Text

Yuki Shiraito

POLSCI 798 Advanced Topics in Quantitative Methodology
University of Michigan

Winter 2019

Analysis of Text

- Measurement using text (Catalinac, 2014. "From Pork to Policy.")
  - Does intraparty competition leads to particularistic campaigning?
  - Measuring the type of campaigns: Topics in manifestos
  - Whether mention particularistic or programmatic goods

Number	Topic Label	%	Classification
1	postal privatization	98.89	prog
2	from concrete to people	98.89	prog
3	appropriator for the district	0.379	**part**
4	fixer-upper for the district	0.379	**part**
5	political reform, protect the constitution	98.89	prog
6	foreign and national security policy	98.89	prog
7	no more unfair taxes, peace constitution	98.89	prog
8	building a society kind to women	51.00	prog
9	primary industries and tourism	0.379	**part**
10	transportation	0.379	**part**
11	no tax increases, no U.S.-Japan alliance	98.89	prog
12	economic recovery	46.21	prog
13	vision for Japan	98.89	prog
14	politics for the civilian, not for bureaucrats	98.89	prog
15	political and administrative reform	98.89	prog

Catalinac (2014) Table D4 in Online Appendix

Modeling Text Data

- Notation
  - *Vocabulary*: Set of all unique words, $\{1, \ldots, L\}$
  - *Word*: L-dimensional standard basis vector, $(0, \ldots, 0, 1, 0, \ldots, 0)^\top$
  - *Document*: Sequece of N_j words, $\mathbf{W}_j = (\mathrm{W}_{1j}, \ldots, \mathrm{W}_{N_j,j})$
  - *Corpus*: Collection of J documents, $(\mathbf{W}_1, \ldots, \mathbf{W}_J)$

- Bag-of-words assumption: Document as *unordered* set of words
  - *Term-document matrix (TDM)*:

$$TDM \equiv \begin{pmatrix} \sum_{i=1}^{N_1} \mathrm{W}_{i1} & \cdots & \sum_{i=1}^{N_J} \mathrm{W}_{iJ} \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} \text{\# of term 1 in doc 1} & \cdots & \text{\# of term 1 in doc J} \\ \vdots & \ddots & \vdots \\ \text{\# of term V in doc 1} & \cdots & \text{\# of term V in doc J} \end{pmatrix}}_{V \times J \text{ matrix of integers}}$$

  - TDM fully represents the data $\Leftrightarrow$ Only term frequencies matter
  - Simple, easy to work with
  - Important dependence across words may be ignored

Latent Dirichlet Allocation

- Lantent Dirichlet Allocation (LDA)
  - One of *topic models*
  - For word i in document j, latent topic Z_{ij} is assigned
  - Z_{ij}: K-dimensional standard basis vector
  - Z_{ij} determines the distribution of W_{ij}
  - The model:
    - Topic assignment of word i, j

$$Z_{ij} \overset{\text{indep.}}{\sim} \text{Multinomial}(\theta_j)$$

    - Observed word W_{ij}

$$W_{ij} \mid Z_{ijk} = 1 \overset{\text{indep.}}{\sim} \text{Multinomial}(\eta_k)$$

  - Prior:
    - Topic proportions of document $\theta_j, j = 1, \ldots, J$

$$\theta_j \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\alpha)$$

    - Word distribution of topic $\eta_k, k = 1, \ldots, K$

$$\eta_k \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\beta)$$

Dirichlet and Multinomial Distributions

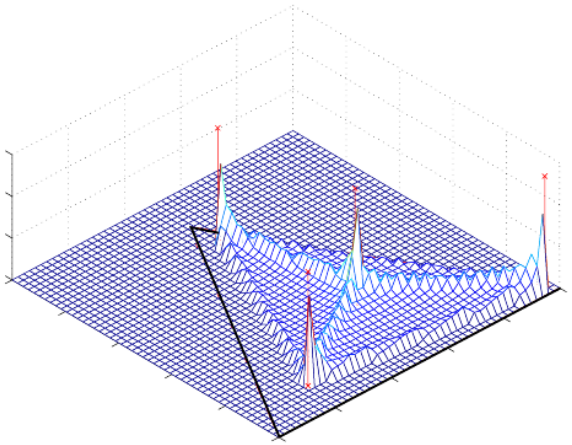
- Multinomial distribution with one trial:

$$p(x) = p_1^{x_1} \cdots p_M^{x_M}$$

  - x: Vector of length M, one element is 1 and $M-1$ elements are 0
  - $M = 2$: Bernoulli distribution
- Dirichlet distribution

$$p(x) = \frac{\Gamma(\sum_{m=1}^{M} \xi_m)}{\prod_{m=1}^{M} \Gamma(\xi_m)} x_1^{\xi_1 - 1} \cdots x_M^{\xi_M - 1}$$

