Statistical Learning Models for Text and Graph Data
Topic Models

Yangqiu Song

Hong Kong University of Science and Technology

yqsong@cse.ust.hk

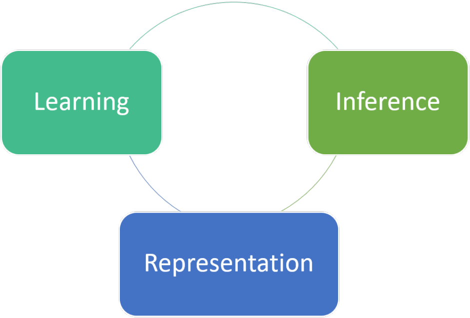
October 30, 2019

$*$Contents are based on materials created by Noah Smith, Xiaojin (Jerry) Zhu, Chengxiang Zhai, David Mackay, Yoav Goldberg

Reference Content

- Noah Smith. CSE 517: Natural Language Processing
  https://courses.cs.washington.edu/courses/cse517/16wi/
- Xiaojin (Jerry) Zhu. CS 769: Advanced Natural Language Processing.
  http://pages.cs.wisc.edu/~jerryzhu/cs769.html
- Yoav Goldberg. Introduction to Natural Language Processing.
  http://u.cs.biu.ac.il/~89-680/

Course Organization



- Representation: language models, word embeddings, topic models, knowledge graphs
- Learning: supervised learning,unsupervised learning, semi-supervised learning, distant supervision, indirect supervision, sequence models, deep learning, optimization techniques
- Inference: constraint modeling, joint inference, search algorithms

Overview

1. Language Models: Recap
2. Topic Models
3. Probabilistic Latent Semantic Analysis (PLSA)
4. Latent Dirichlet Allocation (LDA)
   - Motivation: Bayesian Modeling
   - Background of Monte Carlo Methods
     - Important Sampling
     - Rejection Sampling
     - Metropolis Methods
     - Gibbs Sampling
     - Sampling for EM Algorithm
   - Collapsed Gibbs Sampling for LDA

Overview

Gibbs Sampling

- In the general case of a system with K variables, a single iteration involves sampling one parameter at a time:
  - $x_1^{(t+1)} \sim P(x_1 | x_2^{(t)}, x_3^{(t)}, \ldots, x_K^{(t)})$
  - $x_2^{(t+1)} \sim P(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_K^{(t)})$
  - $x_3^{(t+1)} \sim P(x_3 | x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_K^{(t)})$
  - ...
  - $x_K^{(t+1)} \sim P(x_K | x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{K-1}^{(t+1)})$
- Denote $\mathbf{x}_{\setminus k}^{(t)} = \{x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{k-1}^{(t+1)}, x_{k+1}^{(t)}, \ldots, x_K^{(t)}\}$
- Gibbs sampling can be viewed as a Metropolis method

$$
\begin{aligned}
a_G &= \frac{P^*(\mathbf{x}')Q(\mathbf{x}^{(t)}|x')}{P^*(\mathbf{x}^{(t)})Q(\mathbf{x}'|\mathbf{x}^{(t)})} = \frac{P(\mathbf{x}')P(x_k^{(t)}|\mathbf{x}'_{\setminus k})}{P(\mathbf{x}^{(t)})P(x'_k|\mathbf{x}_{\setminus k}^{(t)})} \\
&= \frac{P(x'_k|\mathbf{x}'_{\setminus k})P(\mathbf{x}'_{\setminus k})P(x_k^{(t)}|\mathbf{x}'_{\setminus k})}{P(x_k^{(t)}|\mathbf{x}_{\setminus k}^{(t)})P(\mathbf{x}_{\setminus k}^{(t)})P(x'_k|\mathbf{x}_{\setminus k}^{(t)})} \overset{\mathbf{x}'_{\setminus k}=\mathbf{x}_{\setminus k}^{(t)}}{=} \frac{P(x'_k|\mathbf{x}'_{\setminus k})P(\mathbf{x}'_{\setminus k})P(x_k^{(t)}|\mathbf{x}'_{\setminus k})}{P(x_k^{(t)}|\mathbf{x}'_{\setminus k})P(\mathbf{x}'_{\setminus k})P(x'_k|\mathbf{x}'_{\setminus k})} = 1
\end{aligned}
$$

- The samples are always accepted

Example of Gibbs Sampling

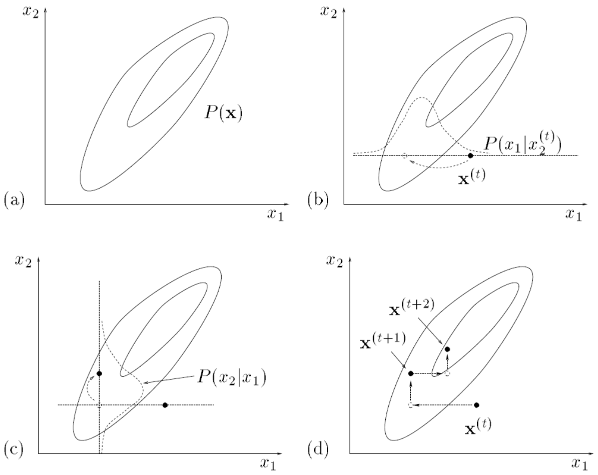


Figure 9. Gibbs sampling. (a) The joint density $P(\mathbf{x})$ from which samples are required. (b) Starting from a state $\mathbf{x}^{(t)}$, x_1 is sampled from the conditional density $P(x_1|x_2^{(t)})$. (c) A sample is then made from the conditional density $P(x_2|x_1)$. (d) A couple of iterations of Gibbs sampling.

