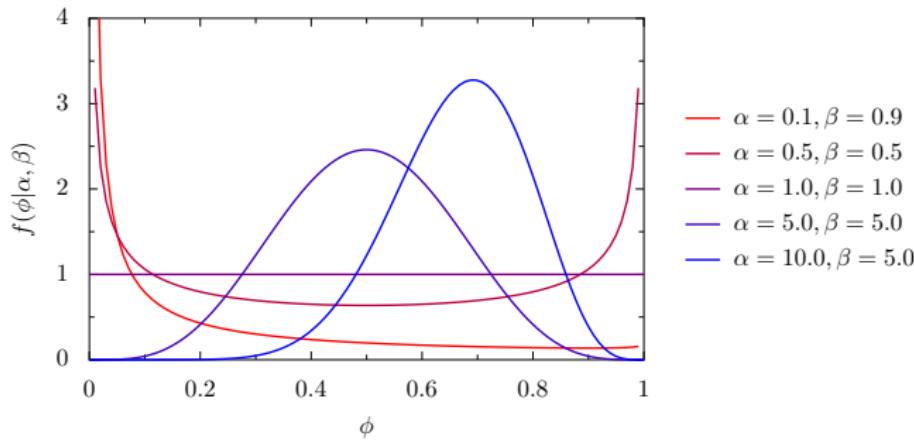
- Latent Dirichlet Allocation
- LDA Inference & Learning
- Extentions of LDA



Beta-Bernoulli Model

- Beta Distribution

$$f(\phi|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$$



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Beta-Bernoulli Model

- Generative Process

$\phi \sim \text{Beta}(\alpha, \beta)$	<i>[draw distribution over words]</i>
For each word $n \in \{1, \dots, N\}$	
$x_n \sim \text{Bernoulli}(\phi)$	<i>[draw word]</i>

- Example corpus (heads/tails)

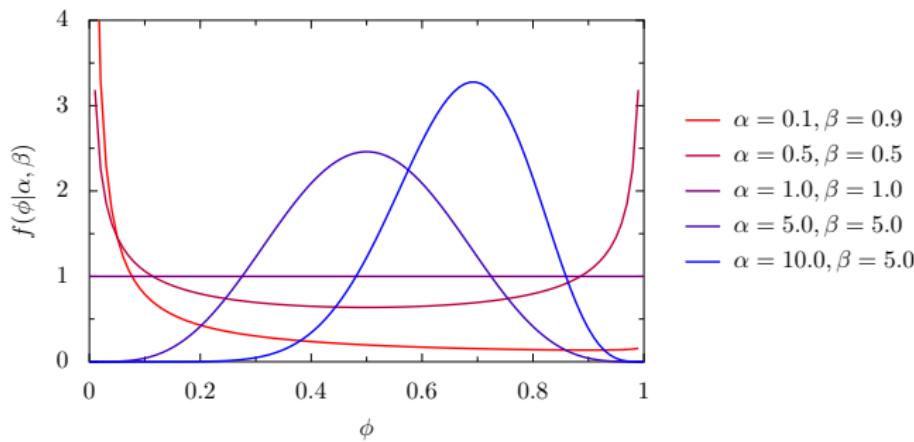
H	T	T	H	H	T	T	H	H	H
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}



Dirichlet-Multinomial Model

- Dirichlet Distribution

$$f(\phi|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

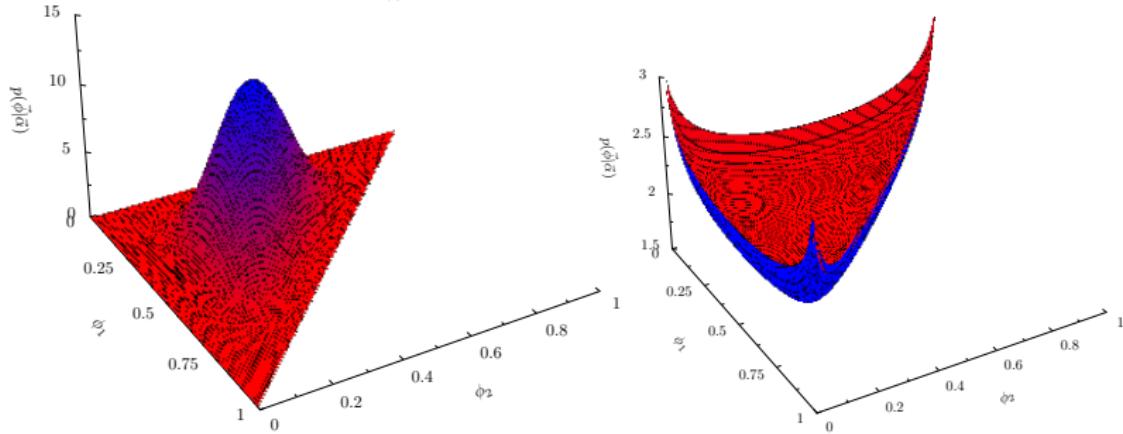


A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Dirichlet-Multinomial Model

- Dirichlet Distribution

$$p(\vec{\phi} | \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K \phi_k^{\alpha_k - 1} \quad \text{where } B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)}$$



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Dirichlet-Multinomial Model

- Generative Process

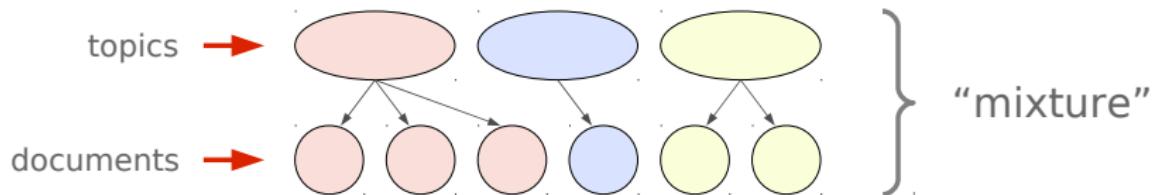
$\phi \sim \text{Dir}(\beta)$	<i>[draw distribution over words]</i>
For each word $n \in \{1, \dots, N\}$	
$x_n \sim \text{Mult}(1, \phi)$	<i>[draw word]</i>

- Example corpus

the	he	is	the	and	the	she	she	is	is
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}

Dirichlet-Multinomial Mixture Model

- Generative Process



- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Dirichlet-Multinomial Mixture Model

- Generative Process

For each topic $k \in \{1, \dots, K\}$:

$$\phi_k \sim \text{Dir}(\beta) \quad [\text{draw distribution over words}]$$

$$\theta \sim \text{Dir}(\alpha) \quad [\text{draw distribution over topics}]$$

For each document $m \in \{1, \dots, M\}$

$$z_m \sim \text{Mult}(1, \theta) \quad [\text{draw topic assignment}]$$

For each word $n \in \{1, \dots, N_m\}$

$$x_{mn} \sim \text{Mult}(1, \phi_{z_m}) \quad [\text{draw word}]$$

- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

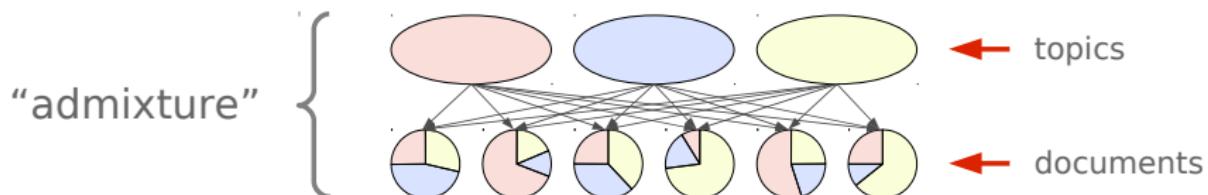
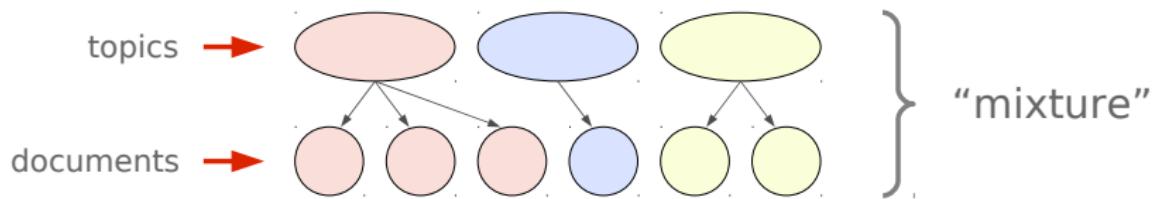
the	and	the
x_{21}	x_{22}	x_{23}

Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

Mixture vs. Admixture (LDA)



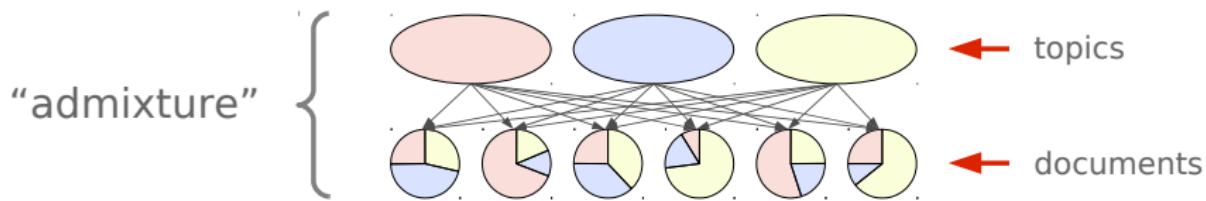
Diagrams from Wallach, JHU 2011, slides

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Latent Dirichlet Allocation

- Generative Process



- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

Document 2

she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Latent Dirichlet Allocation

- Generative Process

For each topic $k \in \{1, \dots, K\}$:

$\phi_k \sim \text{Dir}(\beta)$ [draw distribution over words]

For each document $m \in \{1, \dots, M\}$

$\theta_m \sim \text{Dir}(\alpha)$ [draw distribution over topics]

For each word $n \in \{1, \dots, N_m\}$

$z_{mn} \sim \text{Mult}(1, \theta_m)$ [draw topic assignment]

$x_{mn} \sim \phi_{z_{mi}}$ [draw word]

- Example corpus

the	he	is
x_{11}	x_{12}	x_{13}

Document 1

the	and	the
x_{21}	x_{22}	x_{23}

Document 2

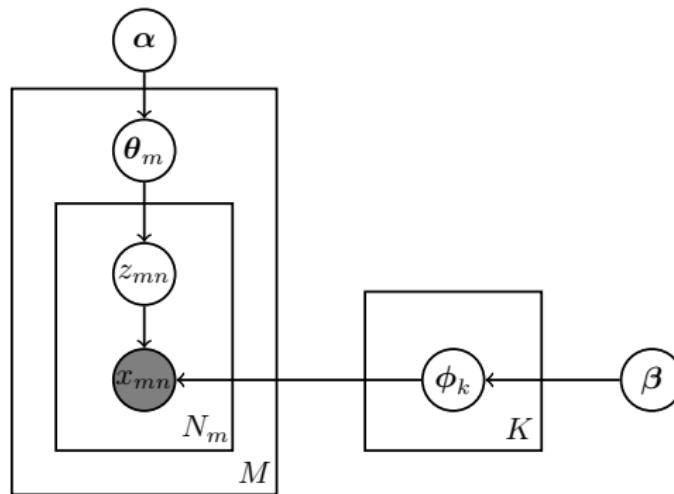
she	she	is	is
x_{31}	x_{32}	x_{33}	x_{34}

