Statistical Computing

June 1, 2021

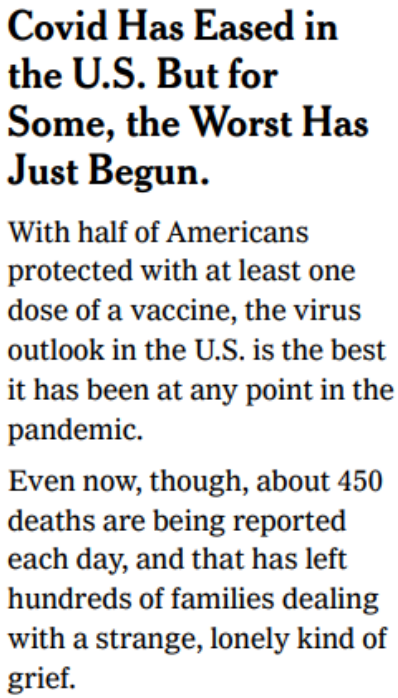
Introduction

- Information from text data.
- Text with topics.
- Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003).
- Variational inference.
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2004).

Text Data

Data information store in the form of news, blogs, Web, pages, scientific articles, and books.

- For example, considering the New York Times, some of the themes might be foreign policy, national affairs, sports.



Covid Has Eased in the U.S. But for Some, the Worst Has Just Begun.

With half of Americans protected with at least one dose of a vaccine, the virus outlook in the U.S. is the best it has been at any point in the pandemic.

Even now, though, about 450 deaths are being reported each day, and that has left hundreds of families dealing with a strange, lonely kind of grief.

Figure: From New York Times 2021/06/01.

Text Data, Cont,

- Electronic Health Records (EHR)

病歷資料 1 (未經處理): lacunar infarction hyperlipidemiahypertension ; ;;;left dysarthria clumsy-hand lacunar infarctionno stroke-related complicationhyperlipidemiahypertensionacute angle closure glaucoma sudden onset of right hand weakness since 7/ 15:00 this 5 uears old male was found having htn and dyslipidemia since last time health exam in april without medical control he has habbit of smoking ppd for 30 years and alcohol 1bt/day for 30 years he was well untill 7/ 15:00 when he felt sudden onset of right hand weakness during driving slurred speech and dizzness were also noted no dysphagia, no aphasia, no double vision, no urine incontinence he went to for help, where ct showed no hemorrhage he was transferred to our er for further treatment

病歷資料 1 (處理後): lacunar infarction dysarthria clumsy hand lacunar infarctionno stroke related angle closure glaucoma sudden onset hand weakness since uears male found htn dyslipidemia since last health exam april without medical control habbit smoking_ppd alcohol bt well untill felt sudden onset hand weakness driving slurred speech dizzness also dysphagia aphasia double vision urine incontinence went help ct hemorrhage transferred er treatment

Figure: Example of EHR Data.

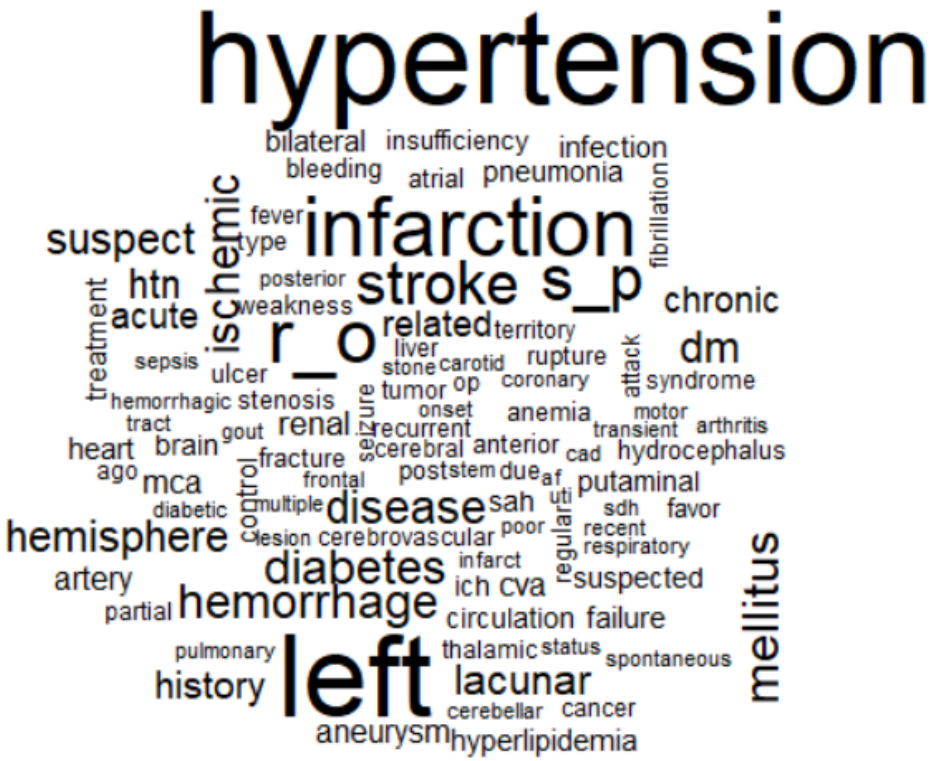
- Electronic Health Records (EHR) with topic