Overview

Mixture Models

$$\mathcal{J}(\Theta^t) = \sum_{m=1}^{M} \log \sum_{z_m} P(\mathbf{x}_m, z_m | \Theta^t)$$

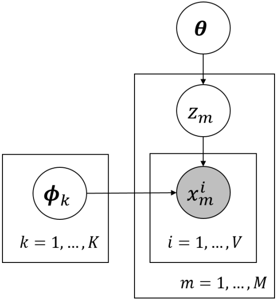


Figure: Mixture Models

EM Algorithm and Sampling

- Change Sum to Integral (to be general and better illustrate the idea)

$$
\begin{aligned}
\mathcal{J}(\Theta^t) &= \sum_{m=1}^{M} \log \int_{\mathbf{z}} P(\mathbf{x}_m, \mathbf{z}|\Theta^t) \\
&= \sum_{m=1}^{M} \log \int_{\mathbf{z}} q_{\mathbf{x}_m,\mathbf{z}}(\Theta) \frac{P(\mathbf{x}_m,\mathbf{z}|\Theta^t)}{q_{\mathbf{x}_m,\mathbf{z}}(\Theta)} \\
&\geq \sum_{m=1}^{M} \int_{\mathbf{z}} q_{\mathbf{x}_m,\mathbf{z}}(\Theta) \log \frac{P(\mathbf{x}_m,\mathbf{z}|\Theta^t)}{q_{\mathbf{x}_m,\mathbf{z}}(\Theta)} \\
&\doteq Q(\Theta, \Theta^t)
\end{aligned}
$$

  where $\int_{\mathbf{z}} q_{\mathbf{x}_m,\mathbf{z}}(\Theta) = 1$ is some distribution
- In E-step, we solve $q_{\mathbf{x}_m,\mathbf{z}}(\Theta) = P(\mathbf{z}|\mathbf{x}_m, \Theta^t)$
- In M-step, we optimize $Q(\Theta^t, \Theta) = \sum_{m=1}^{M} \int_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}_m, \Theta^t) \log P(\mathbf{x}_m, \mathbf{z}|\Theta) + Const$ w.r.t. Θ
- With sampling methods, we can approximate this M-step by a finite sum over samples $\mathbf{z}^r$ from $P(\mathbf{z}^r|\mathbf{x}_m, \Theta^t)$

$$
Q(\Theta^t, \Theta) \approx \sum_{m=1}^{M} \frac{1}{R} \sum_{\mathbf{z}^r \sim P(\mathbf{z}^r|\mathbf{x}_m, \Theta^t)} \log P(\mathbf{x}_m, \mathbf{z}^r|\Theta) + Const
$$

- This procedure is called Monte Carlo EM Algorithm

EM Algorithm and Sampling: Variants

- Monte Carlo EM Algorithm

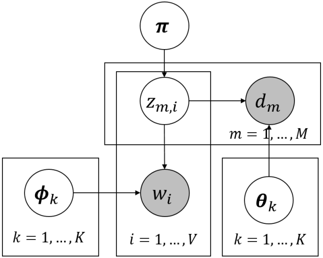
$$Q(\Theta^t, \Theta) \approx \sum_{m=1}^{M} \frac{1}{R} \sum_{\mathbf{z}^r \sim P(\mathbf{z}^r|\mathbf{x}_m, \Theta^t)} \log P(\mathbf{x}_m, \mathbf{z}^r|\Theta) + Const$$

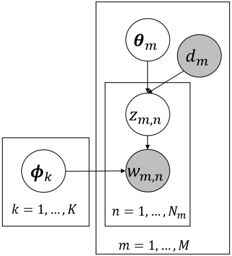
- When we consider a finite mixture model, and draw just one sample at each E-step
  - This is called stochastic EM
  - Here the latent variable $\mathbf{z}$ characterizes which of the K components of the mixture is responsible for generating each data point
  - In the E-step, a sample of $\mathbf{z}$ is taken from the posterior distribution $P(\mathbf{z}|\mathbf{X}, \Theta^t)$ where $\mathbf{X}$ is the data set
  - This effectively makes a hard assignment of each data point to one of the components in the mixture
- If Gibbs sampling is used
  - Instead of drawing a sample from the corresponding conditional distribution, we make a point estimate of the variable given by the maximum of the conditional distribution
  - Then we obtain the iterated conditional modes (ICM) algorithm
  - For finite mixture models, it's similar to K-means

Overview

Alternative Way for PLSA to Generate Texts

$$
\begin{aligned}
P(\mathcal{D},\mathcal{W}) &= \prod_{m=1}^{M}\prod_{i=1}^{N_m}\sum_{k=1}^{K} P(z_{m,i}=k)P(d_m|\boldsymbol{\theta}_k)P(w_i|\boldsymbol{\phi}_k) \\
&= \prod_{m=1}^{M}\prod_{i=1}^{V}\left(\sum_{k=1}^{K} P(z_{m,i}=k)P(d_m|\boldsymbol{\theta}_k)P(w_i|\boldsymbol{\phi}_k)\right)^{c_{d_m}(w_i)}
\end{aligned}
$$





$$
P(\mathcal{D},\mathcal{W}) = \prod_{m=1}^{M}\prod_{i=1}^{V} P(d_m)\left(\sum_{k=1}^{K} P(z_{m,i}=k|\boldsymbol{\theta}_m)P(w_i|\boldsymbol{\phi}_k)\right)^{c_{d_m}(w_i)}
$$