Document 3



Latent Dirichlet Allocation

- Plate Diagram

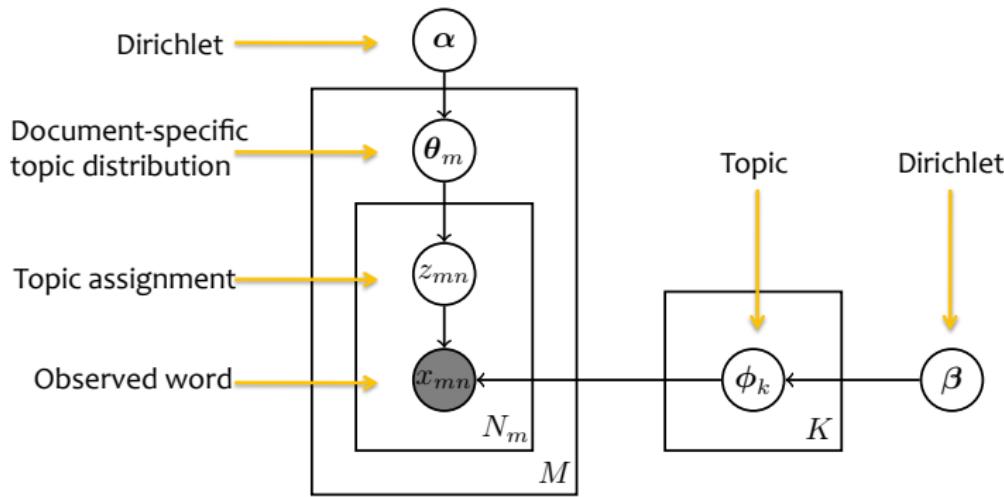


A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Latent Dirichlet Allocation

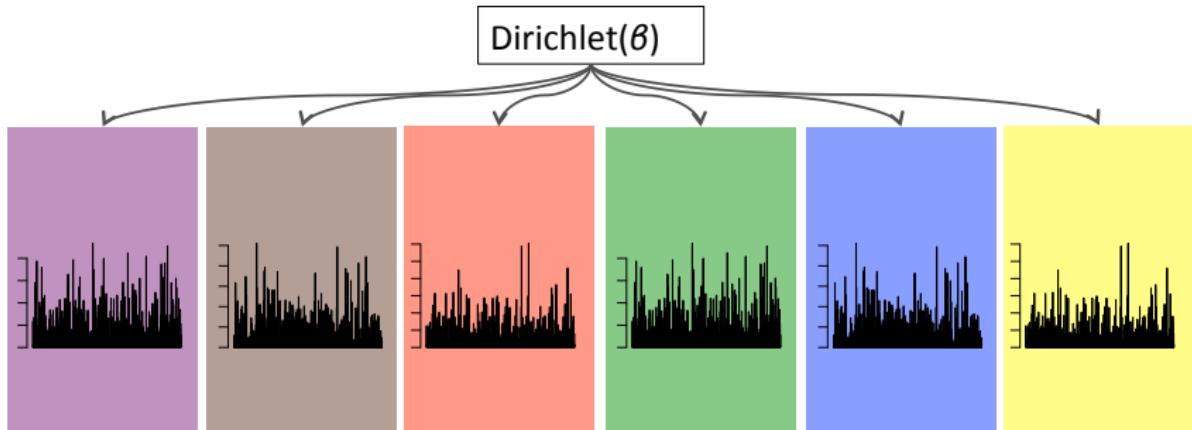
- Plate Diagram



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



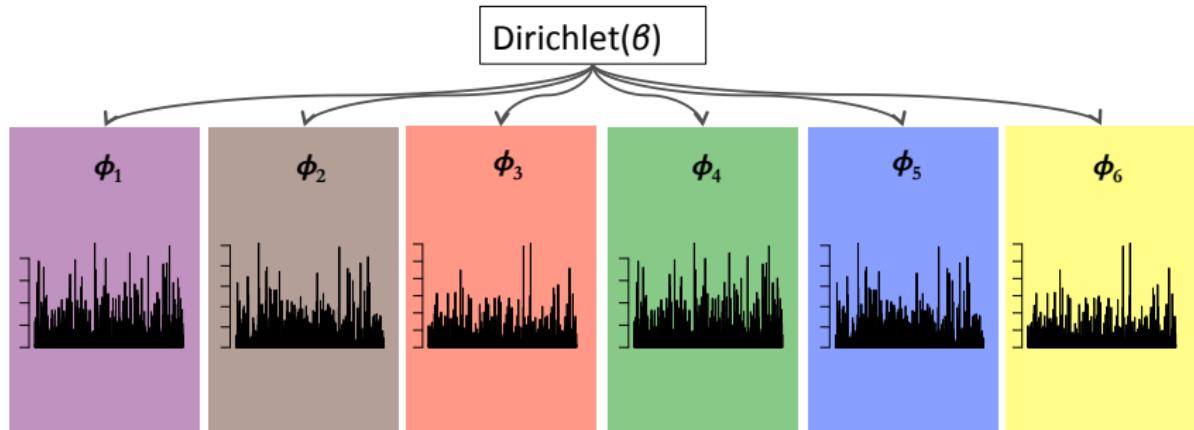
LDA for Topic Modeling



- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by ϕ_k



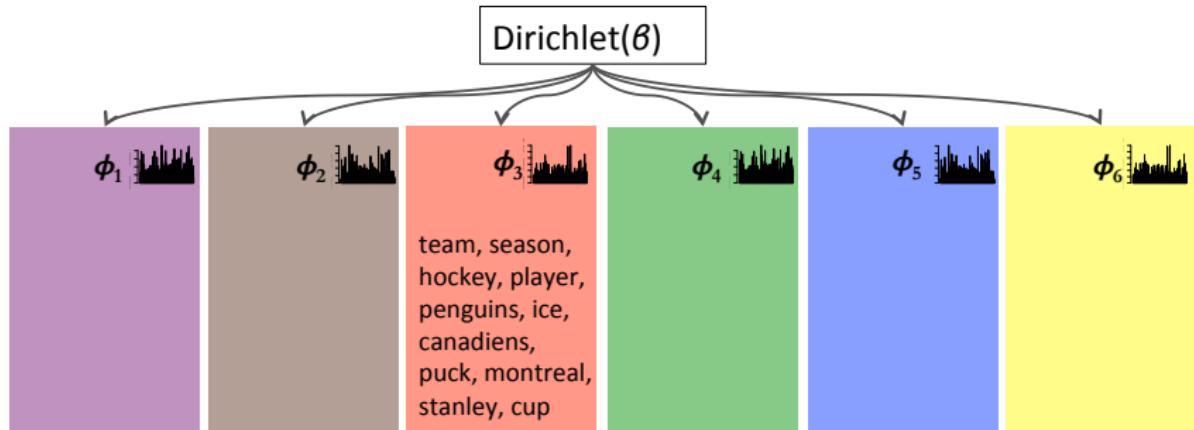
LDA for Topic Modeling



- The **generative story** begins with only a **Dirichlet prior** over the topics.
- Each **topic** is defined as a **Multinomial distribution** over the vocabulary, parameterized by ϕ_k



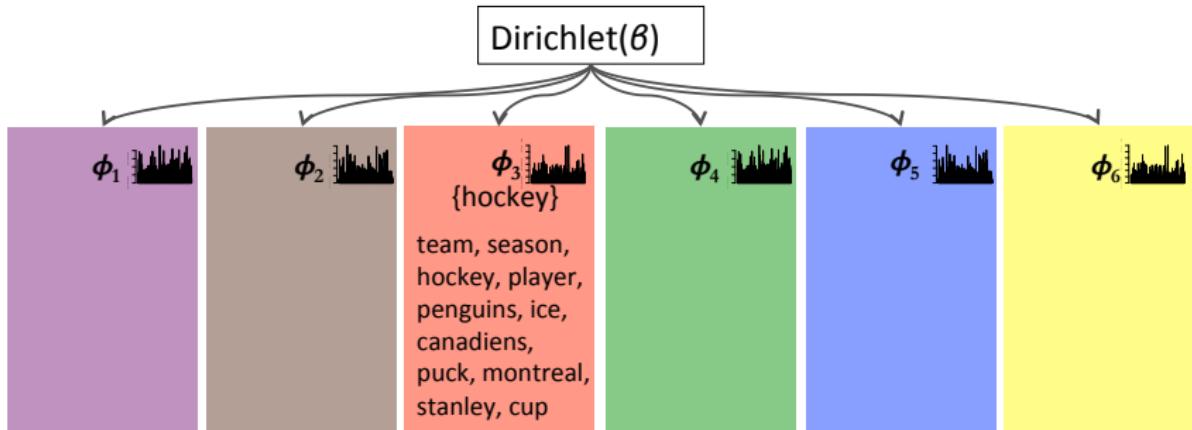
LDA for Topic Modeling



- A topic is visualized as its **high probability words**.



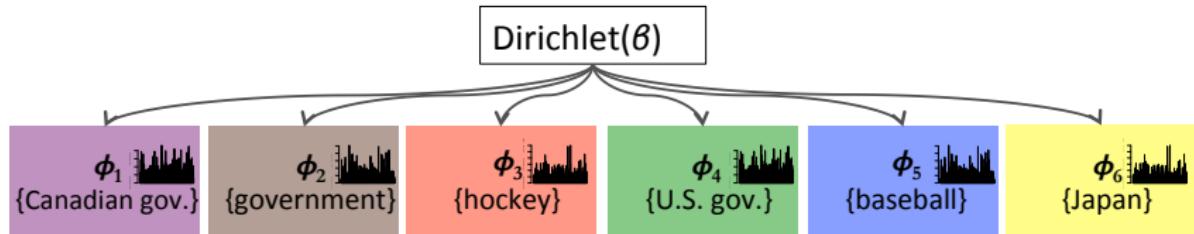
LDA for Topic Modeling



- A topic is visualized as its **high probability words**.
- A pedagogical **label** is used to identify the topic.



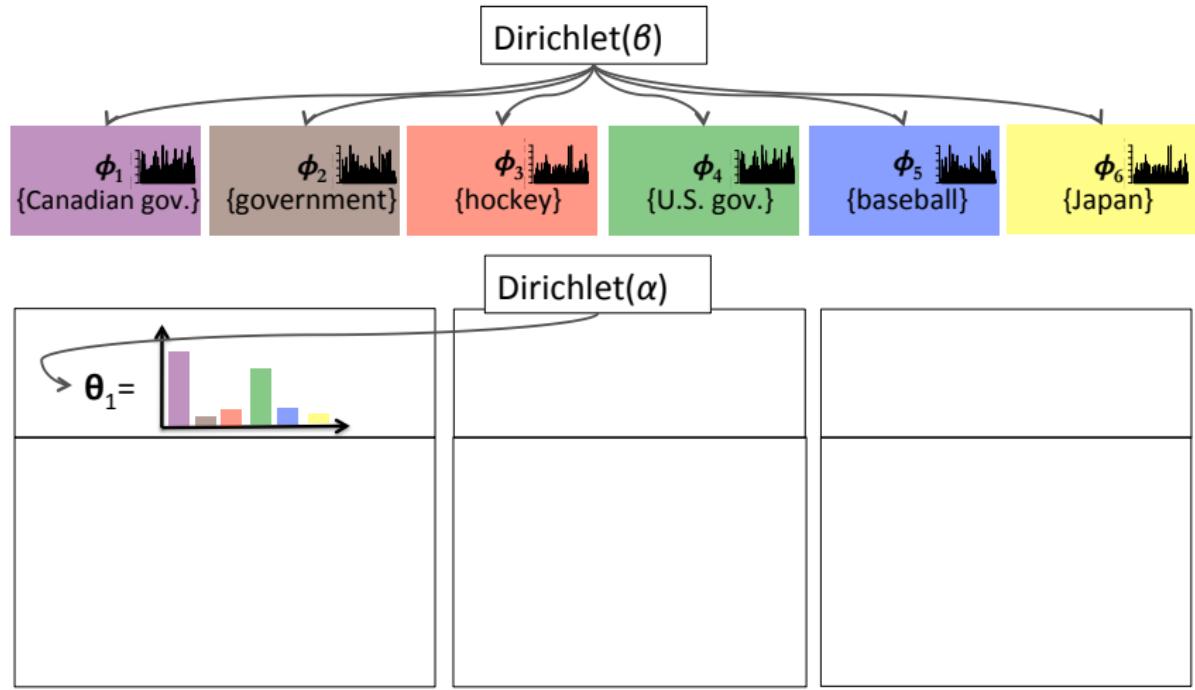
LDA for Topic Modeling



- A topic is visualized as its high probability words.
- A pedagogical **label** is used to identify the topic.