  - x: Vector in the $(M-1)$-simplex, Δ^{M-1}
  - $x \in \Delta^{M-1} \Leftrightarrow x_m \geq 0$ for all m and $\sum_{m=1}^{M} x_m = 1$
  - $M = 2$: Beta distribution



Blei, Ng, and Jordan (2003) Figure 2.

Dirichlet-Multinomial Model

- Dirichlet: Conjugate prior for the multinomial
  $\rightsquigarrow$ analytically tractable compound (mixture) distribution
- Dirichlet-Multinomial model

$$Y \sim \text{Dirichlet}(\xi)$$

$$X \mid Y \sim \text{Multinomial}(Y)$$

  - Marginal probability of X (Gamma tricks!):

$$\begin{aligned}
p(X) &= \int p(X, y)dy = \int p(X \mid y)p(y)dy \\
&= \int \cdots \int y_1^{X_1} \cdots y_M^{X_M} \frac{\Gamma(\sum_{m=1}^M \xi_m)}{\prod_{m=1}^M \Gamma(\xi_m)} y_1^{\xi_1 - 1} \cdots y_M^{\xi_M - 1} dy_1 \cdots dy_M \\
&= \frac{\Gamma(\sum_{m=1}^M \xi_m)}{\prod_{m=1}^M \Gamma(\xi_m)} \int \cdots \int \underbrace{y_1^{\xi_1 + X_1 - 1} \cdots y_M^{\xi_M + X_M - 1}}_{\text{Unnormalized Dirichlet density}} dy_1 \cdots dy_M \\
&= \frac{\Gamma(\sum_{m=1}^M \xi_m)}{\prod_{m=1}^M \Gamma(\xi_m)} \frac{\prod_{m=1}^M \Gamma(\xi_m + X_m)}{\Gamma(\sum_{m=1}^M (\xi_m + X_m))} = \frac{\xi_{m^*}}{\sum_{m=1}^M \xi_m}
\end{aligned}$$

  where m^* is such that $X_{m^*} = 1$

Collapsing Parameters in LDA

- Joint posterior for LDA

$$p(\mathbf{Z}, \theta, \eta \mid \mathbf{W}) \propto \left(\prod_{k=1}^{K} p(\eta_k) \right) \prod_{j=1}^{J} p(\theta_j) \prod_{i=1}^{N_j} p(Z_{ij} \mid \theta_j) p(W_{ij} \mid Z_{ij}, \eta)$$

- High dimensionality of parameters in LDA
  - η_k: K vectors of length L
  - θ_j: J vectors of length K
- Want to collapse (integrate) parameters out

$$p(\mathbf{Z} \mid \mathbf{W}) = \int \int p(\mathbf{Z}, \theta, \eta \mid \mathbf{W}) d\theta d\eta$$

$$\propto \prod_{j=1}^{J} \int p(\theta_j) \prod_{i=1}^{N_j} p(\mathrm{Z}_{ij} \mid \theta_j) d\theta_j$$

$$\times \int \cdots \int \left(\prod_{k=1}^{K} p(\eta_k) \right) \prod_{j=1}^{J} \prod_{i=1}^{N_j} p(\mathrm{W}_{ij} \mid \mathrm{Z}_{ij}, \eta) d\eta_1 \cdots d\eta_K$$

Collapsed Gibbs Sampler

- Collapsed Gibbs Sampler
  - $s = 0$ Initialize $\mathbf{Z}^{(0)}$
  - $s \geq 1$ Loop over all words in the corpus:
    - Draw $Z_{ij}^{(s)}$ conditional on all the other topic assignments
- *Very* simple and fast steps: Random draws from $\{1, \ldots, K\}$

