Collapsed Gibbs sampling for latent Dirichlet allocation

正田 備也

masada@rikkyo.ac.jp

Contents

Joint distribution

$$
\begin{aligned}
& p(\boldsymbol{X}, \boldsymbol{Z}, \Theta, \Phi; \alpha, \beta) \\
& = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d; \alpha) \times \prod_{k=1}^{K} p(\boldsymbol{\phi}_k; \beta) \times \prod_{d=1}^{D} \prod_{i=1}^{n_d} p(z_{d,i} | \boldsymbol{\theta}_d) p(x_{d,i} | \boldsymbol{\phi}_{z_{d,i}})
\end{aligned} \quad (1)
$$

Marginalization

$$
\begin{aligned}
p(\boldsymbol{X}, \boldsymbol{Z}; \alpha, \beta) &= \int p(\boldsymbol{X}, \boldsymbol{Z}, \Theta, \Phi; \alpha, \beta) d\Theta d\Phi \\
&= \int \prod_{d=1}^{D} \left[p(\boldsymbol{\theta}_d; \alpha) \prod_{i=1}^{n_d} p(z_{d,i} | \boldsymbol{\theta}_d) \right] d\Theta \\
&\quad \times \int \left[\prod_{k=1}^{K} p(\boldsymbol{\phi}_k; \beta) \prod_{d=1}^{D} \prod_{i=1}^{n_d} p(x_{d,i} | \boldsymbol{\phi}_{z_{d,i}}) \right] d\Phi \\
&= \prod_{d} \int p(\boldsymbol{z}_d, \boldsymbol{\theta}_d; \alpha) d\boldsymbol{\theta}_d \times \int p(\boldsymbol{X}, \Phi | \boldsymbol{Z}; \alpha, \beta) d\Phi \\
&= \prod_{d} p(\boldsymbol{z}_d; \alpha) \times p(\boldsymbol{X} | \boldsymbol{Z}; \beta) \qquad (2)
\end{aligned}
$$

Topic assignments probability

$$p(\boldsymbol{z}_d, \boldsymbol{\theta}_d; \alpha) = p(\boldsymbol{\theta}_d; \alpha) \prod_{i=1}^{n_d} p(z_{d,i}|\boldsymbol{\theta}_d) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_{d,k}^{\alpha_k - 1} \times \frac{(\sum_k n_{d,k})!}{\prod_k n_{d,k}!} \prod_k \theta_{d,k}^{n_{d,k}}$$
$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{n_d!}{\prod_k n_{d,k}!} \prod_k \theta_{d,k}^{n_{d,k}+\alpha_k-1} \tag{3}$$

where $n_{d,k} \equiv \sum_{i=1}^{n_d} \delta(z_{d,i} = k)$.

$$\int \left[p(\boldsymbol{\theta}_d; \alpha) \prod_{i=1}^{n_d} p(z_{d,i}|\boldsymbol{\theta}_d) \right] d\boldsymbol{\theta}_d = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{n_d!}{\prod_k n_{d,k}!} \int \prod_k \theta_{d,k}^{n_{d,k}+\alpha_k-1} \boldsymbol{\theta}_d$$
$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{n_d!}{\prod_k n_{d,k}!} \frac{\prod_k \Gamma(n_{d,k} + \alpha_k)}{\Gamma(n_d + \sum_k \alpha_k)} \tag{4}$$

$$\therefore p(\boldsymbol{z}_d; \alpha) = \int p(\boldsymbol{z}_d, \boldsymbol{\theta}_d; \alpha) d\boldsymbol{\theta}_d = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{n_d!}{\prod_k n_{d,k}!} \frac{\prod_k \Gamma(n_{d,k} + \alpha_k)}{\Gamma(n_d + \sum_k \alpha_k)} \tag{5}$$

Topic posterior

$$
\begin{aligned}
p(\boldsymbol{\theta}_d | \boldsymbol{z}_d; \alpha) &= \frac{p(\boldsymbol{z}_d, \boldsymbol{\theta}_d; \alpha)}{p(\boldsymbol{z}_d; \alpha)} \\
&= \frac{\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{n_d!}{\prod_k n_{d,k}!} \prod_k \theta_{d,k}^{n_{d,k}+\alpha_k-1}}{\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{n_d!}{\prod_k n_{d,k}!} \frac{\prod_k \Gamma(n_{d,k}+\alpha_k)}{\Gamma(n_d+\sum_k \alpha_k)}} \\
&= \frac{\Gamma(n_d + \sum_k \alpha_k)}{\prod_k \Gamma(n_{d,k} + \alpha_k)} \prod_k \theta_{d,k}^{n_{d,k}+\alpha_k-1}
\end{aligned} \tag{6}
$$