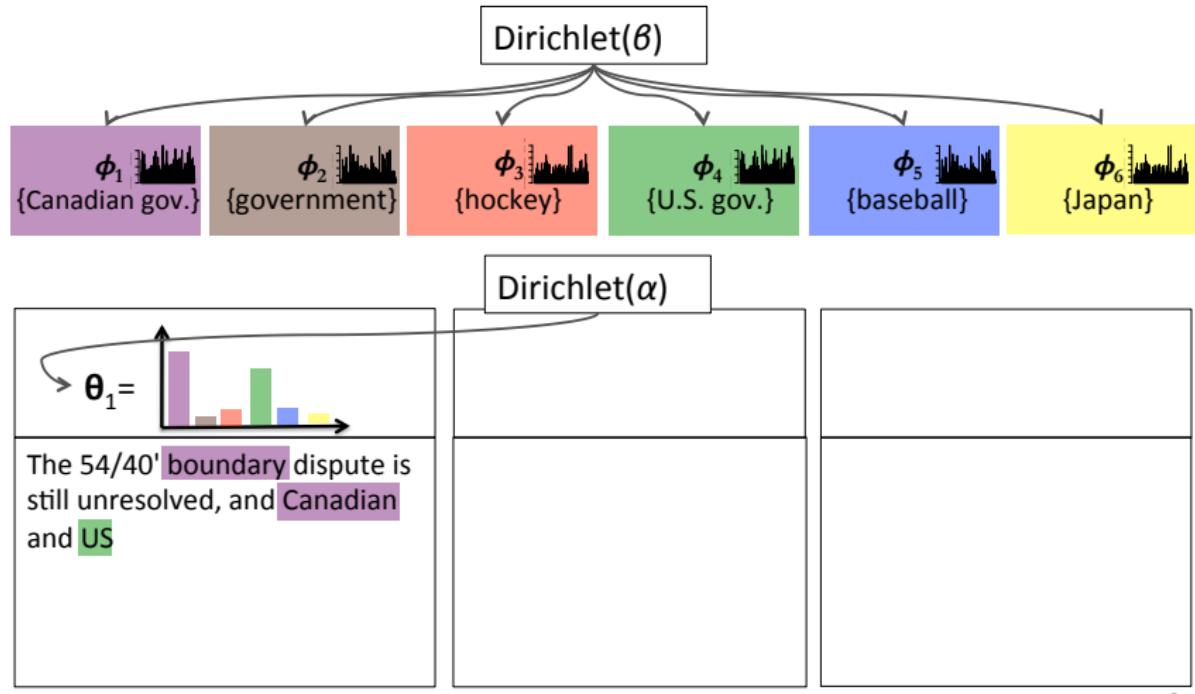
LDA for Topic Modeling



47

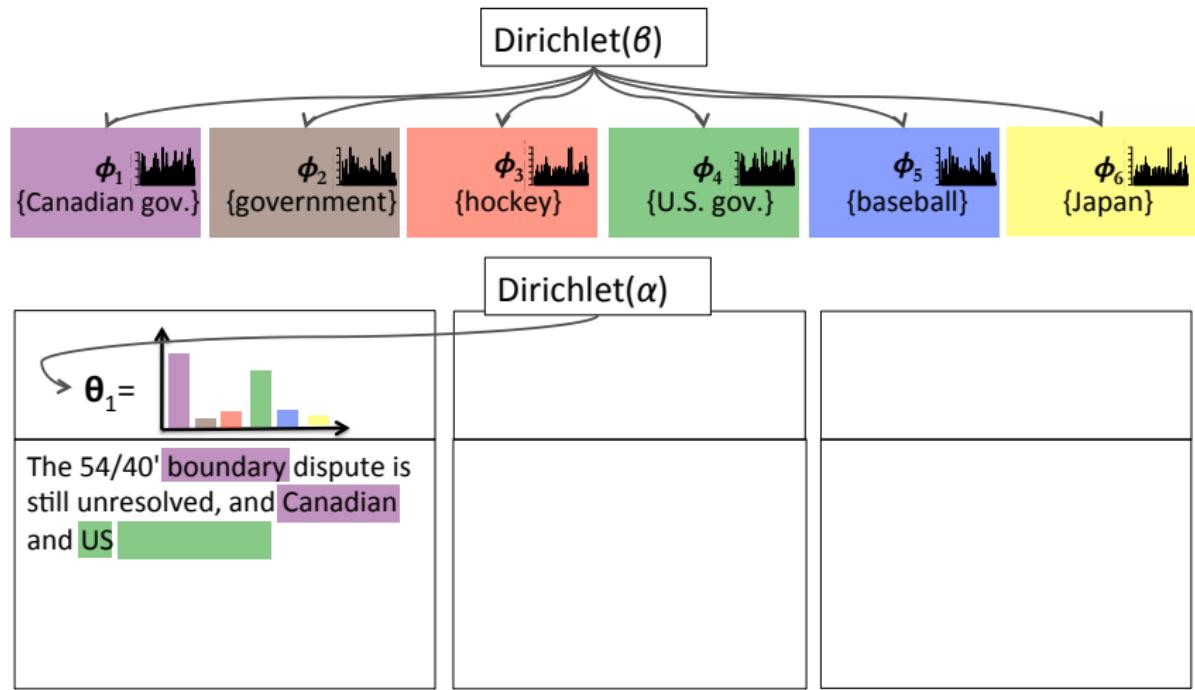


LDA for Topic Modeling





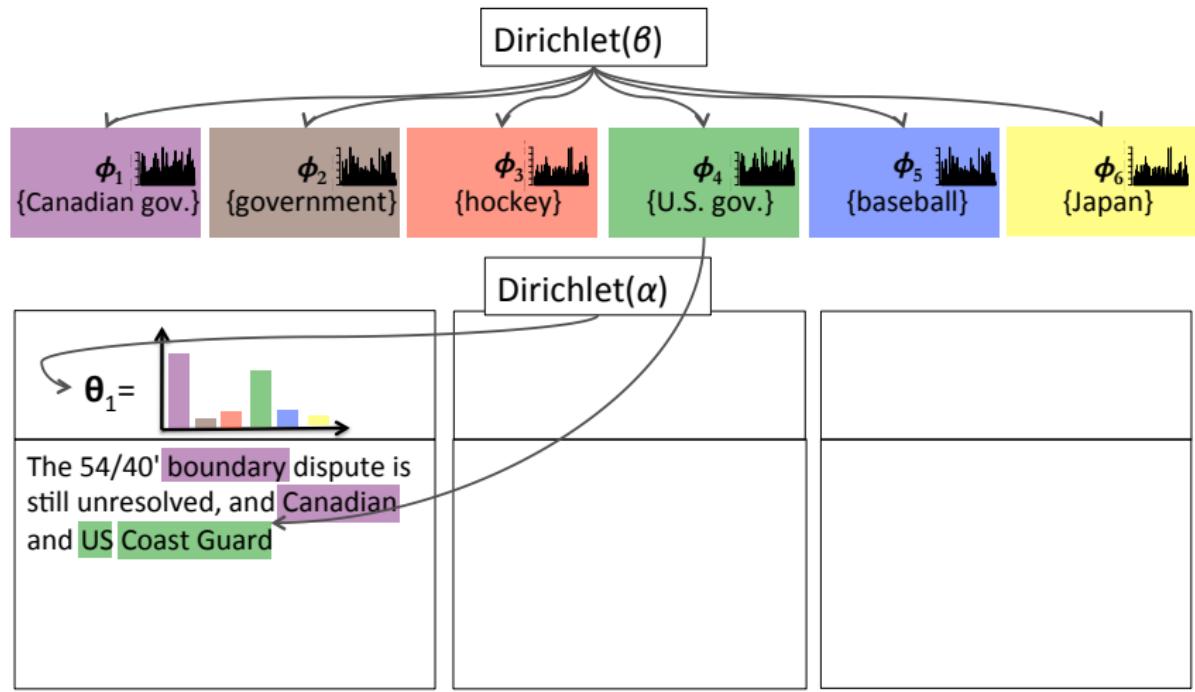
LDA for Topic Modeling



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



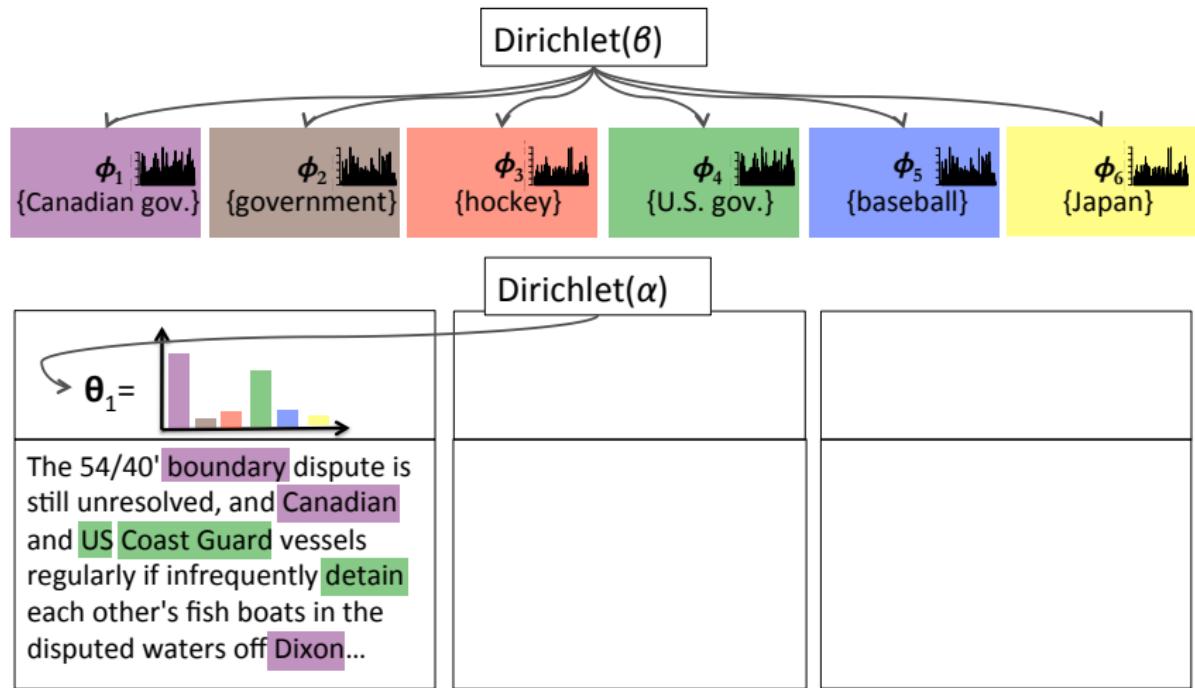
LDA for Topic Modeling



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



LDA for Topic Modeling

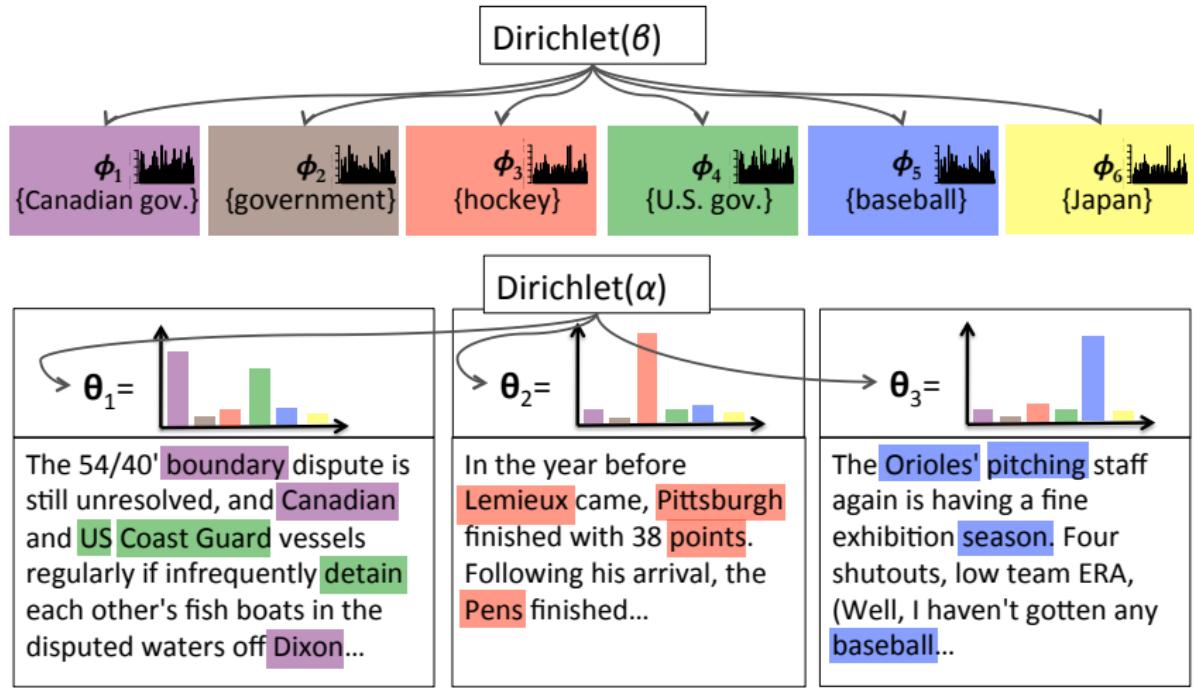


51

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



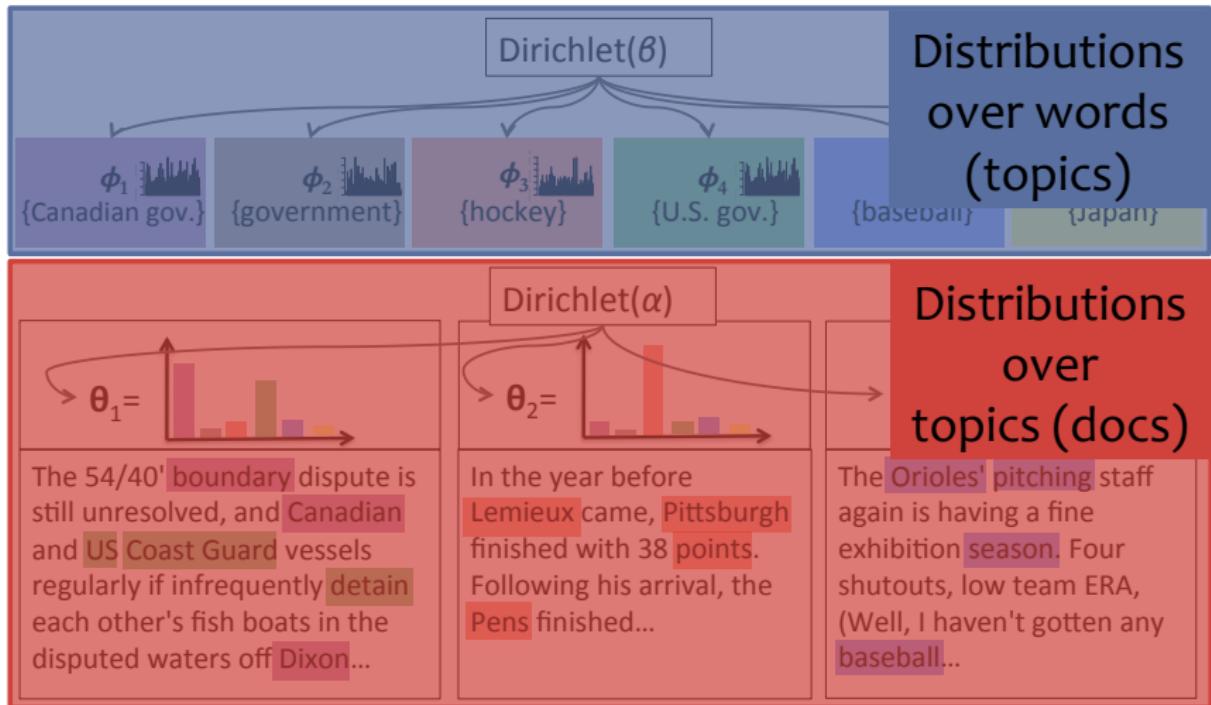
LDA for Topic Modeling



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



LDA for Topic Modeling



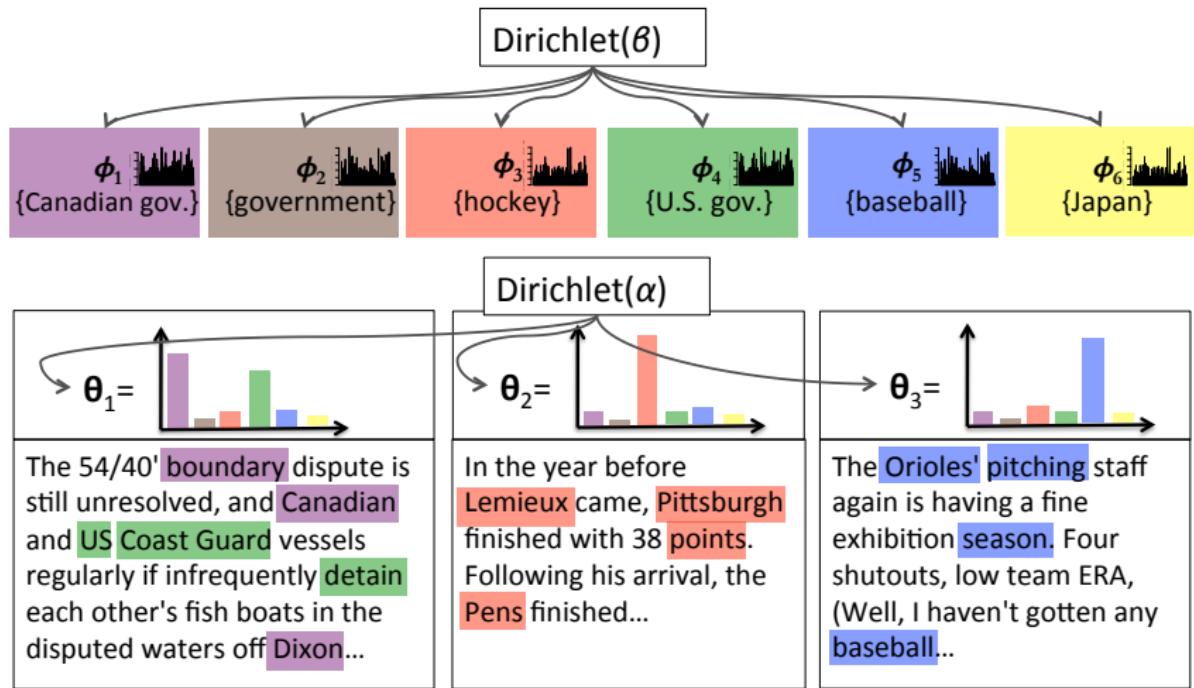
53

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling





LDA for Topic Modeling



54

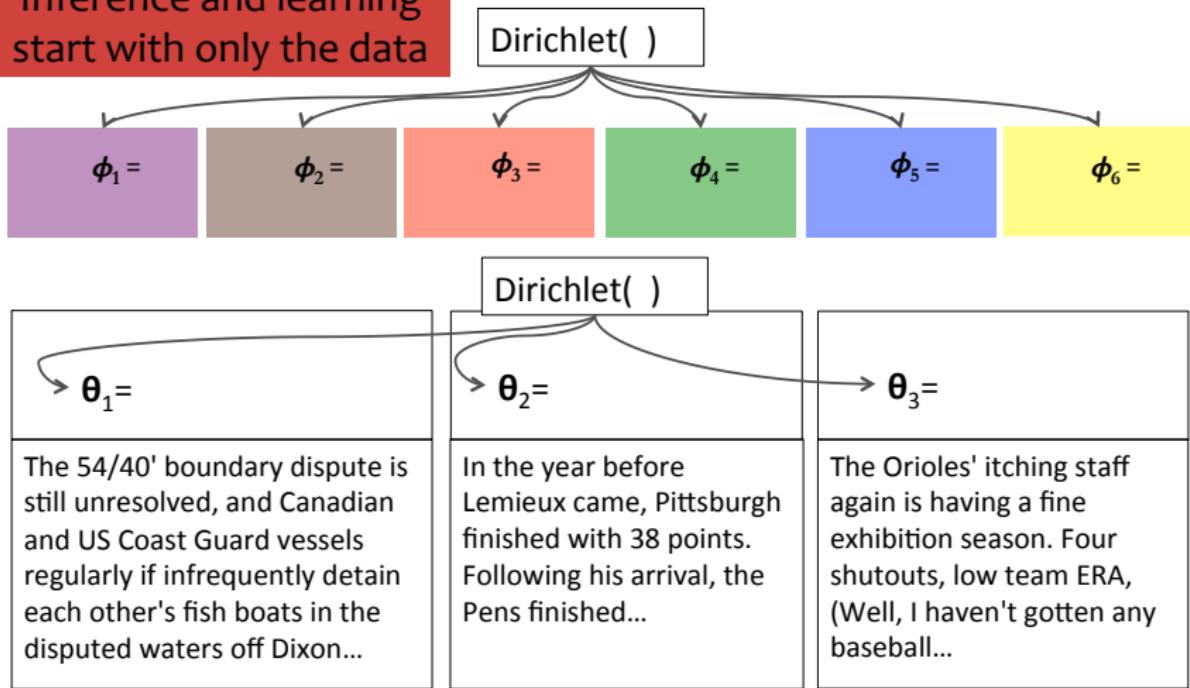
A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling





LDA for Topic Modeling

Inference and learning
start with only the data



55

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Latent Dirichlet Allocation

Questions:

- Is this a believable story for the generation of a corpus of documents?
- Why might it work well anyway?



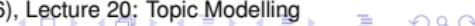
Latent Dirichlet Allocation

Why does LDA “work”?

- LDA trades off two goals.
 - ① For each document, allocate its words to as few topics as possible.
 - ② For each topic, assign high probability to as few terms as possible.
- These goals are at odds.
 - Putting a document in a single topic makes #2 hard:
All of its words must have probability under that topic.
 - Putting very few words in each topic makes #1 hard:
To cover a document's words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.