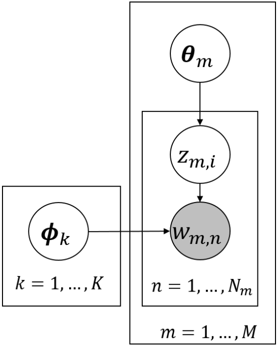


Figure: PLSA

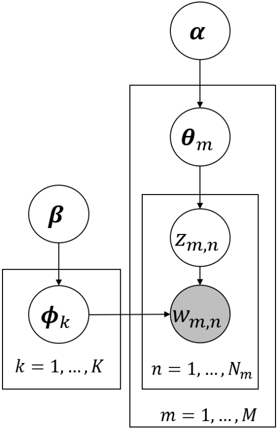


Figure: LDA

Generative Process of Latent Dirichlet Allocation

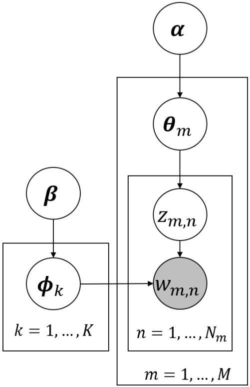


Figure: LDA

- For all clusters/components $k \in [1, K]$:
  - Choose mixture components $\phi_k \sim \mathrm{Dir}(\phi|\beta)$
- For all documents $m \in [1, M]$:
  - Choose $N_m \sim \mathrm{Poisson}(\xi)$
  - Choose mixture probability $\boldsymbol{\theta}_m \sim \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$
  - For all words $n \in [1, N_m]$ in document d_m:
    - Choose a component index $z_{m,n} \sim \mathrm{Mult}(z|\boldsymbol{\theta}_m)$
    - Choose a word $w_{m,n} \sim \mathrm{Mult}(w|\phi_{z_{m,n}})$

A More Detailed Look of LDA

- The probability distribution of the kth latent topic that generates a word is a multinomial distribution

$$
\begin{aligned}
& P(w|z = k, \phi_k) \sim \mathrm{Mult}(w|\phi_k) \\
= \quad & \mathrm{Mult}(w|\phi_{k,1}, \phi_{k,2}, \ldots, \phi_{k,V}) = \prod_{i=1}^{V} \phi_{k,i}^{\delta_{w=v_i}}
\end{aligned}
$$

  where
  - $\phi_k = (\phi_{k,1}, \phi_{k,2}, \ldots, \phi_{k,V})^T \in \mathbb{R}^V$
  - $P(w = v_i|z = k) = P(v_i|z_k) = \phi_{k,i}$
  - The delta function is $\delta_{w=u_i} = 1$ if $w = v_i$; and 0 otherwise
  - We also denote the parameter for the topic mixture probabilities as $\mathbf{\Phi} = (\phi_1, \phi_2, \ldots, \phi_K)^T \in \mathbb{R}^{K\times V}$ where we have K topics

A More Detailed Look of LDA

- The probability distribution that a document generates a topic is:

$$P(z|\boldsymbol{\theta}_m) \sim \mathrm{Mult}(z|\boldsymbol{\theta}_m) = \mathrm{Mult}(w|\theta_{m,1}, \theta_{m,2}, \ldots, \theta_{m,K}) = \prod_{k=1}^{K} \theta_{m,k}^{\delta_{z=k}}$$

where

- $\boldsymbol{\theta}_m = (\theta_{m,1}, \theta_{m,2}, \ldots, \theta_{m,K})^T \in \mathbb{R}^K$
- $P(z = k|d = m) = P(z_k|d_m) = \theta_{m,k}$
- Here we omit the document id in $P(z|\boldsymbol{\theta}_m) = P(z|d_m, \boldsymbol{\theta}_m)$ since $\boldsymbol{\theta}_m$ has the document index m
- We also use $P(z_k|d_m)$ for short rather than the complete form $P(z = k|d = m, \boldsymbol{\theta}_m)$ sometimes
- The delta function is $\delta_{z=k} = 1$ if $z = k$; and 0 otherwise
- We also denote the parameter for the document mixture probabilities as $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)^T \in \mathbb{R}^{M \times K}$ where we have M documents

A More Detailed Look of LDA

- For a full Bayesian view of this mixture model, we add the conjugate Dirichlet priors to both multinomial distributions

$$P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K) \in \mathbb{R}^K$ and

$$P(\boldsymbol{\phi}|\boldsymbol{\beta}) = \mathrm{Dir}(\boldsymbol{\phi}|\boldsymbol{\beta})$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_V) \in \mathbb{R}^V$

A More Detailed Look of LDA

- We formulate the conditional probability of a word $w_{m,n}$ in document d_m given $\boldsymbol{\theta}_m$ and $\boldsymbol{\Phi}$ as:

$$
\begin{aligned}
P(w_{m,n}|\boldsymbol{\theta}_m, \boldsymbol{\Phi}) &= \textstyle\sum_{k=1}^{K} P(w_{m,n}|z_{m,n}=k, \boldsymbol{\Phi})P(z_{m,n}=k|\boldsymbol{\theta}_m) \\
&= \textstyle\sum_{k=1}^{K} P(w_{m,n}|\phi_k)P(z_{m,n}=k|\boldsymbol{\theta}_m)
\end{aligned}
$$

  - This means for each document, we generate a set of topics and each topic generate a word
  - The probability of a word given a document and parameters is also a multinomial distribution

A More Detailed Look of LDA

- Now we can show the data likelihood given a document condition on hyper-parameters:

$$P(\mathcal{W}_m, \mathcal{Z}_m, \boldsymbol{\theta}_m, \boldsymbol{\Phi}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \overbrace{\underbrace{\prod_{n=1}^{N_m} P(w_{m,n}|\phi_k) P(z_{m,n}|\boldsymbol{\theta}_m)}_{\text{word plate}} P(\boldsymbol{\theta}_m|\boldsymbol{\alpha})}^{\text{document plate}} \underbrace{P(\boldsymbol{\Phi}|\boldsymbol{\beta})}_{\text{topic plate}}$$

where $\mathcal{Z}_m = \{z_{m,1}, z_{m,2}, \ldots, z_{m,N_m}\}$ associated with word sequence $\mathcal{W}_m$.

