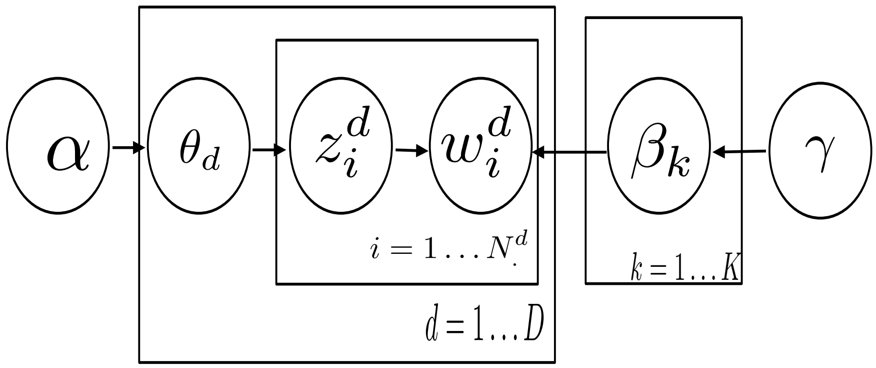

FIGURE 1. Graphical model for LDA model

Lecture LDA

LDA is a hierarchical model used to model text documents. Each document is modeled as a mixture of topics. Each topic is defined as a distribution over the words in the vocabulary. Here, we will denote by K the number of topics in the model. We use D to indicate the number of documents, M to denote the number of words in the vocabulary, and $N^d_\cdot$ to denote the number of words in document d. We will assume that the words have been translated to the set of integers $\{1, \ldots, M\}$ through the use of a static dictionary. This is for convenience only and the integer mapping will contain no semantic information. The generative model for the D documents can be thought of as sequentially drawing a topic mixture θ_d for each document independently from a $\mathrm{Dir}_K(\alpha\vec{1})$ distribution, where $\mathrm{Dir}_K(\vec{\phi})$ is a Dirichlet distribution over the K-dimensional simplex with parameters $[\phi_1, \phi_2, \ldots, \phi_K]$. Each of K topics $\{\beta_k\}_{k=1}^K$ are drawn independently from $\mathrm{Dir}_M(\gamma\vec{1})$. Then, for each of the $i = 1 \ldots N^d$. words in document d, an assignment variable z_i^d is drawn from $\mathrm{Mult}(\theta^d)$. Conditional on the assignment variable z_i^d, word i in document d, denoted as w_i^d, is drawn independently from $\mathrm{Mult}(\beta_{z_i^d})$. The graphical model for the process can be seen in Figure 1.

The model is parameterized by the vector valued parameters $\{\theta_d\}_{d=1}^D$, and $\{\beta_k\}_{k=1}^K$, the parameters $\{Z_i^d\}_{d=1,\ldots,D, i=1,\ldots,N^d_\cdot}$, and the scalar positive parameters α and γ. The model is formally written as:

$$
\begin{aligned}
\theta_d &\sim \mathrm{Dir}_K(\alpha\vec{1}) \\
\beta_k &\sim \mathrm{Dir}_M(\gamma\vec{1}) \\
z_i^d &\sim \mathrm{Mult}(\theta_d) \\
w_i^d &\sim \mathrm{Mult}(\beta_{z_i^d})
\end{aligned}
$$

Again, each of the K β_k parameters represents a unique "topic". Here, the mathematical realization of a topic is a multinomial distribution over the words in the vocabulary. You can imagine that topics having to do with football will have high probability on words like "throw", "ball", "running", and "concussion", while a topic like machine learning will have high probability on words like "algorithm", "learning", and "empirical". For each document there is a mixture of topics which represent it. This allows documents to be about multiple "topics". Since we are sharing topics accross documents we have a chance at learning which documents are like each other based on the topics they are about.

We will now take a quick detour to discuss the Dirichlet distribution and the gamma function. Recall that if $\theta \sim \mathrm{Dir}_K(\alpha)$ then

$$P(\theta) = \frac{\Gamma(\sum_i \alpha_i)}{\Pi_i \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_K^{\alpha_K - 1}.$$

Note that the pdf for the Dirichlet distribution makes use of the gamma function Γ. Now reecall that

$$\begin{aligned}
\Gamma(\eta + 1) &= \int_0^\infty e^{-t} t^\eta dt \\
&= -t^\eta e^{-t}|_0^\infty + \eta \int_0^\infty e^{-t} t^{\eta - 1} dt \\
&= \eta \Gamma(\eta)
\end{aligned}$$

The recursive relationship is derived here using integration by parts.

We can now return to the LDA model and consider the joint likelihood of the model. We can write the joint likelihood $L(\{Z_i^d\}, \{\theta_d\}, \{\beta_k\})$ as :

$$\left[\Pi_{d=1}^D \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \theta_{d,1}^{\alpha-1} \dots \theta_{d,K}^{\alpha-1}\right] \left[\Pi_{d=1}^D \Pi_{i=1}^{N_\cdot^d} \theta_{d,z_i^d}\right] \left[\Pi_{k=1}^K \frac{\Gamma(M\gamma)}{\Gamma^M(\gamma)} \beta_{k,1}^{\gamma-1} \dots \beta_{k,M}^{\gamma-1}\right] \left[\Pi_{d=1}^D \Pi_{i=1}^{N_\cdot^d} \beta_{z_i^d, w_i^d}\right]$$

If we define $N_k^d = \sum_{i=1}^{N_\cdot^d} I(z_{d,i} == k)$ and $W_m^k = \sum_{d=1}^D \sum_{i=1}^{N_\cdot^d} I(w_i^d == m \ \&\& \ z_i^d == k)$ we will be able to write the joint likelihood in a more compressed form. Note that N_k^d is the number of words in document d assigned to topic k and W_m^k is the number of words of type m assigned to topic k. Furthermore, we will use a . in the subscript or superscript to indicate marginal counts. Thus, $N_\cdot^d$ is the number of words in document d and $W_\cdot^k$ is the total number of words assigned to topic k.