Slide from David Blei, MI 55, 2012

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling





Latent Dirichlet Allocation

How does this relate to my other favorite model for capturing low-dimensional representations of a corpus?

- Builds on latent semantic analysis (Deerwester et al., 1990; Hofmann, 1999)
- It is a mixed-membership model (Erosheva, 2004).
- It relates to PCA and matrix factorization (Jakulin and Buntine, 2002)
- Was independently invented for genetics (Pritchard et al., 2000)

Slide from David Blei MI 55 2012

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling





Content

3

Topic Modeling

- Latent Dirichlet Allocation
- LDA Inference & Learning
- Extentions of LDA



Unsupervised Learning

Three learning paradigms:

1. Maximum likelihood

$$\arg \max_{\theta} p(X|\theta)$$

2. Maximum a posteriori (MAP)

$$\arg \max_{\theta} p(\theta|X) \propto p(X|\theta)p(\theta)$$

3. Bayesian approach

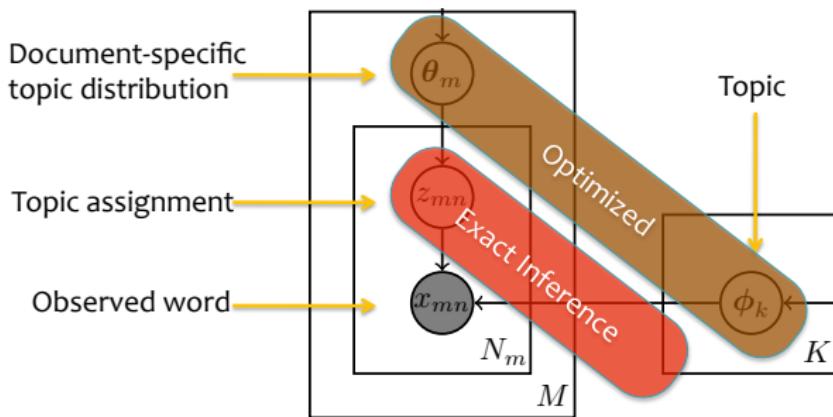
Estimate the posterior:

$$p(\theta|X) = \dots$$



LDA Inference

- Standard EM (Maximum Likelihood)

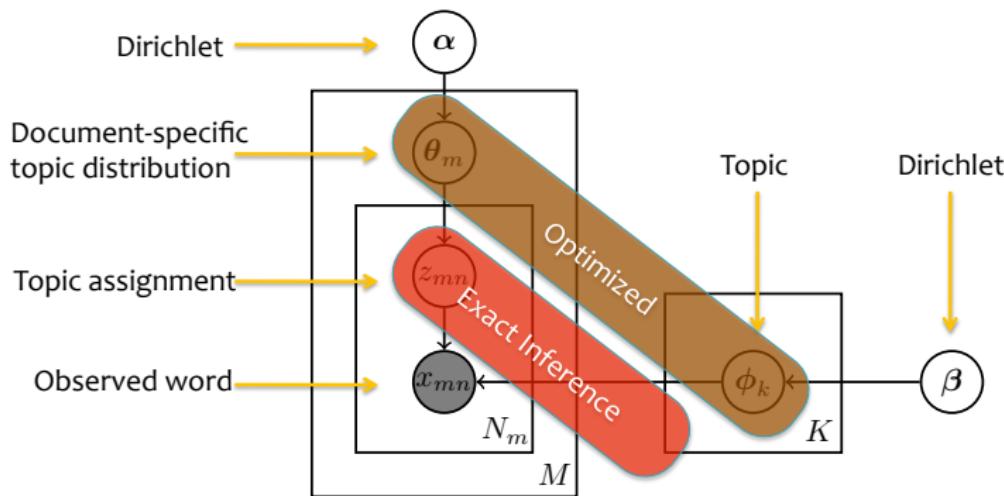


A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



LDA Inference

- Standard EM (MAP)

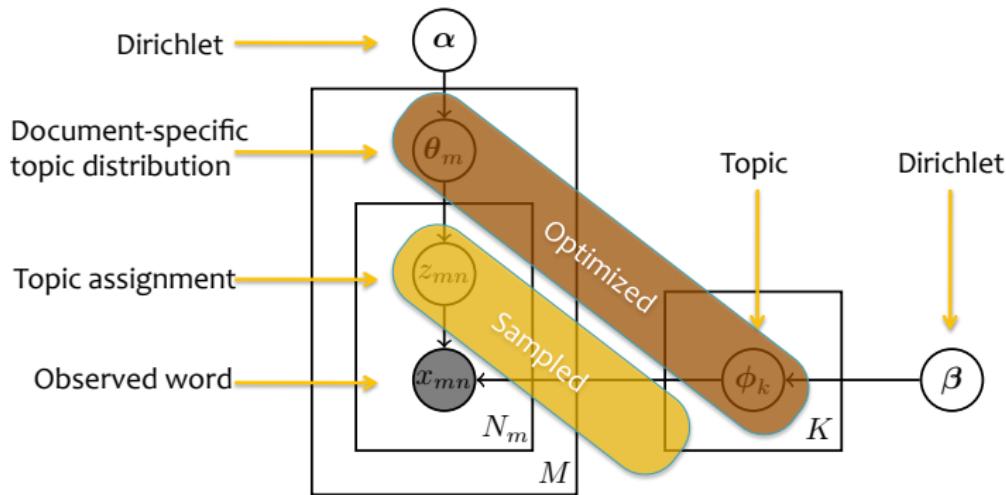


A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



LDA Inference

- Monte Carlo EM

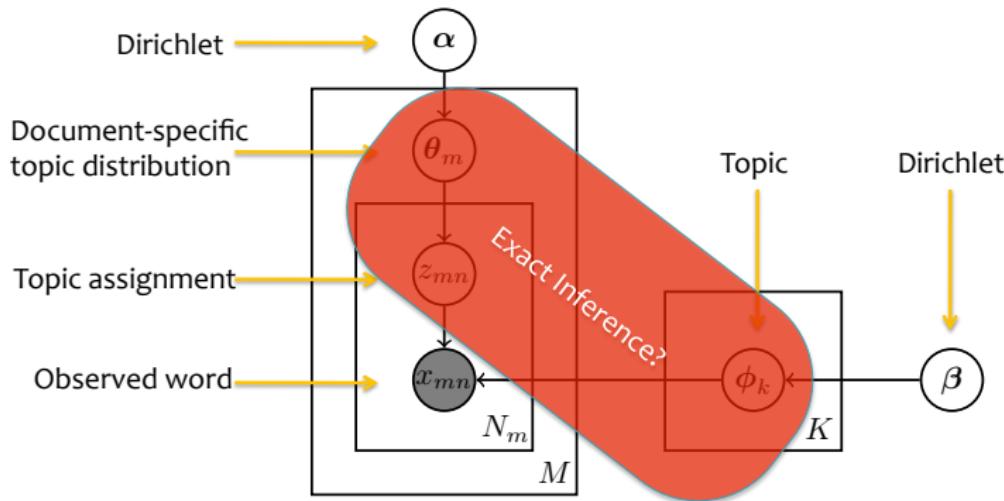


A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



LDA Inference

- Bayesian Approach

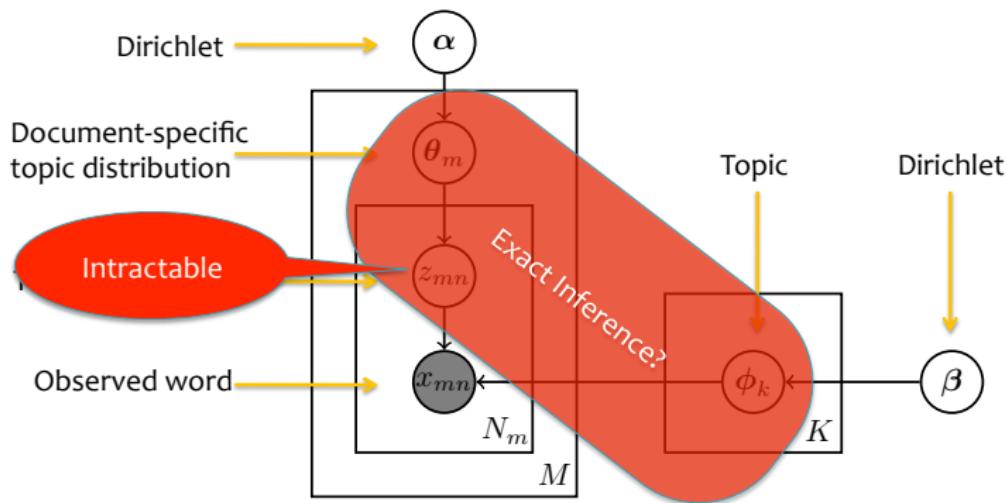


A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



LDA Inference

- Bayesian Approach



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



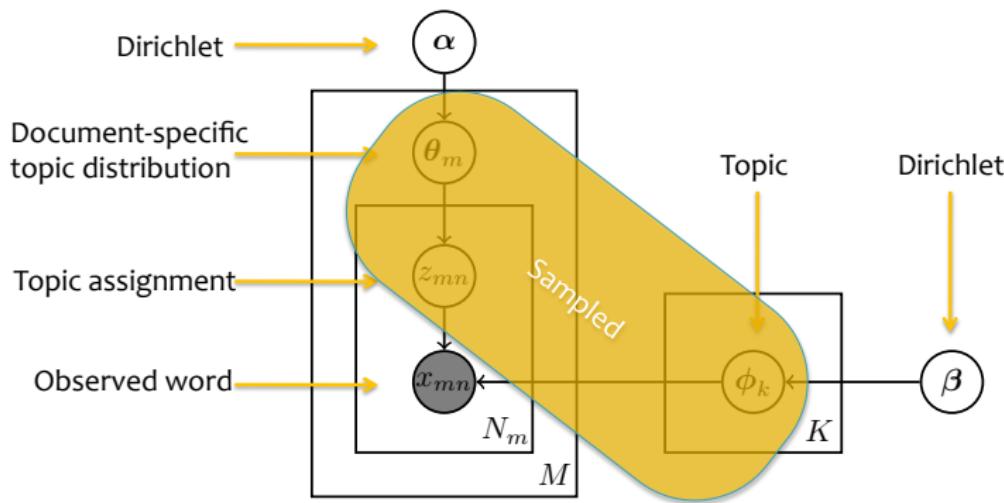
Exact Inference in LDA

- Exactly computing the posterior is intractable in LDA
 - Junction tree algorithm: exact inference in general graphical models
 1. “moralization” converts directed to undirected
 2. “triangulation” breaks 4-cycles by adding edges
 3. Cliques arranged into a junction tree
 - Time complexity is exponential in size of cliques
 - LDA cliques will be large (at least $O(\# \text{ topics})$), so complexity is $O(2^{\# \text{ topics}})$
- Exact MAP inference in LDA is NP-hard for a large number of topics (Sontag & Roy, 2011)