lacunar infarction dysarthria clumsy hand lacunar infarctionno stroke related angle closure glaucoma sudden onset hand weakness since uears male found htn dyslipidemia since last health exam april without medical control habbit smoking_ppd alcohol bt well untill felt sudden onset hand weakness driving slurred speech dizzness also dysphagia aphasia double vision urine incontinence went help ct hemorrhage transferred er treatment

- Transform text into numbers

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0	0	0	0.557438	0	0	0.4339256	0	0	0

Data Visualization



Notation for Text Data

- A word w is an element of dictionary $\{1, ..., V\}$.
- A document $\vec{w}$ is represented by a sequence of N words:
  $\vec{w} = (w_1, ..., w_N)$, $w_n \in \{1, ..., V\}$.
- A corpus $\boldsymbol{D}$ is a collection of D documents: $\boldsymbol{D} = \{\vec{w}^{(1)}, ..., \vec{w}^{(D)}\}$.

Mixture Model with Dirichlet Prior

In mixture model, the probability of N-words document $\vec{w}$ is parameterized by hyper parameters $\{\boldsymbol{\alpha}\}$ of the Dirichlet prior. The probability can be evaluated via conditioning on the latent variable $\vec{\theta}$ as follows.

- $$\begin{aligned} P(\vec{w} \mid \boldsymbol{\alpha}) &= \int P(\vec{w} \mid \vec{\theta}) P(\vec{\theta} \mid \boldsymbol{\alpha}) d\vec{\theta} \\ &= \int P(\vec{w} \mid \vec{\theta}) P(\vec{\theta} \mid \boldsymbol{\alpha}) d\vec{\theta} \\ &= \int \left\{ \prod_{i=1}^{N} P\left(w_i \mid \vec{\theta}\right) \right\} P(\vec{\theta} \mid \boldsymbol{\alpha}) d\vec{\theta}. \end{aligned}$$

- The two-level model does not explicitly consider the topic issue. Each document may represent a mix of topics.

Latent Dirichlet Allocation (LDA)

- "Bag of Words" Models.
- Probabilistic modeling.
- Assume that all the words within a document are exchangeable.
- The count of the words is matter, not the order.
- LDA: Each document is a mixture of topics.
- LDA assumes that a word w is observed, there is an associated topic (latent variable) z.
- Incorporating the (multinomial) distribution of the word over topic, that is $P(w \mid z, \boldsymbol{\beta})$, where the matrix $\boldsymbol{\beta} = [\beta_{ij}]$ parameterizes this distribution and β_{ij} stands for the probability of the j-th word under the i-th topic.

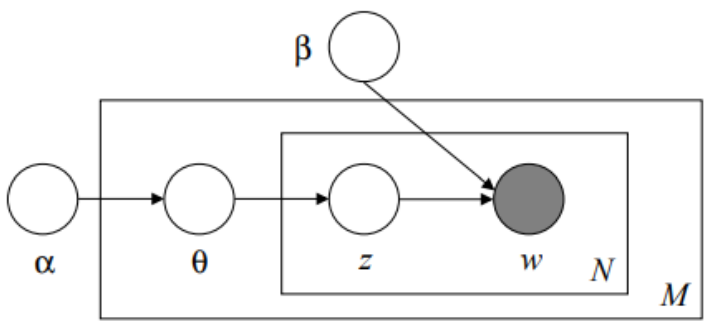


Figure: From Blei, Ng, and Jordan, 2003.

Latent Dirichlet Allocation (LDA), Cont.