Word tokens probability

$$
\begin{aligned}
p(\boldsymbol{X}, \Phi | \boldsymbol{Z}; \beta) &= \prod_{k=1}^{K} p(\boldsymbol{\phi}_k; \beta) \prod_{d=1}^{D} \prod_{i=1}^{n_d} p(x_{d,i} | \boldsymbol{\phi}_{z_{d,i}}) = \prod_{k=1}^{K} \left[p(\boldsymbol{\phi}_k; \beta) \prod_{d=1}^{D} \prod_{i=1}^{n_d} p(x_{d,i} | \boldsymbol{\phi}_k)^{\delta(z_{d,i}=k)} \right] \\
&= \prod_{k=1}^{K} \left[\frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \prod_{w=1}^{W} \phi_{k,w}^{\beta_w - 1} \times \frac{n_k!}{\prod_w n_{k,w}!} \prod_w \phi_{k,w}^{n_{k,w}} \right] \\
&= \prod_{k=1}^{K} \left[\frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{n_k!}{\prod_w n_{k,w}!} \prod_w \phi_{k,w}^{n_{k,w} + \beta_w - 1} \right]
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
\int p(\boldsymbol{X}, \Phi | \boldsymbol{Z}; \beta) d\Phi &= \prod_{k=1}^{K} \left[\frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{n_k!}{\prod_w n_{k,w}!} \int \prod_w \phi_{k,w}^{n_{k,w} + \beta_w - 1} d\boldsymbol{\theta}_k \right] \\
&= \prod_{k=1}^{K} \left[\frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{n_k!}{\prod_w n_{k,w}!} \frac{\prod_w \Gamma(n_{k,w} + \beta_w)}{\Gamma(n_k + \sum_w \beta_w)} \right]
\end{aligned} \tag{8}
$$

Word posterior

$$
\begin{aligned}
p(\Phi|\boldsymbol{X}, \boldsymbol{Z}, \beta) &= \frac{p(\boldsymbol{X}, \Phi|\boldsymbol{Z}; \beta)}{p(\boldsymbol{X}|\boldsymbol{Z}; \beta)} = \frac{p(\boldsymbol{X}, \Phi|\boldsymbol{Z}; \beta)}{\int p(\boldsymbol{X}, \Phi|\boldsymbol{Z}; \beta) d\Phi} \\
&= \frac{\prod_{k=1}^{K} \left[\frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{n_k!}{\prod_w n_{k,w}!} \prod_w \phi_{k,w}^{n_{k,w}+\beta_w-1} \right]}{\prod_{k=1}^{K} \left[\frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{n_k!}{\prod_w n_{k,w}!} \frac{\prod_w \Gamma(n_{k,w}+\beta_w)}{\Gamma(n_k+\sum_w \beta_w)} \right]} \\
&= \prod_{k=1}^{K} \left[\frac{\Gamma(n_k + \sum_w \beta_w)}{\prod_w \Gamma(n_{k,w} + \beta_w)} \prod_w \phi_{k,w}^{n_{k,w}+\beta_w-1} \right]
\end{aligned} \tag{9}
$$

Per-token topic assignment posterior (1/3)

$$
\begin{aligned}
&p(z_{d,i} = k | x_{d,i} = v, \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}, \alpha, \beta) = \frac{p(z_{d,i} = k, x_{d,i} = v, \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta)}{p(x_{d,i} = v, \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta)} \\
&= \frac{p(z_{d,i} = k, x_{d,i} = v, \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta)}{\sum_{k'=1}^{K} p(z_{d,i} = k', x_{d,i} = v, \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta)} \\
&\propto p(z_{d,i} = k, x_{d,i} = v, \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) \\
&= p(x_{d,i} = v, z_{d,i} = k | \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) p(\boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) \\
&= p(x_{d,i} = v | z_{d,i} = k, \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) p(z_{d,i} = k | \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) p(\boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) \\
&\propto p(x_{d,i} = v | z_{d,i} = k, \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) p(z_{d,i} = k | \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) \qquad (10)
\end{aligned}
$$

Per-token topic assignment posterior (2/3)

$$p(z_{d,i} = k | \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) = \int p(z_{d,i} = k | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i | \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) d\boldsymbol{\theta}_i \tag{11}$$

$$\begin{aligned} p(\boldsymbol{\theta}_i | \boldsymbol{X}^{\setminus d,i}, \boldsymbol{Z}^{\setminus d,i}; \alpha, \beta) &= p(\boldsymbol{\theta}_d | \boldsymbol{z}_d^{\setminus d,i}; \alpha) \\ &= \frac{\Gamma(n_d^{\setminus d,i} + \sum_k \alpha_k)}{\prod_k \Gamma(n_{d,k}^{\setminus d,i} + \alpha_k)} \prod_k \theta_{d,k}^{n_{d,k}^{\setminus d,i} + \alpha_k - 1} \end{aligned} \tag{12}$$