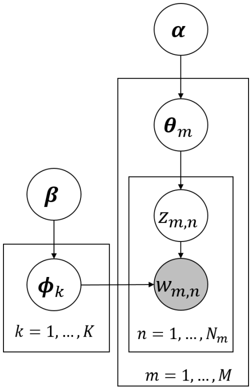


Figure: LDA

- Therefore, the complete likelihood for all documents are given by:

$$P(\mathcal{W}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{m=1}^{M} \int_{\boldsymbol{\Phi}} P(\boldsymbol{\Phi}|\boldsymbol{\beta}) \int_{\boldsymbol{\theta}_m} P(\boldsymbol{\theta}_m|\boldsymbol{\alpha})$$

$$\left(\prod_{n=1}^{N_m} \sum_{k=1}^{K} P(w_{m,n}|\phi_k) P(z_{m,n} = k|\boldsymbol{\theta}_m) \right) \mathrm{d}\boldsymbol{\theta}_m \mathrm{d}\boldsymbol{\Phi}$$

Learning for LDA

- Inference a topic model given a set of training documents involves estimation of document-topic distribution $\boldsymbol{\theta}$'s and topic-word distribution ϕ's
- MAP estimation is intractable due to the interaction between both parameters and also the hyper-parameters
- Thus, approximated methods can be used, such as MCMC (Griffiths and Steyvers (2004)) and variational techniques (Blei et al. (2003))
- Both methods finally produce the estimation of $\boldsymbol{\theta}$'s and ϕ's

Collapsed Gibbs Sampling for LDA

- The collapsed sampling integrate out the parameters of $\boldsymbol{\theta}$'s and $\boldsymbol{\phi}$'s and only sample the latent topic variables by assigning topics to words
- The central idea of Gibbs sampling is to recover the joint marginal (integrating out the parameters) distribution given hyper-parameters:

$$
\begin{aligned}
P(\mathcal{Z}|\mathcal{W},\boldsymbol{\alpha},\boldsymbol{\beta}) &= \frac{P(\mathcal{W},\mathcal{Z}|\boldsymbol{\alpha},\boldsymbol{\beta})}{P(\mathcal{W}|\boldsymbol{\alpha},\boldsymbol{\beta})} \\
&= \frac{\prod_{m=1}^{M}\prod_{n=1}^{N_m} P(w_{m,n},z_{m,n}|\boldsymbol{\alpha},\boldsymbol{\beta})}{\prod_{m=1}^{M}\prod_{n=1}^{N_m}\sum_{k=1}^{K} P(w_{m,n}|\boldsymbol{\alpha},\boldsymbol{\beta})} \\
&= \frac{P(\mathcal{W}|\mathcal{Z},\boldsymbol{\beta})P(\mathcal{Z}|\boldsymbol{\alpha})}{P(\mathcal{W}|\boldsymbol{\alpha},\boldsymbol{\beta})}
\end{aligned}
$$

- Gibbs sampling uses the procedure that samples one variable conditioned on all the other to approximate this distribution

$$
P(z_{m,n}|\mathcal{Z}_{\setminus z_{m,n}},\mathcal{W},\boldsymbol{\alpha},\boldsymbol{\beta})
$$

  to sample a topic associated with a word. The notation $\mathcal{Z}_{\setminus z_{m,n}}$ means the topic assignment set without $z_{m,n}$

Dirichlet Distribution

- Recall the Dirichlet distribution:
$P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \triangleq \frac{\Gamma(\sum_{i=1}^{V}\alpha_i)}{\prod_{i=1}^{V}\Gamma(\alpha_i)} \prod_{i=1}^{V} \theta_i^{\alpha_i - 1} \triangleq \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{i=1}^{V} \theta_i^{\alpha_i - 1}$
  - The "Dirichlet Delta function" $\Delta(\boldsymbol{\alpha})$ is introduced for convenience
  - $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_V)^\top \in \mathbb{R}^V$
  - The Gamma function satisfies $\Gamma(x+1) = x\Gamma(x)$
    - For integer variable, Gamma function is $\Gamma(x) = (x-1)!$
    - For real numbers, it is $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$
- Note that $\int_{\boldsymbol{\theta}} d\boldsymbol{\theta} \prod_{i=1}^{V} \theta_i^{\alpha_i - 1} = \Delta(\boldsymbol{\alpha})$ because
$\int_{\boldsymbol{\theta}} d\boldsymbol{\theta} P(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \int_{\boldsymbol{\theta}} d\boldsymbol{\theta} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{i=1}^{V} \theta_i^{\alpha_i - 1} = 1$

$P(\mathcal{W}|\mathcal{Z},\boldsymbol{\beta})$

- We introduce
  - u_{k,v_i} to represent the count for the word v_i being observed as topic k
- The multinomial distribution of words given topics is

$$\begin{aligned} P(\mathcal{W}|\mathcal{Z},\mathbf{\Phi}) &= \prod_{m=1}^{M}\prod_{n=1}^{N_m} P(w_{m,n}|z_{m,n},\mathbf{\Phi}) \\ &= \prod_{m=1}^{M}\prod_{n=1}^{N_m} \phi_{z_{m,n},w_{m,n}} \\ &= \prod_{k=1}^{K}\prod_{i=1}^{V} \phi_{k,i}^{u_{k,v_i}} \end{aligned}$$