LDA is a three-level generative model

- in which there is a topic level between the word level and the belief level and in LDA, $\vec{\theta}$ becomes topic belief.
- LDA assumes that whenever a key word w is observed, there is an associated topic z, hence an N -words document $\vec{w}$ is associated with a topic sequence $\vec{z}$ of length N.
- By conditioning on $\vec{z}$ and conditional independence, the parameterized probability $P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta})$ of seeing $\vec{w}$ can be written as

$$\begin{aligned} P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta}) &= \sum_{\vec{z}} P(\vec{w} \mid \vec{z}, \vec{\alpha}, \boldsymbol{\beta}) P(\vec{z} \mid \vec{\alpha}, \boldsymbol{\beta}) \\ &= \sum_{\vec{z}} P(\vec{w} \mid \vec{z}, \boldsymbol{\beta}) P(\vec{z} \mid \vec{\alpha}) \end{aligned}$$

- By chain rule and conditional independence, we have

$$\begin{aligned} P(\vec{w} \mid \vec{z}, \boldsymbol{\beta}) &= P\left(w_1, \ldots, w_N \mid z_1, \ldots, z_N, \boldsymbol{\beta}\right) \\ &= P\left(w_1 \mid w_2, \ldots, w_N, z_1, \ldots, z_N, \boldsymbol{\beta}\right) P\left(w_2, \ldots, w_N \mid z_1, \ldots, z_N, \boldsymbol{\beta}\right) \\ &= P\left(w_1 \mid z_1, \boldsymbol{\beta}\right) P\left(w_2, \ldots, w_N \mid z_2, \ldots, z_N, \boldsymbol{\beta}\right) \\ &= P\left(w_1 \mid z_1, \boldsymbol{\beta}\right) \cdots P\left(w_N \mid z_N, \boldsymbol{\beta}\right). \end{aligned}$$

Generation of Latent Dirichlet Allocation (LDA)

For each topic k,

- Draw a Dirichlet distribution over words $\beta_k \sim Dir(\eta)$.

For each document d,

- Draw a Dirichlet distribution over topics $\theta_d \sim Dir(\alpha)$.
- For each word $w_{d,n}$
  - Draw a topic $Z_{d,n}$ following from a multinomial distribution (θ_d) with $Z_{d,n} \in [1, K]$.
  - Draw a word $w_{d,n}$ following from a multinomial distribution $(\beta_{Z_{d,n}})$.

Latent Dirichlet Allocation (LDA), Cont.

- Thus, we have

$$P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta}) = \sum_{\vec{z}} \prod_{n=1}^{N} P\left(w_n \mid z_n, \boldsymbol{\beta}\right) P(\vec{z} \mid \vec{\alpha})$$

- For $P(\vec{z} \mid \vec{\alpha})$, by conditioning on $\vec{\theta}$ and conditional independence, we get

$$\begin{aligned} P(\vec{z} \mid \vec{\alpha}) &= \int P(\vec{z} \mid \vec{\theta}, \vec{\alpha}) P(\vec{\theta} \mid \vec{\alpha}) d\vec{\theta} \\ &= \int \prod_{n=1}^{N} P\left(z_n \mid \vec{\theta}\right) P(\vec{\theta} \mid \vec{\alpha}) d\vec{\theta}. \end{aligned}$$

- Plugging the result of $P(\vec{z} \mid \vec{\alpha})$ and exchanging the order of integration and summation yields

$$\begin{aligned} P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta}) &= \sum_{\vec{z}} \prod_{n=1}^{N} P\left(w_n \mid z_n, \boldsymbol{\beta}\right) \int \prod_{n=1}^{N} P\left(z_n \mid \vec{\theta}\right) P(\vec{\theta} \mid \vec{\alpha}) d\vec{\theta} \\ &= \sum_{\vec{z}} \int \prod_{n=1}^{N} P\left(w_n \mid z_n, \boldsymbol{\beta}\right) P\left(z_n \mid \vec{\theta}\right) P(\vec{\theta} \mid \vec{\alpha}) d\vec{\theta} \\ &= \int \left\{ \sum_{\vec{z}} \prod_{n=1}^{N} P\left(w_n \mid z_n, \boldsymbol{\beta}\right) P\left(z_n \mid \vec{\theta}\right) \right\} P(\vec{\theta} \mid \vec{\alpha}) d\vec{\theta} \end{aligned}$$