- Posterior probability that $Z_{i^*j^*k} = 1$, i.e., the topic of $W_{i^*j^*}$ is k:

$$p(Z_{i^*j^*k} = 1 \mid \mathbf{Z}_{-(i^*,j^*)}, \mathbf{W})$$

$$\propto \underbrace{\int p(\theta_{j^*}) p(Z_{i^*j^*k} = 1 \mid \theta_{j^*}) \prod_{i \neq i^*} p(Z_{ij^*} \mid \theta_{j^*}) d\theta_{j^*}}_{\text{Dirichlet-Multinomial conditional prior: } p(Z_{i^*j^*k}=1 \mid \mathbf{Z}_{-(i^*,j^*)})}$$

$$\times \underbrace{\int p(\eta_k) p(W_{i^*j^*} \mid Z_{i^*j^*k} = 1, \eta_k) \prod_{\{(i,j) \neq (i^*,j^*) : Z_{ijk}=1\}} p(W_{ij} \mid Z_{ij}, \eta_k) d\eta_k}_{\text{Dirichlet-Multinomial conditional likelihood: } p(W_{i^*j^*} \mid Z_{i^*j^*k}=1, \mathbf{Z}_{-(i^*,j^*)}, \mathbf{W}_{-(i^*,j^*)})}$$

- Conditional prior $p(Z_{i^*j^*k} = 1 \mid \mathbf{Z}_{-(i^*,j^*)})$:

$$\int \underbrace{\frac{\Gamma(\sum_{k'=1}^K a_{k'})}{\prod_{k'=1}^K \Gamma(a_{k'})} \theta_{j^*,1}^{a_1-1} \cdots \theta_{j^*,K}^{a_K-1}}_{p(\theta_{j^*}):\ \text{Dirichlet}} \underbrace{\theta_{j^*,k}}_{p(Z_{i^*j^*,k}=1|\theta_{j^*})}$$

$$\times \underbrace{\theta_{j^*,1}^{\sum_{i\neq i^*} Z_{ij^*,1}} \cdots \theta_{j^*,K}^{\sum_{i\neq i^*} Z_{ij^*,K}}}_{\prod_{i\neq i^*} p(Z_{ij^*}|\theta_{j^*}):\ \text{Multinomial}} d\theta_{j^*}$$

$$\propto \frac{\Gamma(a_k + \sum_{i\neq i^*} Z_{ij^*k} + 1) \prod_{k'\neq k} \Gamma(a_{k'} + \sum_{i\neq i^*} Z_{ij^*k'})}{\Gamma(a_k + \sum_{i\neq i^*} Z_{ij^*k} + 1 + \sum_{k'\neq k}(a_{k'} + \sum_{i\neq i^*} Z_{ij^*k'}))}$$

$$\propto \frac{(a_k + \sum_{i\neq i^*} Z_{ij^*k})\Gamma(a_k + \sum_{i\neq i^*} Z_{ij^*k}) \prod_{k'\neq k} \Gamma(a_{k'} + \sum_{i\neq i^*} Z_{ij^*k'})}{\Gamma(\sum_{k'=1}^K a_{k'} + \underbrace{1 + \sum_{k'=1}^K \sum_{i\neq i^*} Z_{ij^*k'}}_{=N_{j^*}:\ \text{Does not depend on } k})}$$

$$\propto (a_k + \sum_{i\neq i^*} Z_{ij^*k}) \prod_{k'=1}^K \Gamma(a_{k'} + \sum_{i\neq i^*} Z_{ij^*k'}) \propto a_k + \sum_{i\neq i^*} Z_{ij^*k}$$

- $\sum_{i\neq i^*} Z_{ij^*k}$: # of other words with topic k in the same document

