A Comparison of Posterior Approximation Algorithms for Latent Dirichlet Allocation

Cindy Cook

STAT540: Final Project

9 December 2015

Latent Dirichlet Allocation

Given a corpus C, what we know,

the vocabulary, $V = \{w_{d,n}\}$ and the number of topics K.

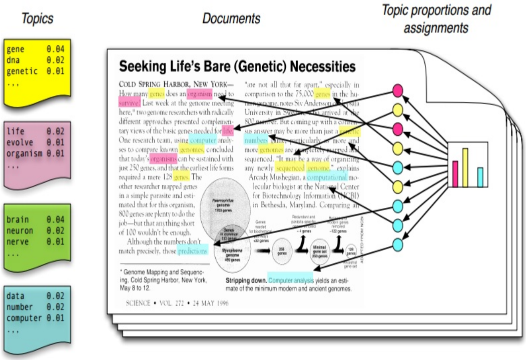


Figure: Article from the *Journal of Science*; Image from David Blei's 2012 topic model video lecture series.

- Each **topic** k is a distribution over the vocabulary.
- Each **document** d is a mixture of corpus wide topics.
- Each **word** $w_{d,n}$ is drawn from one of these topics.

Latent Dirichlet Allocation

Generative Process:

- For each topic k:
  $\beta^k \sim Dir(\eta)$
- For each document d:
  - $\theta^d \sim Dir(\alpha)$
  - For each $w_{d,n}$ in d:
    - Draw $z_{d,n} \sim Multi(\theta^d)$
    - Draw $w_{d,n} \sim Multi(\beta^{z_{d,n}})$

Prior:

$$p(\beta, \theta | \eta, \alpha) = p(\beta|\eta)p(\theta|\alpha)$$

Likelihood:

$$p(w|\beta, \theta) = \sum_z (p(w|z, \beta))p(z|\theta)$$

Posterior:

$$p(\beta, \theta | w, \eta, \alpha) = \frac{p(\beta,\theta|\eta,\alpha)p(w|\beta,\theta)}{\int_\beta \int_\theta \left[(p(\beta|\eta)p(\theta|\alpha) \sum_z (p(w|z,\beta)p(z|\theta)) \right] d\beta d\theta}$$

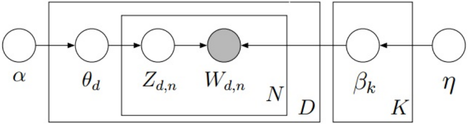


Figure: Directed Graphical Model

Variational EM

In general since the posterior $p(\phi|Y)$ is intractable we use a surrogate $q(\phi)$.

1. From Jensen's inequality $\log p(Y) \geq \int (q(\phi) \log (\frac{p(\phi,Y)}{q(\phi)}) d\phi) = L(q)$.
2. From (Bishop, 2006) $\log p(Y) = L(q) + KL(q||p)$.
3. Finding $q(\phi)$
   1. Exponential family assumptions:
      1. $p(\phi_k|Y, \phi_{-k}) \sim$ exp. family
      2. $q(\phi_k) \sim$ same exp. family
   2. Mean field assumption: $q(\phi) = \prod_{k=1}^{K} q(\phi_k)$.
4. After some calculus, parameters which minimize KL dist can be iteratively found: $q(\phi_k^{new}) = \frac{\exp(E_{j \neq k}[\log p(\phi, Y)])}{\int \exp(E_{j \neq k}[\log p(\phi, Y)]) d\phi_k}$.

$$q(\phi) = \prod_{k=1}^{K} q(\beta|\tilde{\eta}) \prod_{d=1}^{D} q(\theta|\tilde{\alpha}) \prod_{n=1}^{V} q(z_{d,n}|\tilde{\gamma}).$$

Algorithm:

1. Initialize $\tilde{\eta}^{(0)}, \tilde{\alpha}^{(0)}, \tilde{\gamma}^{(0)}$
2. (E-step) Minimize $KL(q||p)$
   Use coordinate ascent algorithm to optimize parameters.
3. (M-step) Maximize $L(q)$ wrt α, η, γ
   Find MLE for β from expected counts (i.e. sufficient stats)
4. If $L(q)$ converged then stop
5. Else return to (E-step)

Collapsed Gibbs Sampling

Simple Gibbs Sampler

- Initialize $\beta^{(0)}, \theta^{(0)}, z_{d,n}^{(0)}$
- Repeat until convergence for $t = 1, 2, \ldots$
  - Draw $\beta^{(t+1)} \sim p(\beta^{(t+1)} | z_{d,n}^{(t)}, w_{d,n})$
  - Draw $\theta^{(t+1)} \sim p(\theta^{(t+1)} | z_{d,n}^{(t)}, w_{d,n})$
  - Draw $z_{d,i}^{(t+1)} \sim p(z_{d,i}^{(t+1)} | \beta^{(t+1)}, \theta_d^{(t+1)}, z_{d,-i}^{(t+1)}, w_{d,n})$

Simple due to the conjugacy of the priors, but VERY slow!!

Integrate out β and θ.

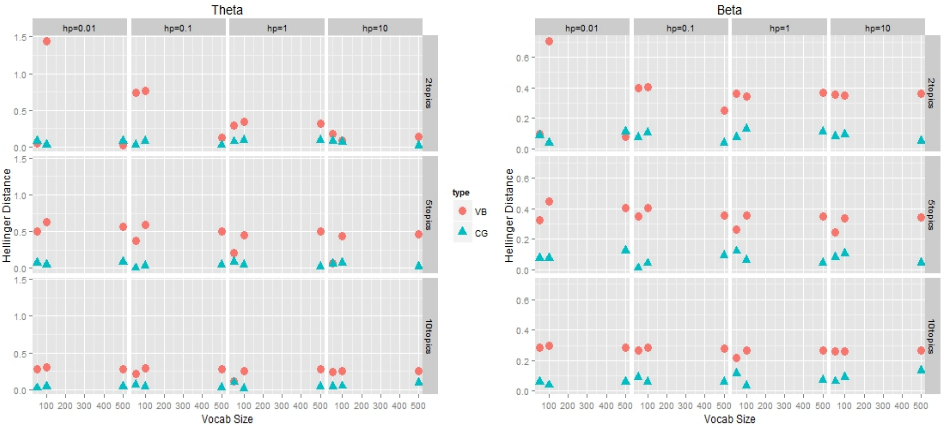
Collapsed Gibbs Sampler

- Initialize $z_{d,n}^{(0)}$
- Repeat for $t = 1, 2, \ldots$
  - Draw $z_{d,i}^{(t+1)} \sim p(z_{d,i}^{(t+1)} | z_{d,-i}^{(t+1)}, w_{d,n})$
- Stop when $H(P, Q) < tol$
- Solve for $\hat{\beta}$
- Solve for $\hat{\theta}$

Faster, but still slow with large vocabularies.
Asymptomatically accurate.

Simulation Study

- Simulated 36 C's according to the LDA generative process with:
- $D = 10$; $\alpha = \eta = \{0.01, 0.1, 1, 10\}$; $K = \{2, 5, 10\}$; $|V| = \{50, 100, 500\}$



Conclusions

- Discussion:
  - CGS took significantly longer to run then VEM
  - Need to run CGS longer for smaller values of K
  - VEM is biased for both θ and β
  - Recommend cautiously using VEM if user needs results immediately. If time permits, always use CGS.
- Future Directions:
  - Not a fair comparison; should look into collapsed VEM
  - Initial Values
  - LDA may not be appropriate to begin with