Likelihood Function of Observed Data

- Note that

$$\sum_{\vec{z}} \prod_{n=1}^{N} P\left(w_n \mid z_n, \boldsymbol{\beta}\right) P\left(z_n \mid \vec{\theta}\right) \quad = \quad \prod_{n=1}^{N} \sum_{z_n} P\left(w_n \mid z_n, \boldsymbol{\beta}\right) P\left(z_n \mid \vec{\theta}\right).$$

- The final form of $P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta})$ can be simplified as

$$\begin{aligned} P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta}) &= \int \left(\prod_{n=1}^{N} \sum_{z_n} P\left(w_n \mid z_n, \boldsymbol{\beta}\right) P\left(z_n \mid \vec{\theta}\right) \right) P(\vec{\theta} \mid \vec{\alpha}) d\vec{\theta} \\ &= \int \left(\prod_{n=1}^{N} \sum_{z} P\left(w_n \mid z, \boldsymbol{\beta}\right) P(z|\vec{\theta}) \right) P(\vec{\theta} \mid \vec{\alpha}) d\vec{\theta} \end{aligned}$$

- Unfortunately, it does not have explicit formula.
- To estimate $\vec{\alpha}, \boldsymbol{\beta}$ from a corpus $\boldsymbol{D} = \left\{\vec{w}^{(1)}, \ldots, \vec{w}^{(D)}\right\}$, maximizing the log likelihood of observed data

$$\ln P(\boldsymbol{D} \mid \vec{\alpha}, \boldsymbol{\beta}) \quad = \quad \sum_{d=1}^{D} \ln P\left(\vec{w}^{(d)} \mid \vec{\alpha}, \boldsymbol{\beta}\right),$$

  is computationally intractable.
- However, the variational inference method can provide a computationally tractable lower bound.

Variational Inference

The idea of the variational inference is to apply the Jensen's inequality to obtain a lower bound of $\ln P(\vec{w} \mid \vec{\alpha}, \beta)$.

- Let $q(\vec{z}, \vec{\theta})$ be a joint probability density function of $\vec{z}, \vec{\theta}$, applying the idea of importance sampling yields

$$\begin{aligned}
\ln P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta}) &= \ln \int \sum_{\vec{z}} P(\vec{w}, \vec{z}, \vec{\theta} \mid \vec{\alpha}, \boldsymbol{\beta}) d\vec{\theta} \\
&= \ln \int \sum_{\vec{z}} \frac{P(\vec{w}, \vec{z}, \vec{\theta} \mid \vec{\alpha}, \beta)}{q(\vec{z}, \vec{\theta})} q(\vec{z}, \vec{\theta}) d\vec{\theta} \\
&\geq \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln \frac{P(\vec{w}, \vec{z}, \vec{\theta} \mid \vec{\alpha}, \beta)}{q(\vec{z}, \vec{\theta})} d\vec{\theta} \\
&= L(\vec{\alpha}, \boldsymbol{\beta}),
\end{aligned}$$

  where the last step is by Jensen's inequality.

- $L(\vec{\alpha}, \boldsymbol{\beta})$ is also referred to as Evidence Lower Bound (ELBO).

Variational Inference, Cont.

- Note that

$$
\begin{aligned}
L(\vec{\alpha}, \boldsymbol{\beta}) &= \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln \frac{P(\vec{w}, \vec{z}, \vec{\theta} \mid \vec{\alpha}, \beta)}{q(\vec{z}, \vec{\theta})} d\vec{\theta} \\
&= \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln \frac{P(\vec{z}, \vec{\theta} \mid \bar{w}, \vec{\alpha}, \beta) P(\vec{w} \mid \vec{\alpha}, \beta)}{q(\vec{z}, \vec{\theta})} d\vec{\theta} \\
&= \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \left(\ln \frac{P(\vec{z}, \vec{\theta} \mid \vec{w}, \vec{\alpha}, \beta)}{q(\vec{z}, \vec{\theta})} + \ln P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta}) \right) d\vec{\theta} \\
&= -\int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln \frac{q(\vec{z}, \vec{\theta})}{P(\vec{z}, \vec{\theta} \mid \bar{w}, \vec{\alpha}, \beta)} d\vec{\theta} + \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta}) d\vec{\theta} \\
&= -KL\{q(\vec{z}, \vec{\theta}) \| P(\vec{z}, \vec{\theta} \mid \vec{w}, \vec{\alpha}, \boldsymbol{\beta})\} + \ln P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta})
\end{aligned}
$$