- Conditional likelihood $p(W_{ijl} = 1 \mid Z_{i^*j^*k} = 1, \mathbf{Z}_{-(i^*,j^*)}, \mathbf{W}_{-(i^*,j^*)})$

$$\int \frac{\Gamma(\sum_{l'=1}^{L} \beta_{l'})}{\prod_{l'=1}^{L} \Gamma(\beta_{l'})} \eta_{k1}^{\beta_1 - 1} \cdots \eta_{kL}^{\beta_L - 1} \eta_{kl} \times \eta_{k1}^{\sum_{\{(i,j) \neq (i^*,j^*) : Z_{ijk}=1\}} W_{ij1}} \cdots \eta_{kL}^{\sum_{\{(i,j) \neq (i^*,j^*) : Z_{ijk}=1\}} W_{ijL}} d\eta_k$$

$$\propto \frac{\beta_l + \sum_{\{(i,j) \neq (i^*,j^*) : Z_{ijk}=1\}} W_{ijl}}{\sum_{l=1}^{L} \beta_l + \sum_{l=1}^{L} \sum_{\{(i,j) \neq (i^*,j^*) : Z_{ijk}=1\}} W_{ijl}}$$

- $\sum_{\{(i,j) \neq (i^*,j^*) : Z_{ijk}=1\}} W_{ijl}$: # of the same unique word with topic k

- Gibbs step for $Z_{ij}^{(s)}$

$$p(Z_{i^*j^*k} = 1 \mid \mathbf{Z}_{-(i^*,j^*)}, \mathbf{W}) \propto \left(a_k + \sum_{i \neq i^*} Z_{ij^*k} \right) \frac{\beta_l + \sum_{\{(i,j) \neq (i^*,j^*) : Z_{ijk}=1\}} W_{ijl}}{\sum_{l=1}^{L} \beta_l + \sum_{l=1}^{L} \sum_{\{(i,j) \neq (i^*,j^*) : Z_{ijk}=1\}} W_{ijl}}$$

- Simple multinomial sampling (`sample.int()` in **R**) for each word
- Memory efficient: Only need to store $K \times L$ integers, often sparse

EM Algorithm

- Large number of latent variables in LDA
  $\rightsquigarrow$ what about the EM algorithm?
- Log joint posterior density:

$$\log p(\mathbf{Z}, \theta, \eta \mid \mathbf{W}) = \sum_{k=1}^{K} \sum_{l=1}^{L} (\beta_l - 1) \log \eta_{kl} + \sum_{j=1}^{J} \left\{ \sum_{k=1}^{K} (\alpha_k - 1) \log \theta_{jk} + \sum_{i=1}^{N_j} \sum_{k=1}^{K} z_{ijk} \left(\log \theta_{jk} + \sum_{l=1}^{L} w_{ijl} \log \eta_{kl} \right) \right\} + \text{constant}$$

- Consider the EM algorithm where:
  1. E-step: Compute $Q(\theta, \eta \mid \theta^{(s-1)}, \eta^{(s-1)})$
  2. M-step: Maximize $Q(\theta, \eta \mid \theta^{(s-1)}, \eta^{(s-1)})$ w.r.t. θ and η

- Q-function:

$$Q(\theta, \eta \mid \theta^{(s-1)}, \eta^{(s-1)}) \equiv \mathbb{E}_{p(\mathbf{Z} \mid \theta^{(s-1)}, \eta^{(s-1)}, \mathbf{W})}[\log p(\mathbf{Z}, \theta, \eta \mid \mathbf{W})]$$

$$= \sum_{k=1}^{K} \sum_{l=1}^{L} (\beta_l - 1) \log \eta_{kl} + \sum_{j=1}^{J} \left\{ \sum_{k=1}^{K} (\alpha_k - 1) \log \theta_{jk} \right.$$

$$\left. + \sum_{i=1}^{N_j} \sum_{k=1}^{K} \mathbb{E}_{p(\mathbf{Z} \mid \theta^{(s-1)}, \eta^{(s-1)}, \mathbf{W})}[Z_{ijk}] \left(\log \theta_{jk} + \sum_{l=1}^{L} W_{ijl} \log \eta_{kl} \right) \right\}$$

$$+ \text{constant}$$

- Posterior of $\mathbf{Z}$ conditional on θ and η:

$$p(\mathbf{Z} \mid \theta, \eta, \mathbf{W}) \propto \prod_{j=1}^{J} \prod_{i=1}^{N_j} \prod_{k=1}^{K} \left(\theta_{jk} \prod_{l=1}^{L} \eta_{kl}^{W_{ijl}} \right)^{Z_{ijk}}$$

$\Rightarrow Z_{ij}$'s are conditionally independent given θ, η, and $\mathbf{W}$