LDA Inference

- Explicit Gibbs Sampler

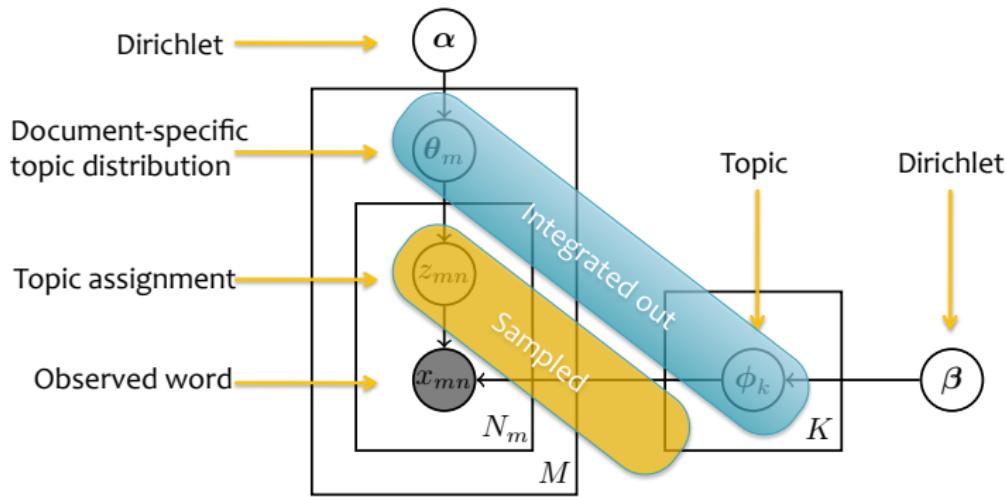


A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



LDA Inference

- Collapsed Gibbs Sampler



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Sampling

Goal:

- Draw samples from the posterior $p(Z|X, \alpha, \beta)$
- Integrate out topics ϕ and document-specific distribution over topics θ

Algorithm:

- While not done...
 - For each document, m :
 - For each word, n :
 - » Resample a single topic assignment using the full conditionals for z_{mn}



Sampling

- What can we do with samples of z_{mn} ?
 - Mean of z_{mn}
 - Mode of z_{mn}
 - Estimate posterior over z_{mn}
 - Estimate of topics ϕ and document-specific distribution over topics θ



Gibbs Sampling for LDA

- Full conditionals

$$p(z_i = k | Z^{-i}, X, \alpha, \beta) = \frac{n_{kt}^{-i} + \beta_t}{\sum_{v=1}^T n_{kv}^{-i} + \beta_v} \cdot \frac{n_{mk}^{-i} + \alpha_k}{\sum_{j=1}^K n_{mj}^{-i} + \alpha_j}$$

where t, m are given by i

n_{kt} = # times topic k appears with type t

n_{mk} = # times topic k appears in document m



Gibbs Sampling for LDA

- Sketch of the derivation of the full conditionals

$$\begin{aligned} p(z_i = k | Z^{-i}, X, \alpha, \beta) &= \frac{p(X, Z | \alpha, \beta)}{p(X, Z^{-i} | \alpha, \beta)} \\ &\propto p(X, Z | \alpha, \beta) \\ &= p(X | Z, \beta) p(Z | \alpha) \\ &= \int_{\Phi} p(X | Z, \Phi) p(\Phi | \beta) d\Phi \int_{\Theta} p(Z | \Theta) p(\Theta | \alpha) d\Theta \\ &= \left(\prod_{k=1}^K \frac{B(\vec{n}_k + \beta)}{B(\beta)} \right) \left(\prod_{m=1}^M \frac{B(\vec{n}_m + \alpha)}{B(\alpha)} \right) \\ &= \frac{n_{kt}^{-i} + \beta_t}{\sum_{v=1}^T n_{kv}^{-i} + \beta_v} \cdot \frac{n_{mk}^{-i} + \alpha_k}{\sum_{j=1}^K n_{mj}^{-i} + \alpha_j} \\ &\quad \text{where } t, m \text{ are given by } i \end{aligned}$$



Dirichlet-Multinomial Model

- The Dirichlet is conjugate to the Multinomial

$\phi \sim \text{Dir}(\beta)$	[draw distribution over words]
For each word $n \in \{1, \dots, N\}$	
$x_n \sim \text{Mult}(1, \phi)$	[draw word]

- The posterior of ϕ is $p(\phi|X) = \frac{p(X|\phi)p(\phi)}{P(X)}$
- Define the count vector \mathbf{n} such that n_t denotes the number of times word t appeared
- Then the posterior is also a Dirichlet distribution:
 $p(\phi|X) \sim \text{Dir}(\beta + \mathbf{n})$



Dirichlet-Multinomial Model

- Why conjugacy is so useful

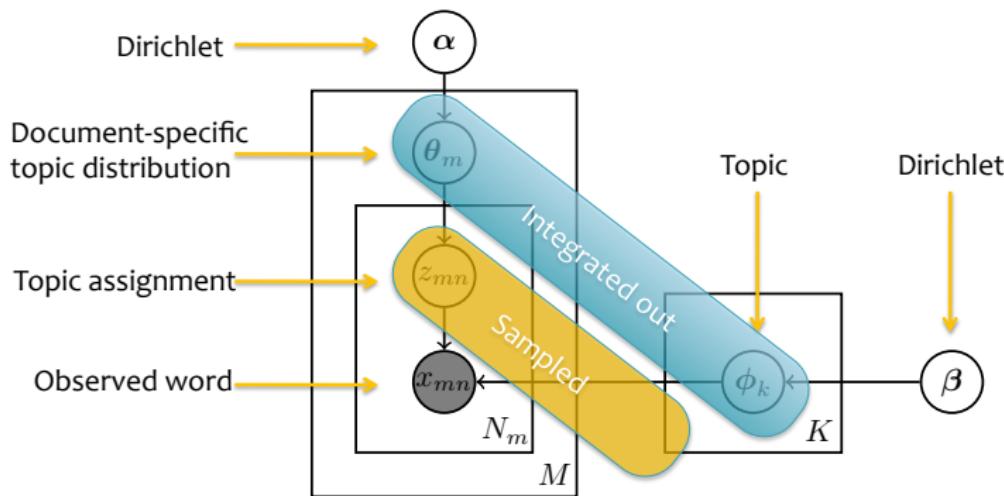
$$\begin{aligned} p(X|\boldsymbol{\alpha}) &= \int_{\phi} p(X|\vec{\phi})p(\vec{\phi}|\boldsymbol{\alpha}) d\phi \\ &= \int_{\phi} \left(\prod_{v=1}^V \phi_v^{n_v} \right) \left(\frac{1}{B(\boldsymbol{\alpha})} \prod_{v=1}^V \phi_v^{\alpha_v - 1} \right) d\phi \\ &= \frac{1}{B(\boldsymbol{\alpha})} \int_{\phi} \prod_{v=1}^V \phi_v^{n_v + \alpha_v - 1} d\phi \\ &= \frac{1}{B(\boldsymbol{\alpha})} \int_{\phi} \frac{B(\vec{n} + \boldsymbol{\alpha})}{B(\vec{n} + \boldsymbol{\alpha})} \prod_{v=1}^V \phi_v^{n_v + \alpha_v - 1} d\phi \\ &= \frac{B(\vec{n} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \underbrace{\int_{\phi} \frac{1}{B(\vec{n} + \boldsymbol{\alpha})} \prod_{v=1}^V \phi_v^{n_v + \alpha_v - 1} d\phi}_{Dir(\vec{n} + \boldsymbol{\alpha})} \\ &= \frac{B(\vec{n} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \end{aligned}$$

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



LDA Inference

- Collapsed Gibbs Sampler



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Gibbs Sampling for LDA

Algorithm

```
// initialisation
zero all count variables,  $n_m^{(k)}, n_m, n_k^{(t)}, n_k$ 
for all documents  $m \in [1, M]$  do
    for all words  $n \in [1, N_m]$  in document  $m$  do
        sample topic index  $z_{m,n}=k \sim \text{Mult}(1/K)$ 
        increment document-topic count:  $n_m^{(k)} += 1$ 
        increment document-topic sum:  $n_m += 1$ 
        increment topic-term count:  $n_k^{(t)} += 1$ 
        increment topic-term sum:  $n_k += 1$ 
```

Figure from Heinrich (2008)

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Gibbs Sampling for LDA

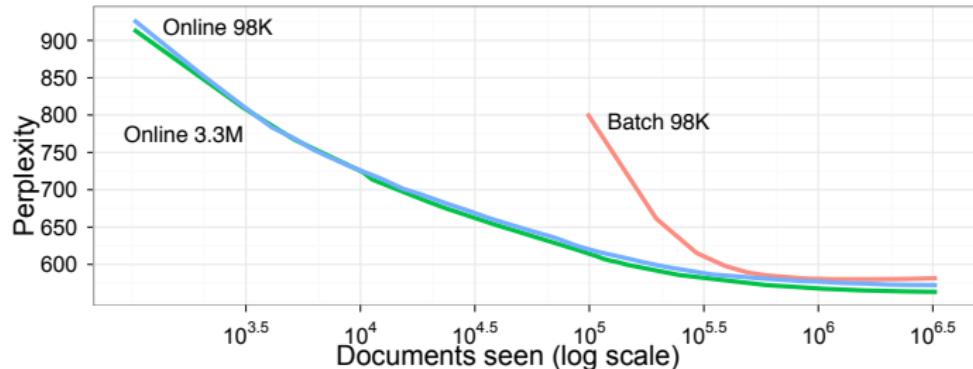
Algorithm

```
// Gibbs sampling over burn-in period and sampling period
while not finished do
    for all documents  $m \in [1, M]$  do
        for all words  $n \in [1, N_m]$  in document  $m$  do
            // for the current assignment of  $k$  to a term  $t$  for word  $w_{m,n}$ :
            decrement counts and sums:  $n_m^{(k)} -= 1; n_m -= 1; n_k^{(t)} -= 1; n_k -= 1$ 
            // multinomial sampling acc. to Eq. 78 (decrements from previous step):
            sample topic index  $\tilde{k} \sim p(z_i | \vec{z}_{\neg i}, \vec{w})$ 
            // for the new assignment of  $z_{m,n}$  to the term  $t$  for word  $w_{m,n}$ :
            increment counts and sums:  $n_m^{(\tilde{k})} += 1; n_m += 1; n_{\tilde{k}}^{(t)} += 1; n_{\tilde{k}} += 1$ 
```

Figure from Heinrich (2008)

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Online Variational Inference for LDA



Documents analyzed	2048	4096	8192	12288	16384	32768	49152	65536
Top eight words	systems road made service announced national west language	systems health communication service billion language care road	service systems health companies market communication company billion	service systems companies business company billion health industry	service systems companies business company billion market industry	business service companies industry company management systems services	business service companies industry company services company management	business industry service companies services company management public



Content

3

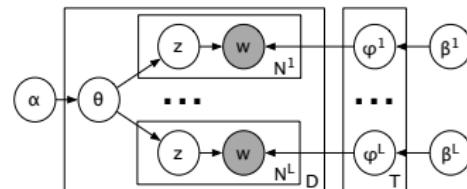
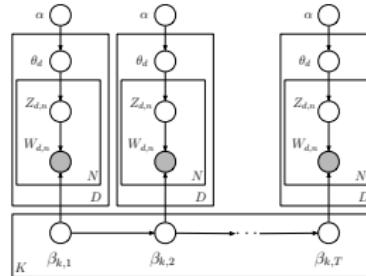
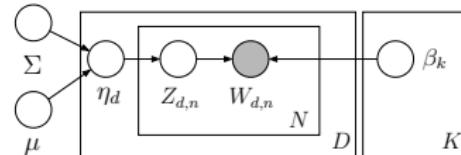
Topic Modeling

- Latent Dirichlet Allocation
- LDA Inference & Learning
- Extentions of LDA



Extensions to the LDA Model

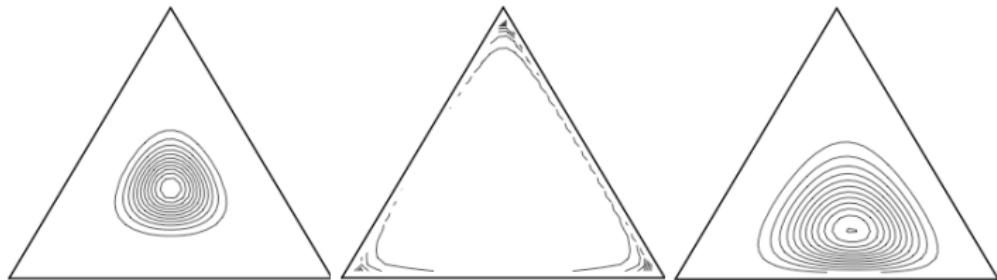
- Correlated topic models
 - Logistic normal prior over topic assignments
 - Dynamic topic models
 - Learns topic changes over time
 - Polylingual topic models
 - Learns topics aligned across multiple languages
- ...