- As a result,

$$\ln P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta}) = L(\vec{\alpha}, \boldsymbol{\beta}) + KL\{q(\vec{z}, \vec{\theta}) \| P(\vec{z}, \vec{\theta} \mid \vec{w}, \vec{\alpha}, \boldsymbol{\beta})\}).$$

- $\ln P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta})$ is the sum of the evidence lower bound and the KL distance between $q(\vec{z}, \vec{\theta})$ and the true posterior.
- In variational inference, finding a $L(\vec{\alpha}, \boldsymbol{\beta})$ that is as much close to $\ln P(\vec{w} \mid \vec{\alpha}, \boldsymbol{\beta})$ as possible.
- This can be done by finding a (parameterized) $q(\vec{z}, \vec{\theta})$ that has as small KL distance to $P(\vec{z}, \vec{\theta} \mid \vec{w}, \vec{\alpha}, \boldsymbol{\beta})$ as possible.

Variational Inference, Cont.

- Using the fact that

$$\begin{aligned} P(\vec{w}, \vec{z}, \vec{\theta} \mid \vec{\alpha}, \boldsymbol{\beta}) &= P(\vec{w}, \vec{z} \mid \vec{\theta}, \vec{\alpha}, \boldsymbol{\beta}) P(\vec{\theta} \mid \vec{\alpha}, \boldsymbol{\beta}) \\ &= P(\vec{w} \mid \vec{z}, \vec{\theta}, \vec{\alpha}, \boldsymbol{\beta}) P(\vec{z} \mid \vec{\theta}, \vec{\alpha}, \boldsymbol{\beta}) P(\vec{\theta} \mid \vec{\alpha}) \\ &= P(\vec{w} \mid \vec{z}, \boldsymbol{\beta}) P(\vec{z} \mid \vec{\theta}) P(\vec{\theta} \mid \vec{\alpha}). \end{aligned}$$

- we have

$$\begin{aligned} L(\vec{\alpha}, \boldsymbol{\beta}) &= \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln \frac{P(\vec{w} \mid \vec{z}, \beta) P(\vec{z} \mid \vec{\theta}) P(\vec{\theta} \mid \vec{\alpha})}{q(\vec{z}, \vec{\theta})} d\vec{\theta} \\ &= E_q\{\ln P(\vec{w} \mid \vec{z}, \boldsymbol{\beta})\} + E_q\{\ln P(\vec{z} \mid \vec{\theta})\} + E_q\{\ln P(\vec{\theta} \mid \vec{\alpha})\} - E_q\{\ln q(\vec{z}, \vec{\theta})\}. \end{aligned}$$

- Assume that Mean Field approximation and let $q(\vec{z}, \vec{\theta})$ be parameterized:

$$\begin{aligned} q(\vec{z}, \vec{\theta}) &= q\left(\vec{z}, \vec{\theta} \mid \vec{\phi}^{(1)}, \ldots, \vec{\phi}^{(N)}, \vec{\gamma}\right) \\ &= q\left(\vec{z} \mid \vec{\phi}^{(1)}, \ldots, \vec{\phi}^{(N)}\right) q(\vec{\theta} \mid \vec{\gamma}) \\ &= q\left(z_1, \ldots, z_N \mid \vec{\phi}^{(1)}, \ldots, \vec{\phi}^{(N)}\right) q(\vec{\theta} \mid \vec{\gamma}) \\ &= q(\vec{\theta} \mid \vec{\gamma}) \prod_{n=1}^{N} q\left(z_n \mid \vec{\phi}^{(n)}\right), \end{aligned}$$

  where $\vec{\gamma}$ is assumed as Dirichlet parameter and $\vec{\phi}^{(n)}$ is assumed as a multinomial parameter.