- Simple E-step:

$$\widetilde{Z}_{ijk^*}^{(s-1)} \equiv \mathbb{E}_{p(\mathbf{Z} \mid \theta^{(s-1)}, \eta^{(s-1)}, \mathbf{W})}[Z_{ijk^*}] = \frac{\theta_{jk^*}^{(s-1)} \eta_{k^* l^*}^{(s-1)}}{\sum_{k=1}^{K} \theta_{jk}^{(s-1)} \eta_{kl^*}^{(s-1)}}$$

- Constrained optimization w.r.t. θ and η:

$$\max_{\theta,\eta} Q(\theta, \eta \mid \theta^{(s-1)}, \eta^{(s-1)}) - \sum_{j=1}^{J} \lambda_{\theta_j} \left(\sum_{k=1}^{K} \theta_{jk} - 1 \right) - \sum_{k=1}^{K} \lambda_{\eta_k} \left(\sum_{l=1}^{L} \eta_{kl} - 1 \right)$$

  where λ_{θ_j} and λ_{η_k} are Lagrange multipliers

- First order conditions:

$$\frac{a_k - 1 + \sum_{i=1}^{N_j} \widetilde{Z}_{ijk}^{(s-1)}}{\theta_{jk}} - \lambda_{\theta_j} = 0 \text{ for all } j \text{ and } k$$

$$\sum_{k=1}^{K} \theta_{jk} - 1 = 0 \text{ for all } j$$

$$\frac{\beta_l - 1 + \sum_{j=1}^{J} \sum_{i=1}^{N_j} \widetilde{Z}_{ijk}^{(s-1)} W_{ijl}}{\eta_{kl}} - \lambda_{\eta_k} = 0 \text{ for all } k \text{ and } l$$

$$\sum_{l=1}^{L} \eta_{kl} - 1 = 0 \text{ for all } k$$

- M-step is simple as well:

$$\theta_{jk}^{(s)} = \frac{a_k - 1 + \sum_{i=1}^{N_j} \widetilde{Z}_{ijk}^{(s-1)}}{\sum_{k'=1}^{K} a_{k'} - K + \sum_{k'=1}^{K} \sum_{i=1}^{N_j} \widetilde{Z}_{ijk'}^{(s-1)}}$$

$$\eta_{kl}^{(s)} = \frac{\beta_l - 1 + \sum_{j=1}^{J} \sum_{i=1}^{N_j} \widetilde{Z}_{ijk}^{(s-1)} W_{ijl}}{\sum_{l'=1}^{L} \beta_{l'} - L + \sum_{l'=1}^{L} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \widetilde{Z}_{ijk}^{(s-1)} W_{ijl'}}$$

- Maximizers $\theta^{(s)}$ and $\eta^{(s)}$ are separately computed
  $\rightsquigarrow$ no need iterative steps as in the two-parameter IRT model

- An EM algorithm for LDA
  1. Initialize $\theta^{(0)}$ and $\eta^{(0)}$
  2. Repeat:
     1. E-step: Compute $\widetilde{Z}_{ijk}^{(s-1)}$ for all (i, j, k)
     2. M-step: Compute $\theta_{jk}^{(s)}$ for all (j, k) and $\eta_{kl}^{(s)}$ for all (k, l)

  until $(\theta^{(s)}, \eta^{(s)})$ converge
- Key: Z_{ij}'s are conditionally independent given θ, η, and $\mathbf{W}$
- Differs from Blei et. al., which approximates $Q(\eta \mid \eta^{(s-1)})$

EM with the Collapsed Posterior?

- The EM algorithm updates high-dimentional θ and η
- What if collapse those parameters?
- Revisit $p(\mathbf{Z} \mid \mathbf{W})$:

$$p(\mathbf{Z} \mid \mathbf{W}) \propto \prod_{j=1}^{J} \frac{\prod_{k=1}^{K} \Gamma(a_k + \sum_{i=1}^{N_j} Z_{ijk})}{\Gamma\left(\sum_{k=1}^{K} (a_k + \sum_{i=1}^{N_j} Z_{ijk})\right)} \times \prod_{k=1}^{K} \frac{\prod_{l=1}^{L} \Gamma(\beta_l + \sum_{j=1}^{J} \sum_{i=1}^{N_j} Z_{ijk} W_{ijl})}{\Gamma(\sum_{l=1}^{L} \beta_l + \sum_{l=1}^{L} \sum_{j=1}^{J} \sum_{i=1}^{N_j} Z_{ijk} W_{ijl})}$$