- By integrating out the parameters $\phi_{k,i}$, we can obtain the target distribution $P(\mathcal{W}|\mathcal{Z},\boldsymbol{\beta})$

$$\begin{aligned} P(\mathcal{W}|\mathcal{Z},\boldsymbol{\beta}) &= \int_{\mathbf{\Phi}} P(\mathcal{W}|\mathcal{Z},\mathbf{\Phi})P(\mathbf{\Phi}|\boldsymbol{\beta})\mathrm{d}\mathbf{\Phi} \\ &= \int_{\mathbf{\Phi}} \prod_{k=1}^{K}\frac{1}{\Delta(\boldsymbol{\beta})}\prod_{i=1}^{V}\phi_{k,i}^{\beta_i+u_{k,v_i}-1}\mathrm{d}\boldsymbol{\phi}_k \\ &= \prod_{k=1}^{K}\frac{\Delta(\mathbf{u}_k+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \end{aligned}$$

where we denote $\mathbf{u}_k = (u_{k,v_1}, u_{k,v_2}, \ldots, u_{k,v_V})^T \in \mathbb{R}^V$

$P(\mathcal{Z}|\boldsymbol{\alpha})$

- We introduce
  - $u_{d_m,k}$ represent the count for the topic k for a word being observed in document d_m
- Similarly, we can formulate the multinomial topic distributions given document parameters.

$$\begin{aligned} P(\mathcal{Z}|\boldsymbol{\Theta}) &= \prod_{m=1}^{M}\prod_{n=1}^{N_m} P(z_{m,n}|d_m,\boldsymbol{\theta}_m) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} \theta_{m,z_{m,n}} \\ &= \prod_{m=1}^{M}\prod_{k=1}^{K} \theta_{m,k}^{u_{d_m,k}} \end{aligned}$$

- By integrating out the parameters $\theta_{m,k}$, we can obtain the other target distribution $P(\mathcal{Z}|\boldsymbol{\alpha})$

$$\begin{aligned} P(\mathcal{Z}|\boldsymbol{\alpha}) &= \int_{\boldsymbol{\Theta}} P(\mathcal{Z}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\Phi} \\ &= \int_{\boldsymbol{\Theta}} \prod_{m=1}^{M} \frac{1}{\Delta(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_{m,k}^{\alpha_k+u_{d_m,k}-1}\mathrm{d}\phi_k \\ &= \prod_{m=1}^{M} \frac{\Delta(\mathbf{u}_{d_m}+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \end{aligned}$$

where we denote $\mathbf{u}_{d_m} = (u_{d_m,1}, u_{d_m,2}, \ldots, u_{d_m,K})^T \in \mathbb{R}^K$.

Joint Distribution

- Given

$$P(\mathcal{W}, \mathcal{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\mathcal{W}|\mathcal{Z}, \boldsymbol{\beta})P(\mathcal{Z}|\boldsymbol{\alpha})$$

- The joint distribution is

$$P(\mathcal{W}, \mathcal{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\Delta(\mathbf{u}_k + \boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \cdot \prod_{m=1}^{M} \frac{\Delta(\mathbf{u}_{d_m} + \boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})}$$

Conditional Distribution

$$
\begin{aligned}
& P(z_{m,n}=k|\mathcal{Z}_{\backslash z_{m,n}},\mathcal{W},\boldsymbol{\alpha},\boldsymbol{\beta}) \\
= \; & P(z_{m,n}=k|w_{m,n}=v_i,\mathcal{Z}_{\backslash z_{m,n}},\mathcal{W}_{\backslash w_{m,n}},\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{P(\mathcal{Z},\mathcal{W}|\boldsymbol{\alpha},\boldsymbol{\beta})}{P(\mathcal{Z}_{\backslash z_{m,n}},\mathcal{W}|\boldsymbol{\alpha},\boldsymbol{\beta})}
\end{aligned}
$$

(Using the fact $w_{m,n} \perp \mathcal{W}_{\backslash w_{m,n}} | \mathcal{Z}_{\backslash z_{m,n}}$ and $P(w_{m,n}|\boldsymbol{\beta})=\sum_{i=1}^{K} P(w_{m,n},z_{m,n}|\boldsymbol{\beta})$ is irrelevant to $z_{m,n}$)

$$
= \frac{P(\mathcal{W}|\mathcal{Z},\boldsymbol{\beta})}{P(\mathcal{W}_{\backslash w_{m,n}}|\mathcal{Z}_{\backslash z_{m,n}},\boldsymbol{\beta})P(w_{m,n}|\boldsymbol{\beta})} \cdot \frac{P(\mathcal{Z}|\boldsymbol{\alpha})}{P(\mathcal{Z}_{\backslash z_{m,n}}|\boldsymbol{\alpha})} \propto \frac{\Delta(\mathbf{u}_k+\boldsymbol{\beta})}{\Delta(\mathbf{u}_{k,\backslash z_{m,n}}+\boldsymbol{\beta})} \cdot \frac{\Delta(\mathbf{u}_{d_m}+\boldsymbol{\alpha})}{\Delta(\mathbf{u}_{d_m,\backslash z_{m,n}}+\boldsymbol{\alpha})}
$$

