A Collapsed Gibbs Sampler for the Relational Topic Model

Arthur Asuncion
July 2009

Abstract

Here is a collapsed Gibbs sampler for the Relational Topic Model of Chang/Blei.

1 Relational Topic Model

Here is RTM's generative process. I use the exponential link probability function for simplicity. I also put a prior on the topics β_k instead of treating them as parameters. I use Chang/Blei's choice of symbols, except I denote the observed words as x instead of w:

$$\theta_d \sim \mathcal{D}[\alpha] \qquad \beta_k \sim \mathcal{D}[\gamma] \qquad z_{d,n} \sim \theta_d \qquad x_{d,n} \sim \beta_{z_{d,n}} \qquad y_{d,d'} \sim \psi_e(y|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu) \tag{1}$$

Here is the full joint distribution of the RTM:

$$\begin{aligned}
&p(x, z, y, \theta, \beta|\alpha, \gamma, \eta, \nu) \\
&= \prod_{nd} p(x_{nd}|\beta_{z_{nd}}) \prod_{nd} p(z_{nd}|\theta_d) \prod_{d,d'} \psi_e(y_{d,d'}|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu) \prod_k p(\beta_k|\gamma) \prod_d p(\theta_d|\alpha) \\
&\propto \prod_{nd} \beta_{1|z_{nd}}^{N_{nd1}} ... \beta_{W|z_{nd}}^{N_{ndW}} \prod_{nd} \theta_{1|d}^{N_{nd1}} ... \theta_{K|d}^{N_{ndK}} \prod_{d,d'} \psi_e(y_{d,d'}|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu) \\
&\quad \prod_k \frac{\Gamma(W\gamma)}{\Gamma(\gamma)^W} \prod_k \beta_{1|k}^{\gamma-1} ... \beta_{W|k}^{\gamma-1} \prod_d \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \theta_{1|d}^{\alpha-1} ... \theta_{D|d}^{\alpha-1} \\
&= \prod_k \beta_{1|k}^{N_{k1}+\gamma-1} ... \beta_{W|k}^{N_{kW}+\gamma-1} \prod_d \theta_{1|d}^{N_{d1}+\alpha-1} ... \theta_{K|d}^{n_{dK}+\alpha-1} \\
&\quad \prod_k \frac{\Gamma(W\gamma)}{\Gamma(\gamma)^W} \prod_d \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{d,d'} \psi_e(y_{d,d'}|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu)
\end{aligned} \tag{2}$$

In the derivation above (2), note that $N_{ndz_{nd}1}$ is 1 if token n in document d is assigned to topic z_{nd} and word-type 1 (otherwise, $N_{ndz_{nd}1} = 0$). Likewise, n_{k1} is the number of tokens that are assigned to topic k and word 1.

We can derive the collapsed joint distribution by integrating out $\{\theta_d\}$ and $\{\beta_k\}$. Due to the conjugacy between the Dirichlet and the discrete distribution, the resulting collapsed distribution can be analytically written down and is simply a product of gamma functions which are the normalizing constants of the prior and posterior Dirichlets:

$$\begin{aligned}
p(w, z, y|\alpha, \gamma, \eta, \nu) = &\prod_d \left[\frac{\Gamma(K\alpha)}{\Gamma(N_d + K\alpha)} \prod_k \frac{\Gamma(N_{dk} + \alpha)}{\Gamma(\alpha)} \right] \prod_k \left[\frac{\Gamma(W\gamma)}{\Gamma(N_k + W\gamma)} \prod_w \frac{\Gamma(N_{kw} + \gamma)}{\Gamma(\gamma)} \right] \\
&\prod_{d,d'} \psi_e(y_{d,d'}|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu)
\end{aligned} \tag{3}$$

In order to obtain the conditional distribution for a particular z_{nd}, we simply need to view all other variables as constant. We can remove terms from the collapsed joint that do not depend on z_{nd}. Making use of facts like $\Gamma(N_{dk}+\alpha) = (N_{dk}^{\neg nd}+\alpha)\Gamma(N_{dk}^{\neg nd}+\alpha)$ and realizing that $\Gamma(N_{dk}^{\neg nd}+\alpha)$ does not rely on z_{nd} (since $\neg nd$ means we have excluded token nd from the count) allows us to obtain the conditional distribution. The resulting conditional distribution is simply the standard LDA conditional distribution of z_{nd} multiplied by the relevant link probability functions:

$$
\begin{aligned}
p(z_{nd}=k|\boldsymbol{z}^{\neg nd}, x_{nd}=w, x^{\neg nd}, y, \alpha, \gamma, \eta, \nu) &\propto \left[(N_{dk}^{\neg nd}+\alpha)\frac{(N_{kw}^{\neg nd}+\gamma)}{(N_{k}^{\neg nd}+W\gamma)}\right] \\
&\quad \prod_{d'\neq d: y_{d,d'}=1} \psi_e(y_{d,d'}=1|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu) \\
&\quad \prod_{d'\neq d: y_{d,d'}=0} \psi_e(y_{d,d'}=0|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu) \qquad (4) \\
&\propto \left[(N_{dk}^{\neg nd}+\alpha)\frac{(N_{kw}^{\neg nd}+\gamma)}{(N_{k}^{\neg nd}+W\gamma)}\right] \\
&\quad \prod_{d'\neq d: y_{d,d'}=1} \frac{\psi_e(y_{d,d'}|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu)}{\psi_e(y_{d,d'}|\boldsymbol{z}_d^{\neg nd}, \boldsymbol{z}_{d'}, \eta, \nu)} \\
&\quad \prod_{d'\neq d: y_{d,d'}=0} \psi_e(y_{d,d'}=0|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu) \qquad (5)
\end{aligned}
$$

In line 5, we divide the unnormalized conditional distribution $\psi_e(y_{d,d'}|\boldsymbol{z}_d^{\neg nd}, \boldsymbol{z}_{d'}, \eta, \nu)$ which are constants with respect to z_{nd}. This does not change the conditional distribution.