The expansion of $E_q\{\ln P(\vec{w} \mid \vec{z}, \boldsymbol{\beta})\})$ in ELBO

The expansion of $E_q\{\ln P(\vec{w} \mid \vec{z}, \boldsymbol{\beta})\})$ in ELBO is

- $$
\begin{aligned}
E_q\{\ln P(\vec{w} \mid \vec{z}, \boldsymbol{\beta})\}) &= \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln P(\vec{w} \mid \vec{z}, \boldsymbol{\beta}) d\vec{\theta} \\
&= \sum_{\vec{z}} \left\{ \prod_{n=1}^{N} q\left(z_n \mid \vec{\phi}^{(n)}\right) \right\} \ln P(\vec{w} \mid \vec{z}, \boldsymbol{\beta}) \int q(\vec{\theta} \mid \vec{\gamma}) d\vec{\theta} \\
&= \sum_{\vec{z}} \left\{ \prod_{n=1}^{N} q\left(z_n \mid \vec{\phi}^{(n)}\right) \right\} \left(\sum_{n=1}^{N} \ln P\left(w_n \mid z_n, \boldsymbol{\beta}\right) \right) \\
&= \sum_{\vec{z}} \left\{ \prod_{n=1}^{N} q\left(z_n \mid \vec{\phi}^{(n)}\right) \right\} \left(\sum_{n=1}^{N} \ln \beta_{z_n, w_n} \right) \\
&= \sum_{\vec{z}} \left\{ \prod_{n=1}^{N} q\left(z_n \mid \vec{\phi}^{(n)}\right) \right\} \left(\ln \beta_{z_1, w_1} + \cdots + \ln \beta_{z_N, w_N} \right) \\
&= \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_i^{(n)} w_n^j \ln \beta_{i,j},
\end{aligned}
$$

where w_n^j is 1 if w_n is the j word in the dictionary $\{1, \ldots, V\}$; otherwise $w_n^j = 0$.

The expansion of $E_q\{\ln P(\vec{z} \mid \vec{\theta})\}$ in ELBO

The expansion of $E_q\{\ln P(\vec{z} \mid \vec{\theta})\}$ in ELBO is

- $$
\begin{aligned}
E_q\{\ln P(\vec{z} \mid \vec{\theta})\} &= \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln P(\vec{z} \mid \vec{\theta}) d\vec{\theta} \\
&= \int q(\vec{\theta} \mid \vec{\gamma}) \sum_{\vec{z}} \left\{ \prod_{n=1}^{N} q\left(z_n \mid \vec{\phi}^{(n)}\right) \right\} \left\{ \sum_{n=1}^{N} \ln P\left(z_n \mid \vec{\theta}\right) \right\} d\vec{\theta} \\
&= \int q(\vec{\theta} \mid \vec{\gamma}) \sum_{n=1}^{N} \sum_{z_n} (\ln \theta_{z_n}) \, \phi_{z_n}^{(n)} d\vec{\theta} \\
&= \int q(\vec{\theta} \mid \vec{\gamma}) \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \ln \theta_i d\vec{\theta} \\
&= \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \int q(\vec{\theta} \mid \vec{\gamma}) \ln \theta_i d\vec{\theta} \\
&= \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \left\{ \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j \right) \right\},
\end{aligned}
$$

where $\Psi(\cdot)$ is the digamma function, the first derivative of the log Gamma function.

The expansion of $E_q\{\ln P(\vec{\theta} \mid \vec{\alpha})\}$ in ELBO

The expansion of $E_q\{\ln P(\vec{\theta} \mid \vec{\alpha})\}$ in ELBO is

- $$
\begin{aligned}
E_q\{\ln P(\vec{\theta} \mid \vec{\alpha})\} &= \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln P(\vec{\theta} \mid \vec{\alpha}) d\vec{\theta} \\
&= \int q(\vec{\theta} \mid \vec{\gamma}) \ln P(\vec{\theta} \mid \vec{\alpha}) d\vec{\theta} \\
&= \int q(\vec{\theta} \mid \vec{\gamma}) \left(\ln \Gamma \left(\sum_{i=1}^{K} \alpha_i \right) - \sum_{i=1}^{K} \ln \Gamma (\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \ln \theta_i \right) d\vec{\theta} \\
&= \ln \Gamma \left(\sum_{i=1}^{K} \alpha_i \right) - \sum_{i=1}^{K} \ln \Gamma (\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \int q(\vec{\theta} \mid \vec{\gamma}) \ln \theta_i d\vec{\theta} \\
&= \ln \Gamma \left(\sum_{i=1}^{K} \alpha_i \right) - \sum_{i=1}^{K} \ln \Gamma (\alpha_i) \\
&+ \sum_{i=1}^{K} (\alpha_i - 1) \left\{ \Psi (\gamma_i) - \Psi \left(\sum_{j=1}^{K} \gamma_j \right) \right\}.
\end{aligned}
$$