(For $w_{m,n}=v_i$ and current coresponding topic is $z_{m,n}=k^*$)

$$
\begin{aligned}
\propto \; & \frac{\Gamma(u_{k,v_i}+\beta_i+(1-\delta_{k=k^*}))}{\Gamma(\sum_{i=1}^{V}(u_{k,v_i}+\beta_i)+(1-\delta_{k=k^*}))} \cdot \frac{\Gamma(\sum_{i=1}^{V}(u_{k,v_i}+\beta_i)-\delta_{k=k^*})}{\Gamma(u_{k,v_i}+\beta_i-\delta_{k=k^*})} \cdot \\
& \frac{\Gamma(u_{d_m,k}+\alpha_k+(1-\delta_{k=k^*}))}{\Gamma(\sum_{k=1}^{K}(u_{d_m,k}+\alpha_k))} \cdot \frac{\Gamma(\sum_{k=1}^{K}(u_{d_m,k}+\alpha_k)-1)}{\Gamma(u_{d_m,k}+\alpha_k-\delta_{k=k^*})} \quad \left(\text{given } \frac{\Gamma(\sum_{i=1}^{V}\alpha_i)}{\prod_{i=1}^{V}\Gamma(\alpha_i)} = \frac{1}{\Delta(\boldsymbol{\alpha})}\right)
\end{aligned}
$$

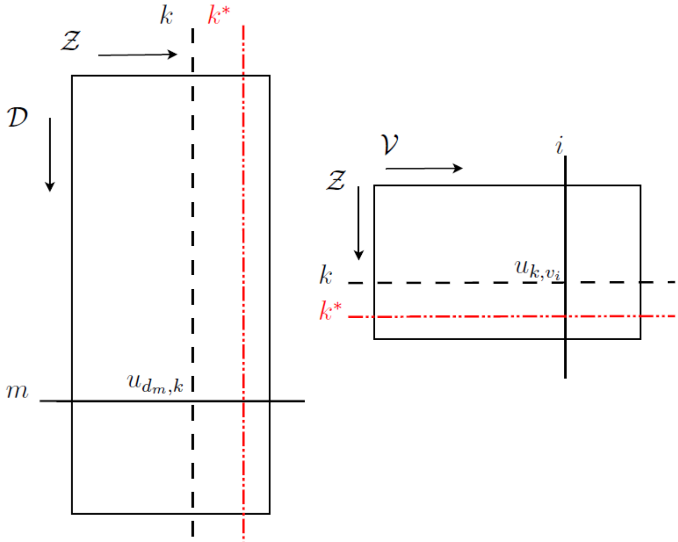
(Using $\Gamma(x+1)=x\Gamma(x)$)

$$
\propto \frac{u_{k,v_i}+\beta_i-\delta_{k=k^*}}{\sum_{i=1}^{V}(u_{k,v_i}+\beta_i)-\delta_{k=k^*}} \cdot \frac{u_{d_m,k}+\alpha_k-\delta_{k=k^*}}{\sum_{k=1}^{K}(u_{d_m,k}+\alpha_k)-1}
$$

($\sum_{k=1}^{K}(u_{d_m,k}+\alpha_k)-1$ is contant for all k's)

$$
\propto \frac{u_{k,v_i}+\beta_i-\delta_{k=k^*}}{\sum_{i=1}^{V}(u_{k,v_i}+\beta_i)-\delta_{k=k^*}} \cdot (u_{d_m,k}+\alpha_k-\delta_{k=k^*})
$$

Matrix Illustration



Sampling Algorithm

Input: Document data set $\mathcal{W}$
repeat
  for all documents $m = 1$ **to** M **do**
    for all words $w_{m,n} = v_i$ where $n = 1$ **to** N_m **do**
      $\diamond$ for the current assignment topic k^* to word $w_{m,n} = v_i$:
        decrement counts: $u_{d_m,k^*} - 1$ and $u_{k^*,v_i} - 1$
      $\diamond$ multinomial sampling topic
      $z_{m,n} = k^{new} \sim p(z_{m,n}|\mathcal{Z}_{\setminus z_{m,n}}, \mathcal{W}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ according to

$$\frac{u_{k,v_i} + \beta_i - \delta_{k=k^*}}{\sum_{i=1}^{V}(u_{k,v_i} + \beta_i) - \delta_{k=k^*}} \cdot (u_{d_m,k} + \alpha_k - \delta_{k=k^*})$$

      $\diamond$ use the new assignment of $z_{m,n}$ to $w_{m,n} = v_i$:
        increment counts: $u_{d_m,k^{new}} + 1$ and $u_{k^{new},v_i} + 1$
    end for
  end for
until Convergence

Parameter Estimation

- Having the sampling counts, we can estimate the posterior of multinomial parameters $\mathbf{\Theta}$ and $\mathbf{\Phi}$ according to the state of the Markov Chain $\mathcal{M} = \{\mathcal{W}, \mathcal{Z}\}$ (MAP estimation)

$$\begin{aligned} & p(\boldsymbol{\theta}_m|\mathcal{M}, \boldsymbol{\alpha}) \\ = \; & \frac{1}{Z_{\boldsymbol{\theta}_m}} \prod_{n=1}^{N_m} p(z_{m,n}|\boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) \\ = \; & \mathrm{Dir}(\boldsymbol{\theta}_m|\mathbf{u}_{d_m} + \boldsymbol{\alpha}) \end{aligned}$$

and

$$\begin{aligned} & p(\boldsymbol{\phi}_k|\mathcal{M}, \boldsymbol{\beta}) \\ = \; & \frac{1}{Z_{\boldsymbol{\phi}_m}} \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(w_{m,n}|z_{m,n} = k, \boldsymbol{\phi}_k) p(\boldsymbol{\phi}_k|\boldsymbol{\beta}) \\ = \; & \mathrm{Dir}(\boldsymbol{\phi}_k|\mathbf{u}_k + \boldsymbol{\beta}) \end{aligned}$$

Parameter Estimation (Cont'd)