Now let us consider the terms relating to exponential probability link function (where $\bar{\boldsymbol{z}}_d = \frac{1}{N_d}\sum_n z_{d,n}$ and each $z_{d,n}$ is represented in 0/1 vector form):

$$
\begin{aligned}
\psi_e(y_{d,d'}=1|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu) &= \exp(\eta^T(\bar{\boldsymbol{z}}_d \circ \bar{\boldsymbol{z}}_{d'}) + \nu) \\
&= \exp(\frac{1}{N_d N_{d'}}\sum_k(\eta_k N_{dk} N_{d'k}) + \nu) \qquad (6)
\end{aligned}
$$

$$
\prod_{d'\neq d: y_{d,d'}=1} \frac{\psi_e(y_{d,d'}=1|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu)}{\psi_e(y_{d,d'}=1|\boldsymbol{z}_d^{\neg nd}, \boldsymbol{z}_{d'}, \eta, \nu)} = \prod_{d'\neq d: y_{d,d'}=1} \frac{\exp(\frac{1}{N_d N_{d'}}\sum_k(\eta_k(N_{dk}^{\neg nd}+\mathcal{I}(z_{nd}=k))N_{d'k}) + \nu)}{\exp(\frac{1}{N_d N_{d'}}\sum_k(\eta_k N_{dk}^{\neg nd} N_{d'k}) + \nu)} \qquad (7)
$$

$$
\begin{aligned}
&= \prod_{d'\neq d: y_{d,d'}=1} \exp(\frac{1}{N_d N_{d'}}\eta_{z_{nd}} N_{d'z_{nd}}) \qquad (8) \\
&= \exp(\frac{1}{N_d}\eta_{z_{nd}} \sum_{d'\neq d: y_{d,d'}=1} \frac{N_{d'z_{nd}}}{N_{d'}}) \qquad (9)
\end{aligned}
$$

$$
\psi_e(y_{d,d'}=0|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu) = 1 - \exp(\frac{1}{N_d N_{d'}}\sum_k(\eta_k N_{dk} N_{d'k}) + \nu) \qquad (10)
$$

$$
\prod_{d'\neq d: y_{d,d'}=0} \psi_e(y_{d,d'}=0|\boldsymbol{z}_d, \boldsymbol{z}_{d'}, \eta, \nu) = \prod_{d'\neq d: y_{d,d'}=0} [1 - \exp(\frac{1}{N_d N_{d'}}\sum_k(\eta_k N_{dk} N_{d'k}) + \nu)] \qquad (11)
$$

$$
\approx \left[1 - \exp(\frac{1}{N_d N_{d': y_{d,d'}=0}}\sum_k(\eta_k N_{dk}(\sum_{d'\neq d: y_{d,d'}=0} \frac{N_{d'k}}{N_{d'}})) + \nu)\right]^{N_{d': y_{d,d'}=0}} \qquad (12)
$$

In line 12 above, I made the approximation $\prod_i(1-\exp(c_i)) \approx (1-\exp(\bar{c}_i))^N$ for computational efficiency. If one is willing to tolerate more computation, more accurate approximations can be used as well.

Putting all the pieces together, here is the conditional distribution for z_{nd}:

$$p(z_{nd} = k|\boldsymbol{z}^{\neg nd}, -) \approx\propto \left[(N_{dk}^{\neg nd} + \alpha)\frac{(N_{kw}^{\neg nd} + \gamma)}{(N_{k}^{\neg nd} + W\gamma)}\right] \tag{13}$$

$$\exp(\frac{1}{N_d}\eta_k \sum_{d' \neq d : y_{d,d'}=1} \frac{N_{d'k}}{N_{d'}}) \tag{14}$$

$$\left[1 - \exp(\frac{1}{N_d N_{d' : y_{d,d'}=0}} \sum_{k'} (\eta_{k'}(N_{dk'})(\sum_{d' \neq d : y_{d,d'}=0} \frac{N_{d'k'}}{N_{d'}})) + \nu)\right]^{N_{d' : y_{d,d'}=0}} \tag{15}$$

Note that line 13 is the "LDA" part of the conditional distribution, line 14 is the term for observed edges, and line 15 is the term for the observed non-edges. The term in line 14 only needs to be computed once per document per sweep. Note that the only approximation made is the term in line 15. Also note that z_{nd} is included within the count $N_{dk'}$ where $k' = k$. We can introduce an additional approximation by only computing the term in line 15 once per document per sweep and caching it – this amounts to using the N_{dk} count at the beginning of the Gibbs sweep.

After each Gibbs sweep over $\{z_{nd}\}$, the hyperparameters η, ν can be optimized via the techniques in the appendix of Chang/Blei (although I haven't looked at this closely yet). Also, α, γ can be optimized by Minka's fixed point updates.

I chose to focus on the exponential probability link function since it allows for the simplification in line 14. It may also be possible to simplify the other probability link functions in Chang/Blei.

Since we now have a collapsed Gibbs sampler over $\{z_{nd}\}$ that is similar to the standard LDA sampler, one can easily extend the RTM. For instance, one can sample from the *Relational Author Topic* model by simply plugging in the "author" term into the conditional distribution.

Finally, in order to turn this into a CVB0-style algorithm, one can use the same conditional distribution, but instead of sampling each z_{nd}, one should maintain a distribution over each z_{nd} and update the count matrices N_{dk} and N_{wk} with the fractional count corresponding exactly to the conditional distribution.