A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Correlated Topic Models



- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

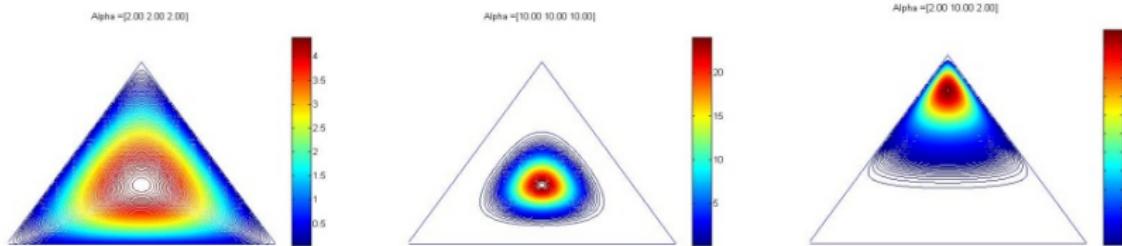
Slide from David Blei, MI 55, 2012

81

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Correlated Topic Models



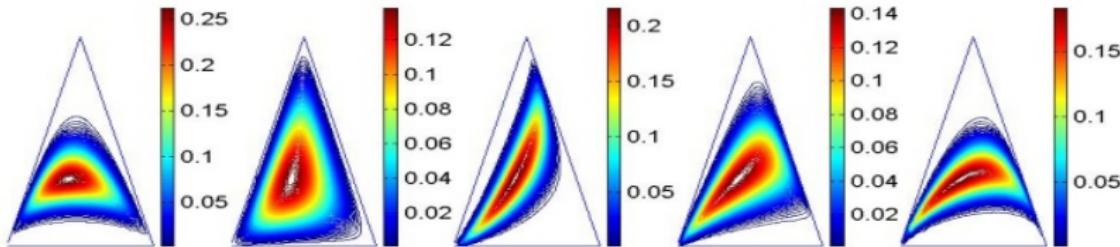
- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

Slide from David Blei MI SS 2012

82

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Correlated Topic Models



- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

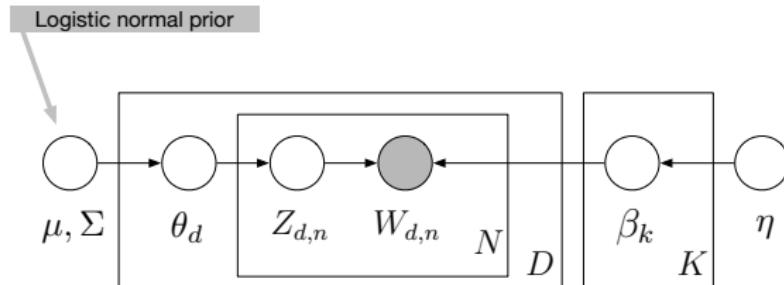
$$\begin{aligned} X &\sim \mathcal{N}_K(\mu, \Sigma) \\ \theta_i &\propto \exp\{x_i\}. \end{aligned}$$

Slide from David Blei MI SS 2012

83

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Correlated Topic Models



- Draw topic proportions from a logistic normal
- This allows topic occurrences to exhibit correlation.
- Provides a “map” of topics and how they are related
- Provides a better fit to text data, but computation is more complex

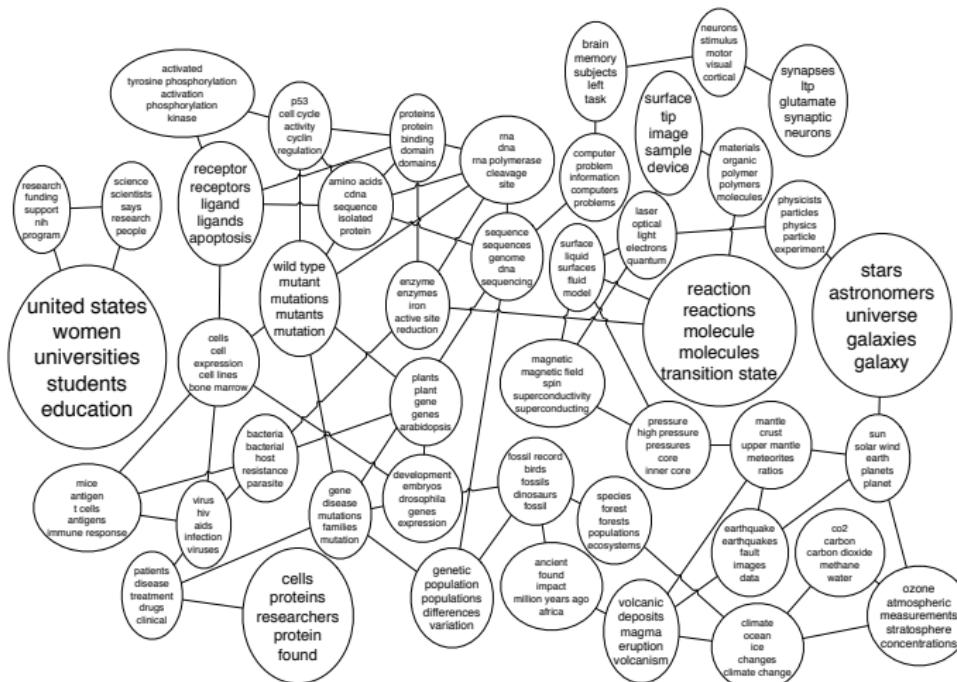
Slide from David Blei, MI 55, 2012

84

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Correlated Topic Models



Slide from David Blei MI SS 2012

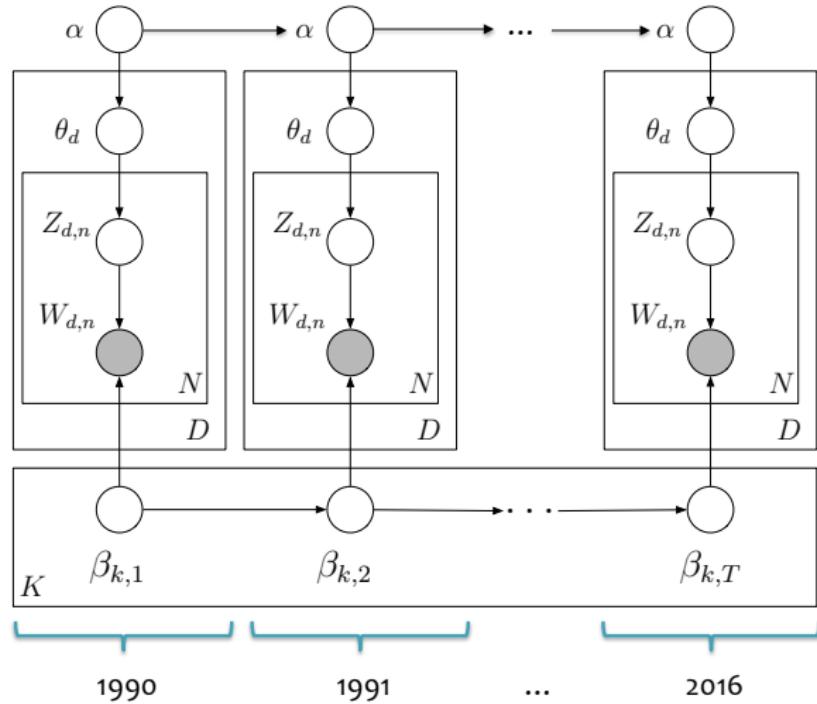
85

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling

Dynamic Topic Models

High-level idea:

- Divide the documents up by year
- Start with a separate topic model for each year
- Then add a dependence of each year on the previous one





Dynamic Topic Models

1789



My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

Inaugural addresses

2009



AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

- LDA assumes that the order of documents does not matter.
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- Further, we may want to track how language changes over time.
- Dynamic topic models let the topics *drift* in a sequence.

Slide from David Blei MI SS 2012

87

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Dynamic Topic Models

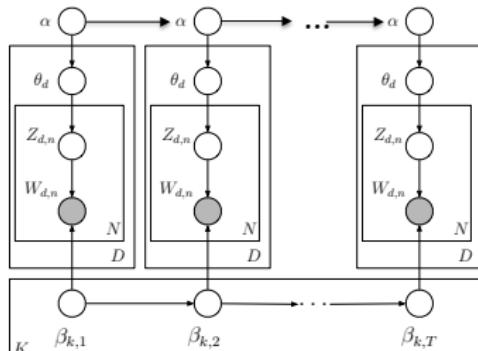
Generative Story

1. Draw topics $\beta_t | \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$. ← Logistic-normal priors
2. Draw $\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$. ←
3. For each document:

 - (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$ ←
 - (b) For each word:
 - i. Draw $Z \sim Mult(\pi(\eta))$. ←
 - ii. Draw $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$. ←

The π function maps from the natural parameters to the mean parameters:

$$\pi(\beta_{k,t})_w = \frac{\exp(\hat{\beta}_{k,t,w})}{\sum_w \exp(\hat{\beta}_{k,t,w})}.$$