- Based on the expectation formulation of Dirichlet distribution $\langle \mathrm{Dir}(\boldsymbol{\alpha}) \rangle = (\alpha_i / \sum_i \alpha_i)_i$, we have:

$$\hat{\theta}_{m,k} = \frac{u_{d_m,k} + \alpha_k}{\sum_{k=1}^{K} (u_{d_m,k} + \alpha_k)}$$

and

$$\hat{\phi}_{k,i} = \frac{u_{k,vi} + \beta_i}{\sum_{i=1}^{V} (u_{k,vi} + \beta_i)}$$

Inference for New Coming Documents

- For a new coming document data set $\tilde{\mathcal{W}}$, we assume that the assigned topic set is $\tilde{\mathcal{Z}}$
- Each word $\tilde{w}_{m,n}$ will be assigned with a topic index $\tilde{z}_{m,n}$ also via Gibbs sampling procedure
- By fixing the training data and parameters $\mathbf{\Theta}$ and $\mathbf{\Phi}$, we first randomly assign a topic to new coming word
- Then, perform sampling based on the following conditional probability:

$$p(\tilde{z}_{m,n} = k | \tilde{w}_{m,n} = v_i, \tilde{\mathcal{Z}}_{\backslash \tilde{z}_{m,n}}, \tilde{\mathcal{W}}_{\backslash \tilde{w}_{m,n}}, \mathcal{M}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$
$$\propto \quad \frac{u_{k,v_i} + \tilde{u}_{k,v_i} + \beta_i - \delta_{k=k^*}}{\sum_{i=1}^{V}(u_{k,v_i} + \tilde{u}_{k,v_i} + \beta_i) - \delta_{k=k^*}} \cdot \frac{\tilde{u}_{\tilde{d}_m,k} + \alpha_k - \delta_{k=k^*}}{\sum_{k=1}^{K}(\tilde{u}_{\tilde{d}_m,k} + \alpha_k) - 1}$$

Inference for New Coming Documents

- If the new coming documents are short, u_{k,v_i} dominates the first term compared with $\tilde{u}_{k,v_i}$, which are randomly assigned
- Thus, repeatedly sampling from this distribution $p(\tilde{z}_{m,n} = k|\cdot)$ and updating $\tilde{u}_{\tilde{d}_m,k}$, topic-word associations are propagated into document-topic association
- For simplicity, we can even omit the topic-word term (Heinrich (2008)):

$$
\begin{aligned}
& p(\tilde{z}_{m,n} = k|\tilde{w}_{m,n} = v_i, \tilde{\mathcal{Z}}_{\backslash \tilde{z}_{m,n}}, \tilde{\mathcal{W}}_{\backslash \tilde{w}_{m,n}}, \mathcal{M}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
\propto \quad & \hat{\phi}_{k,i} \cdot \frac{\tilde{u}_{\tilde{d}_m,k} + \alpha_k - \delta_{k=k^*}}{\sum_{k=1}^{K}(\tilde{u}_{\tilde{d}_m,k} + \alpha_k) - 1}
\end{aligned}
$$

Inference for New Coming Documents

- The topic distribution posterior for new coming documents are:

$$\hat{\tilde{\theta}}_{m,k} = \frac{\tilde{u}_{\tilde{d}_m,k} + \alpha_k}{\sum_{k=1}^{K}(\tilde{u}_{\tilde{d}_m,k} + \alpha_k)}$$

- In practice, we often assume the set of new coming data are much smaller than the training data: $|\tilde{\mathcal{W}}| \ll |\mathcal{W}|$.
- Otherwise, the new data will make the topic-word count distortion as $u_{k,v_i} + \tilde{u}_{k,v_i}$.

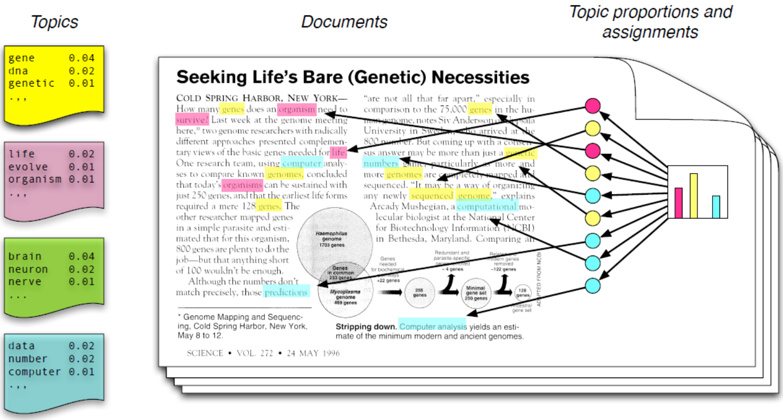
Hyperparameter Estimation

- We can also do hyperparameter estimation using maximum likelihood estimation
  - Refer to (Heinrich (2008))
  - Detailed Dirichlet distribution analysis can be found from (Minka (2000))
    `https://tminka.github.io/papers/dirichlet/minka-dirichlet.pdf`

Variants of LDA

- There are many variants of LDA
  - Online and Incremental Learning
  - Distributed Computing
  - Dynamic Topic Model
  - Author Topic (AT) and Author Recipient Topic (ART) Model
  - Hierarchical Dirichlet Processes
  - Neural Topic Models