Now, re-writing the joint likelihood we find it is

$$\left[\Pi_{d=1}^D \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \theta_{d,1}^{\alpha+N_1^d-1} \dots \theta_{d,K}^{\alpha+N_K^d-1}\right] \left[\Pi_{k=1}^K \frac{\Gamma(M\gamma)}{\Gamma^M(\gamma)} \beta_{k,1}^{\gamma+W_1^k-1} \dots \beta_{k,M}^{\gamma+W_M^k-1}\right]$$

From this we can see that conditioned on $\{Z_i^d\}$

$$\theta^d \sim \mathrm{Dir}_K([\alpha + N_1^d, \alpha + N_2^d, \dots, \alpha + N_K^d])$$

and

$$\beta^k \sim \mathrm{Dir}_M([\gamma + W_1^k, \gamma + W_2^k, \ldots, \gamma + W_M^k])$$

If we wish to write a Gibbs sampler in this model representation the only other conditional distribution we need to consider is that of the Z_i^d. If we consider the original form of the joint likelihood we see that Z_i^d for a fixed i and d shows up only twice. Once, it shows up in a θ_{d,z_i^d} term and once in a $\beta_{z_i^d,w_i^d}$ term. Thus, conditioned on $\{\beta_k\}$ and $\{\theta_d\}$, we see that $P(Z_i^d = \tilde{k}) \propto \theta_{d,\tilde{k}}\beta_{\tilde{k},w_i^d}$. We could now write a Gibbs sampler by sampling the θ and β parameters conditioned on the Z parameters and then sampling the Z parameters conditioned on the θ and $beta$ parameters.

While the above Gibbs sampler will work it actually mixes quite slowly. One thing we can sometimes do in hierarchical models of this type is reduce the parameter space by analytically integerating some of the latent parameters out. Here, we will integerate out both the θ and β parameters and consider only the latent parameters z.

$$\begin{aligned}
L(\{Z_i^d\}) &= \int\int L(\{\theta_k\}, \{\beta_k\}, \{Z_i^d\}) d\{\theta_d\} d\{\beta_k\} \\
&= \left[\Pi_{d=1}^D \left(\frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)}\right)\left(\frac{\Pi_{k=1}^K \Gamma(\alpha + N_k^d)}{\Gamma(K\alpha + N_\cdot^d)}\right)\right]\left[\Pi_{k=1}^K \left(\frac{\Gamma(M\gamma)}{\Gamma^M(\gamma)}\right)\left(\frac{\Pi_{m=1}^M \Gamma(\gamma + W_m^k)}{\Gamma(M\gamma + W_\cdot^k)}\right)\right] \\
&\propto \left[\Pi_{d=1}^D \Pi_{k=1}^K \Gamma(\alpha + N_k^d)\right]\left[\Pi_{k=1}^K \frac{\Pi_{m=1}^M \Gamma(\gamma + W_m^k)}{\Gamma(M\gamma + W_\cdot^k)}\right]
\end{aligned}$$

This integral is not hard to do since we already know the normalizing constant of the Dirichlet distribution. When we integrate the unormalized Dirichlet pdf we must get the inverse of the normalizing constant in the pdf.

Now, to create a Gibbs sampler we need only consider the conditional distribution of z_i^d for each d and i. For this part of the derivation we will consider a fixed z_i^d. We will definte $\tilde{N}_k^d$ and $\tilde{W}_m^k$ the same as before except without the contribution of z_i^d. Therefore, N_k^d is the number of words in document d, other than the i'th word, assigned to topic k. Using these new count variables we can derive the conditional distribution of z_i^d up to a constant of proportionality.

$$\begin{aligned}
P(z_i^d = \tilde{k}) &\propto \Pi_{k=1}^K \left[\Gamma(\alpha + \tilde{N}_k^d + I(k = \tilde{k}))\frac{\Pi_{m=1}^M \Gamma(\gamma + \tilde{W}_m^k + I(w_i^d = m, k = \tilde{k}))}{\Gamma(M\gamma + \tilde{W}_\cdot^k + I(k = \tilde{k}))}\right] \\
&\propto \left(\Pi_{k=1}^K \left[\Gamma(\alpha + \tilde{N}_k^d)\frac{\Pi_{m=1}^M \Gamma(\gamma + \tilde{W}_m^k)}{\Gamma(M\gamma + \tilde{W}_\cdot^k)}\right]\right)\left(\frac{(\alpha + \tilde{N}_{\tilde{k}}^d)(\gamma + \tilde{W}_{w_i^d}^{\tilde{k}})}{M\gamma + \tilde{W}_\cdot^{\tilde{k}}}\right) \\
&\propto \frac{(\alpha + \tilde{N}_{\tilde{k}}^d)(\gamma + \tilde{W}_{w_i^d}^{\tilde{k}})}{M\gamma + \tilde{W}_\cdot^{\tilde{k}}}
\end{aligned}$$

Because Z_i^d can only take values $1, \ldots, K$ it is not hard to normalize this distribution to find the conditional distribution we need for Gibbs sampling. To create a Gibbs sampler in this representation we need only sample each Z_i^d in succession. We can always recover the β and θ parameters if we need them based on the conditional distributions of the β and θ variables conditioned on the Z variables.