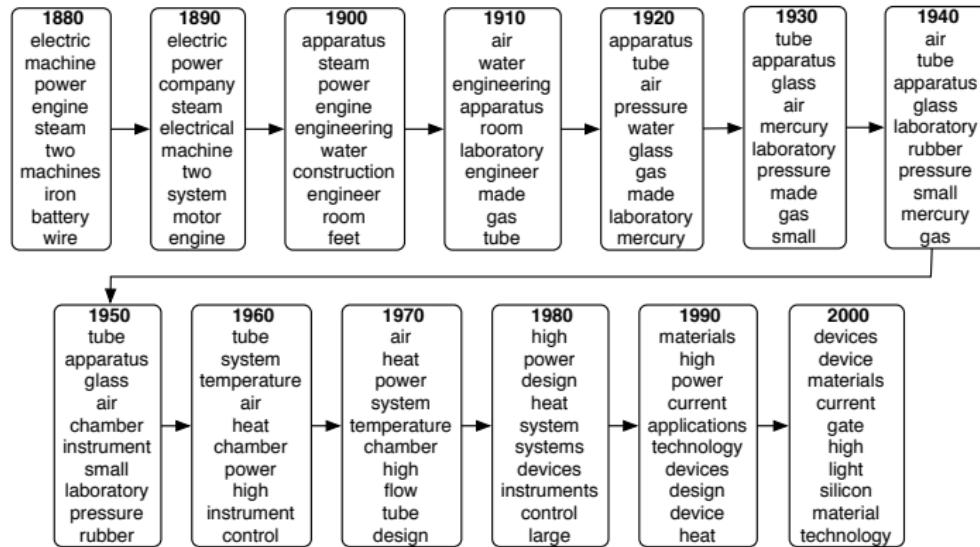
88

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Dynamic Topic Models

Top ten most likely words in a “drifting” topic shown at 10-year increments

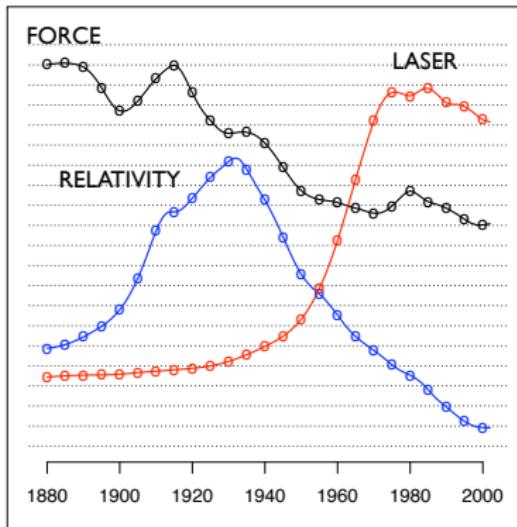




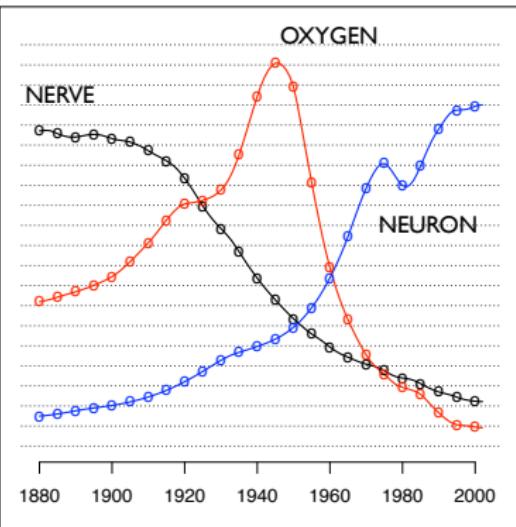
Dynamic Topic Models

Posterior estimate of **word frequency as a function of year** for three words each in two separate topics:

"Theoretical Physics"

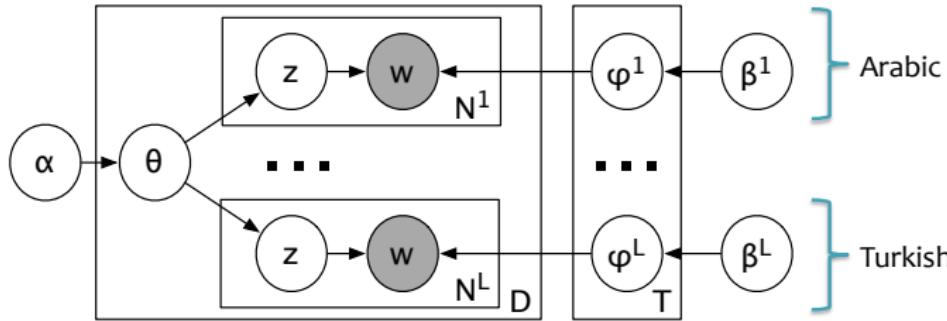


"Neuroscience"



Polylingual Topic Models

- **Data Setting:** Comparable versions of each document exist in multiple languages (e.g. the Wikipedia article for “Barak Obama” in twelve languages)
- **Model:** Very similar to LDA, except that the topic assignments, z , and words, w , are sampled separately for each language.



91



Polylingual Topic Models

Topic 1 (twelve languages)

- CY sadwrn blaned gallair at lloeren mytholeg
- DE space nasa sojus flug mission
- EL διαστημικό sts nasa αγγλ small
- EN **space mission launch satellite nasa spacecraft**
- FA فضایی ماموریت ناسا مدار فضانورد ماهواره
- FI sojuz nasa apollo ensimmäinen space lento
- FR spatiale mission orbite mars satellite spatial
- HE החלל הארץ חלל כדור א תוכנית
- IT spaziale missione programma space sojuz stazione
- PL misja kosmicznej stacji misji space nasa
- RU космический союз космического спутник станции
- TR uzay soyuz ay uzaya salyut sovyetler



Polylingual Topic Models

Topic 2 (twelve languages)

- CY sbaen madrid el la josé sbaeneg
- DE de spanischer spanischen spanien madrid la
- EL ισπανίας ισπανία de ισπανός ντε μαδρίτη
- EN **de spanish spain la madrid y**
- FA ترین de اسپانیا اسپانیایی کوبا مادرید
- FI espanja de espanjan madrid la real
- FR espagnol espagne madrid espagnole juan y
- HE ספרד ספרדית זה מדריד הספרדית קובה
- IT de spagna spagnolo spagnola madrid el
- PL de hiszpański hiszpanii la juan y
- RU де мадрид испании испания испанский de
- TR ispanya ispanyol madrid la küba real



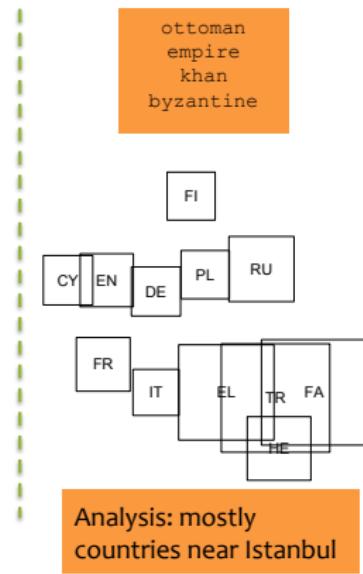
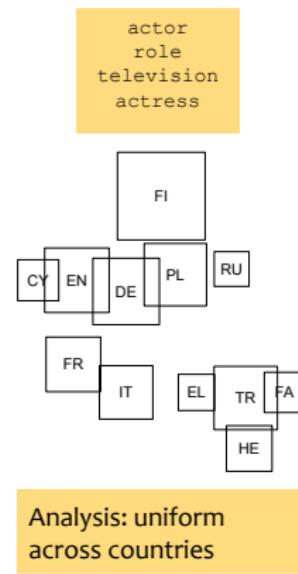
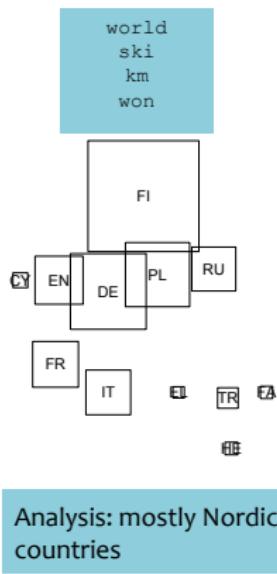
Polylingual Topic Models

Topic 3 (twelve languages)

- CY bardd gerddi iaith beirdd fardd gymraeg
- DE dichter schriftsteller literatur gedichte gedicht werk
- EL ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
- EN **poet poetry literature literary poems poem**
- FA شاعر شعر ادبیات فارسی ادبی آثار
- FI runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
- FR poète écrivain littérature poésie littéraire ses
- HE משורר ספרות שירה סופר שירים המשורר
- IT poeta letteratura poesia opere versi poema
- PL poeta literatury poezji pisarz in jego
- RU поэт его писатель литературы поэзии драматург
- TR şair edebiyat şiir yazar edebiyatı adlı

Polylingual Topic Models

Size of each square represents proportion of tokens assigned to the specified topic.





Supervised LDA

- LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?
- Many data are paired with **response variables**.
 - User reviews paired with a number of stars
 - Web pages paired with a number of “likes”
 - Documents paired with links to other documents
 - Images paired with a category
- **Supervised LDA** are topic models of documents and responses.
They are fit to find topics predictive of the response.

Slide from David Blei, MI 55, 2012

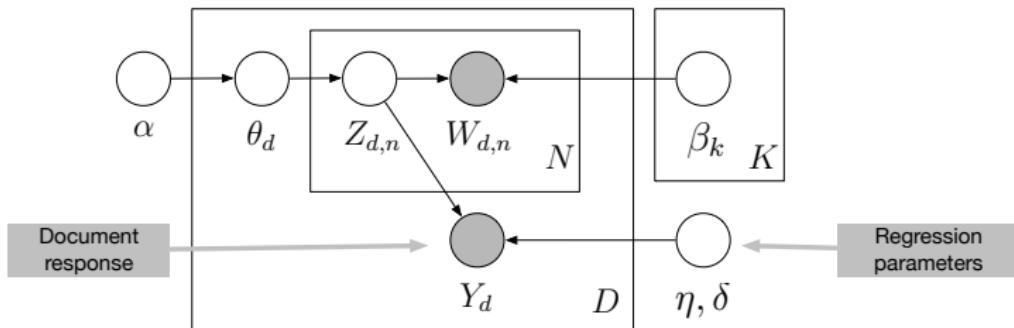
96

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling





Supervised LDA



- ① Draw topic proportions $\theta | \alpha \sim \text{Dir}(\alpha)$.
- ② For each word
 - Draw topic assignment $z_n | \theta \sim \text{Mult}(\theta)$.
 - Draw word $w_n | z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$.
- ③ Draw response variable $y | z_{1:N}, \eta, \sigma^2 \sim N(\eta^\top \bar{z}, \sigma^2)$, where

$$\bar{z} = (1/N) \sum_{n=1}^N z_n.$$

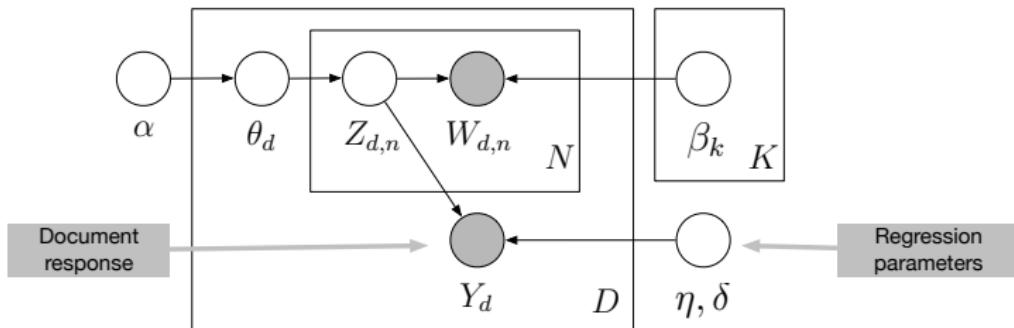
Slide from David Blei MI SS 2012

97

A slide from: Matt Gormley, Carnegie-Mellon University 10-710 (2016), Lecture 20: Topic Modelling



Supervised LDA



- Fit sLDA parameters to documents and responses.
This gives: topics $\beta_{1:K}$ and coefficients $\eta_{1:K}$.
- Given a new document, predict its response using the expected value:

$$\mathbb{E}[Y|w_{1:N}, \alpha, \beta_{1:K}, \eta, \sigma^2] = \eta^\top \mathbb{E}[\bar{Z}|w_{1:N}]$$

- This blends generative and discriminative modeling.

Slide from David Blei MI SS 2012

98



What if we don't know the number of topics, K , ahead of time?

Take 10-708 to learn **Bayesian Nonparametrics**

- New modeling constructs:
 - Chinese Restaurant Process (Dirichlet Process)
 - Indian Buffet Process
- e.g. an **infinite number of topics** in a finite amount of space



Summary: Topic Modeling

- **The Task of Topic Modeling**
 - Topic modeling enables the **analysis of large** (possibly unannotated) **corpora**
 - Applicable to more than just bags of words
 - Extrinsic evaluations are often appropriate for these unsupervised methods
- **Constructing Models**
 - LDA is comprised of **simple building blocks** (Dirichlet, Multinomial)
 - LDA itself can act as a building block **for other models**
- **Approximate Inference**
 - Many different approaches to inference (and learning) can be applied to the same model

100



Content

- 1 ML Recap
- 2 Sampling Basics
- 3 Topic Modeling