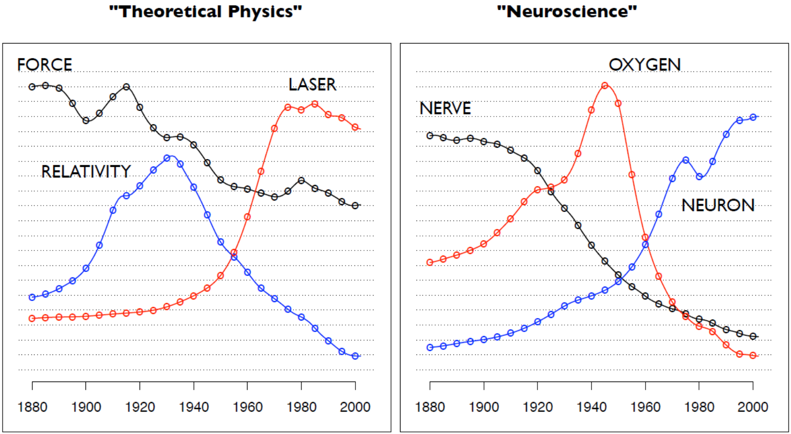
Use of Topic Models



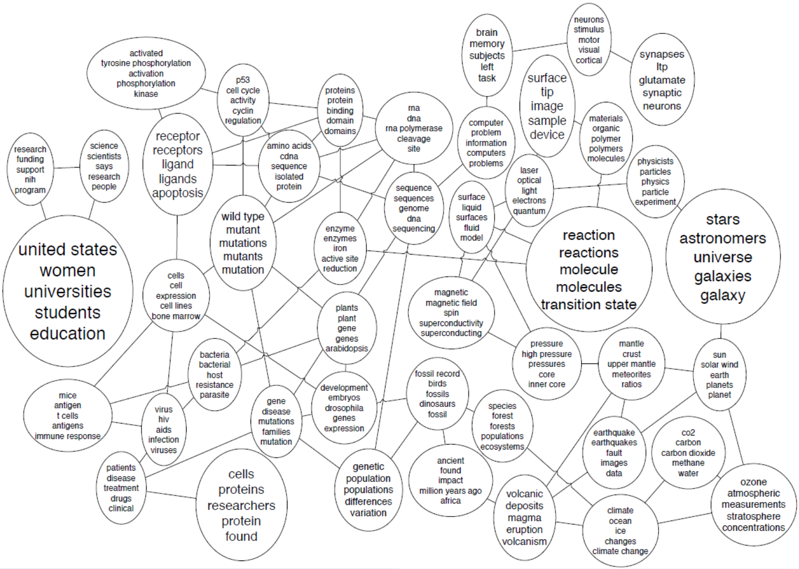
Discover topics from a corpus

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

Model the evolution of topics over time



Model connections between topics



Annotate images

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE

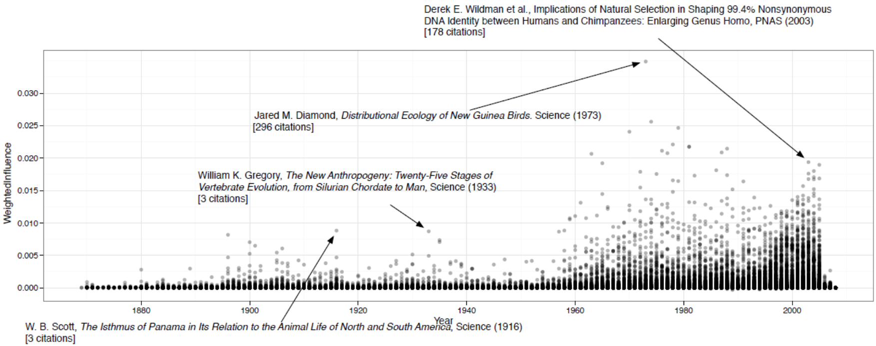
SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

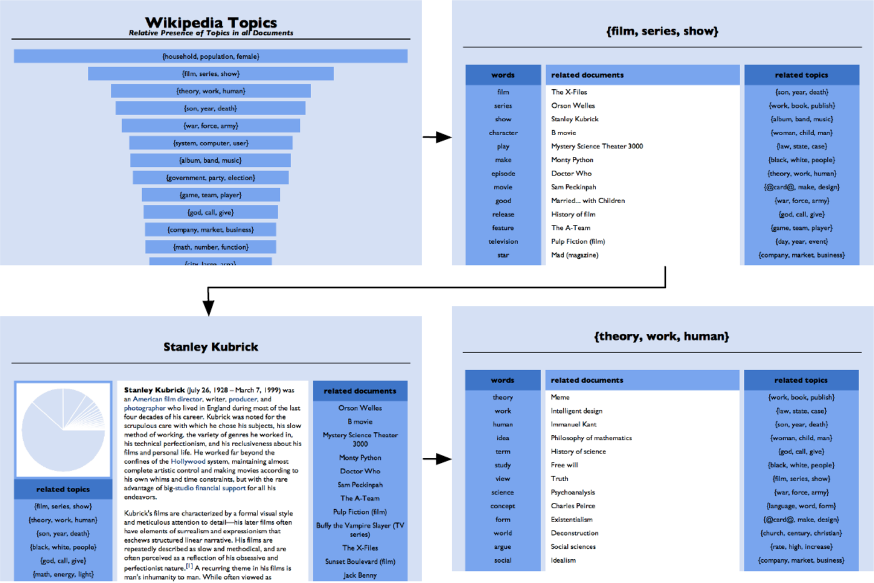
PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

Discover influential articles



Organize and browse large corpora



Further Reading

How to reduce the variance of the gradient in Monte Carlo based variational inference?
Rajesh Ranganath, Sean Gerrish, David M. Blei: Black Box Variational Inference. AISTATS 2014: 814-822. (Ranganath et al. (2014))

References I

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Heinrich, G. (2008). Parameter estimation for text analysis. Technical Report Version 2.4, vsonix GmbH + University of Leipzig, Germany.

Minka, T. P. (2000). Estimating a dirichlet distribution. Technical report.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *AISTATS*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 814–822. JMLR.org.