The expansion of $E_q\{\ln q(\vec{z}, \vec{\theta})\}$ in ELBO

The expansion of $E_q\{\ln q(\vec{z}, \vec{\theta})\}$ in ELBO is

- $$\begin{aligned} E_q\{\ln q(\vec{z}, \vec{\theta})\} &= E_q\left\{\ln\left(q(\vec{\theta} \mid \vec{\gamma}) \prod_{n=1}^{N} q\left(z_n \mid \vec{\phi}^{(n)}\right)\right)\right\} \\ &= E_q\{\ln q(\vec{\theta} \mid \vec{\gamma})\} + \sum_{n=1}^{N} E_q\left\{\ln q\left(z_n \mid \vec{\phi}^{(n)}\right)\right\}. \end{aligned}$$

- $$\begin{aligned} E_q\{\ln q(\vec{\theta} \mid \vec{\gamma})\} &= \ln\Gamma\left(\sum_{i=1}^{K} \gamma_i\right) - \sum_{i=1}^{K} \ln\Gamma(\gamma_i) \\ &+ \sum_{i=1}^{K} (\gamma_i - 1)\left\{\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right\}. \end{aligned}$$

The expansion of $E_q\{\ln q(\vec{z}, \vec{\theta})\}$ in ELBO, Cont.

- $$\begin{aligned}
\sum_{n=1}^{N} E_q\left\{\ln q\left(z_n \mid \vec{\phi}^{(n)}\right)\right\} &= \sum_{n=1}^{N} \int \sum_{\vec{z}} q(\vec{z}, \vec{\theta}) \ln q\left(z_n \mid \vec{\phi}^{(n)}\right) d\vec{\theta} \\
&= \sum_{n=1}^{N} \sum_{\vec{z}} \left\{\prod_{t=1}^{N} q\left(z_t \mid \vec{\phi}^{(t)}\right)\right\} \ln q\left(z_n \mid \vec{\phi}^{(n)}\right) \int q(\vec{\theta} \mid \vec{\gamma}) d\vec{\theta} \\
&= \sum_{n=1}^{N} \sum_{z} \left\{\prod_{t=1}^{N} q\left(z_t | \vec{\phi}^{(t)}\right)\right\} \ln q\left(z_n | \vec{\phi}^{(n)}\right) \\
&= \sum_{n=1}^{N} \sum_{z_n} q\left(z_n \mid \vec{\phi}^{(n)}\right) \ln q\left(z_n \mid \vec{\phi}^{(n)}\right) \\
&= \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \ln \phi_i^{(n)}.
\end{aligned}$$

The expansion of ELBO

The expansion of ELBO is

- $$
\begin{aligned}
L(\vec{\alpha}, \boldsymbol{\beta}) &= L\left(\vec{w}; \vec{\phi}^{(1)}, \ldots, \vec{\phi}^{(N)}, \vec{\gamma}; \vec{\alpha}, \boldsymbol{\beta}\right) \\
&= \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_i^{(n)} w_n^j \ln \beta_{i,j} \\
&+ \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \left\{ \Psi\left(\gamma_i\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right) \right\} \\
&+ \ln \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \ln \Gamma\left(\alpha_i\right) + \sum_{i=1}^{K} \left(\alpha_i - 1\right) \left\{ \Psi\left(\gamma_i\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right) \right\} \\
&- \ln \Gamma\left(\sum_{i=1}^{K} \gamma_i\right) + \sum_{i=1}^{K} \ln \Gamma\left(\gamma_i\right) - \sum_{i=1}^{K} \left(\gamma_i - 1\right) \left\{ \Psi\left(\gamma_i\right) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right) \right\} \\
&- \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_i^{(n)} \ln \phi_i^{(n)}.
\end{aligned}
$$

Algorithm: Maximizing ELBO

The purpose of algorithm is to maximize its evidence lower bound, $\mathbf{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta})$.

- The idea of the algorithm is to start from an initial $(\vec{\alpha}, \boldsymbol{\beta})$, and iteratively improve the estimate via the following alternating E-step and M-step:
- E-step: Given $(\vec{\alpha}, \boldsymbol{\beta})$, find $(\boldsymbol{\phi}, \boldsymbol{\gamma})$ to maximize $\mathbf{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta})$,
- M-step: Given $(\boldsymbol{\phi}, \boldsymbol{\gamma})$ found from the E-step, find $(\vec{\alpha}, \boldsymbol{\beta})$ to maximize $\mathbf{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta})$.
- The two steps are repeated until the value of $\mathbf{L}(\boldsymbol{\phi}, \boldsymbol{\gamma}; \vec{\alpha}, \boldsymbol{\beta})$ converges.
- In E-step, the update equations for $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ are

$$\begin{aligned} \phi_i^{(n,d)} &\propto \beta_{i,w_n^{(d)}} \exp\left\{ \Psi(\gamma_i^{(d)}) - \Psi\left(\sum_{j=1}^{K} \gamma_j^{(d)} \right) \right\}, \\ \gamma_i^{(d)} &= \alpha_i + \sum_{n=1}^{N_d} \phi_i^{(n,d)}. \end{aligned}$$

- Once $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ are obtained in E-step, the update equation for $\boldsymbol{\beta}$ can be determined by using Lagrange multipliers, it's

$$\beta_{i,j} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_i^{(n,d)} I(w_n^{(d)} = j),$$

where I is indicator function.

Algorithm: Maximizing ELBO, Cont.

Note that there is no analytical form of the update equation for $\vec{\alpha}$ in the M-step, the Newton's method can be employed for obtaining updated $\vec{\alpha}$ by maximizing

$$\mathbf{L}_{\vec{\alpha}} = \sum_{d=1}^{D} \left[\ln \Gamma \left(\sum_{i=1}^{K} \alpha_i \right) - \sum_{i=1}^{K} \ln \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \left\{ \Psi(\gamma_i^{(d)}) - \Psi \left(\sum_{j=1}^{K} \gamma_j^{(d)} \right) \right\} \right].$$

Collapsed Gibbs sampling

Let $k = z_n^{(d)}$ and $v = w_n^{(d)}$, the full conditional is reduced to

$$
\begin{aligned}
P(z_n^{(d)} &= k \mid \mathbf{Z} \setminus z_n^{(d)}, w_n^{(d)} = v, \mathbf{D} \setminus w_n^{(d)}, \vec{\alpha}, \boldsymbol{\eta}) \\
&\propto \frac{B(\vec{\eta}^{(k)} + \vec{c}_k)}{B(\vec{\eta}^{(k)} + \vec{c}'_k)} \cdot \frac{B(\vec{\alpha} + \vec{c}^{(d)})}{B(\vec{\alpha} + \vec{c}^{(d)\prime})} \\
&= \frac{\eta_v^{(k)} + c_{kv} - 1}{\sum_{j=1}^{V} \eta_j^{(k)} + \sum_{j=1}^{V} c_{kj} - 1} \cdot \frac{\alpha_k + c_k^{(d)} - 1}{\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{K} c_i^{(d)} - 1} \\
&\propto \frac{(\eta_v^{(k)} + c_{kv} - 1) \cdot (\alpha_k + c_k^{(d)} - 1)}{\sum_{j=1}^{V} \eta_j^{(k)} + \sum_{j=1}^{V} c_{kj} - 1},
\end{aligned}
$$

where $c_i^{(d)}$ is count of topic i in document d, and c_{ij} is count of topic i to word j in all documents.

Collapsed Gibbs sampling, Cont.

Given this full conditional posterior, the Gibbs sampling procedure is as straightforward as follows:

- Step 0: Initialize $\vec{\alpha}, \boldsymbol{\eta}, \mathbf{Z}$.
- Step 1: Iteratively update $\mathbf{Z}$ by drawing $z_n^{(d)}$.
- Step 2: Update $\vec{\alpha}, \boldsymbol{\eta}$ by maximizing the joint likelihood function $P(\mathbf{Z}, \mathbf{D} \mid \vec{\alpha}, \boldsymbol{\eta})$. Go to step 1.

The step 1 usually requires huge number of iterations. The objective function in step 2 is tractable and there is no need for a evidence lower bound.

- $\widehat{E}(\vec{\beta}^{(i)}) = (\vec{\eta}^{(i)} + \vec{c}_i)/(\sum_{j=1}^{V} \eta_j^{(i)} + \sum_{j=1}^{V} c_{ij})$, which represents the word distribution over topic i and
- $\widehat{E}(\vec{\theta}^{(d)}) = (\vec{\alpha} + \vec{c}^{(d)})/(\sum_{i=1}^{K} \alpha_i + \sum_{i=1}^{K} + c_i^{(d)})$, which represents the topic distribution in document d.