$$\propto \prod_{j=1}^{J} \prod_{k=1}^{K} \Gamma(a_k) \prod_{m=1}^{\sum_{i=1}^{N_j} Z_{ijk}} (a_k + m) \times \prod_{k=1}^{K} \frac{\prod_{l=1}^{L} \Gamma(\beta_l) \prod_{m=1}^{\sum_{j=1}^{J} \sum_{i=1}^{N_j} Z_{ijk} W_{ijl}} (\beta_l + m)}{\Gamma(\sum_{l=1}^{L} \beta_l) \prod_{m=1}^{\sum_{l=1}^{L} \sum_{j=1}^{J} \sum_{i=1}^{N_j} Z_{ijk} W_{ijl}} (\sum_{l=1}^{L} \beta_l + m)}$$

Mean-Field Approximation

- $p(\mathbf{Z} \mid \mathbf{W}) \neq \prod_{j=1}^{J} \prod_{i=1}^{N_j} p(Z_{ij} \mid \mathbf{W}) \rightsquigarrow \mathbb{E}_{p(\mathbf{Z}|\mathbf{W})}[Z_{ij}]$ is not easy
- Consider an approximate version of the EM algorithm
  - Intractable due to dependence across Z_{ij}'s
  - *Let's approximate assuming independence*
- Recall the key inequality for the EM algorithm:

$$\log p(\psi \mid \text{Data}) = \log \int \frac{p(\psi, x \mid \text{Data})}{p(x \mid \underline{\psi}, \text{Data})} p(x \mid \underline{\psi}, \text{Data}) dz$$

$$\geq \mathbb{E}_{p(X|\underline{\psi}, \text{Data})} \left[\log \frac{p(\psi, X \mid \text{Data})}{p(X \mid \underline{\psi}, \text{Data})} \right]$$

- Analogously using Jensen's inequality,

$$\log p(\text{Data}) = \log \int \int \frac{p(\psi, x, \text{Data})}{q(x \mid \delta) q(\psi \mid \zeta)} q(x \mid \delta) q(\psi \mid \zeta) dz d\psi$$

$$\geq \mathbb{E}_{q(X|\delta)q(\psi|\zeta)} \left[\log \frac{p(\psi, X, \text{Data})}{q(X \mid \delta) q(\psi \mid \zeta)} \right]$$

  for arbitrary distributions $q(X \mid \delta)$ and $q(\psi \mid \zeta)$
- Independence between ψ and X: Mean-field approximation

Variational Inference

- Variational inference:
  - *Variational lower bound* (VLB)
$$\mathbb{E}_{q(X|\delta)q(\psi|\zeta)}\left[\log \frac{p(\psi, X, \text{Data})}{q(X \mid \delta)q(\psi \mid \zeta)}\right] \leq \log p(\text{Data})$$
  - Maximum at $\log p(\text{Data})$ under $q(X \mid \delta)q(\psi \mid \zeta) = p(\psi, X \mid \text{Data})$
  - Never attained if $p(\psi, X \mid \text{Data}) \neq p(\psi \mid \text{Data})p(X \mid \text{Data})$
  - Maximize w.r.t. δ and $\zeta \rightsquigarrow q(X \mid \delta)q(\psi \mid \zeta)$ "close to" $p(\psi, X \mid \text{Data})$
  - EM as a special case
- Alternating maximization:
$$\begin{aligned} VLB =& \mathbb{E}_{q(X|\delta)}\left[\mathbb{E}_{q(\psi|\zeta)}\left[\log p(\psi, X, \text{Data})\right]\right] \\ & - \mathbb{E}_{q(X|\delta)}\left[\log q(X \mid \delta)\right] - \mathbb{E}_{q(\psi|\zeta)}\left[\log q(\psi \mid \zeta)\right] \end{aligned}$$
  1. Maximization w.r.t. δ:
$$\begin{aligned} & \mathbb{E}_{q(X|\delta)}\left[\mathbb{E}_{q(\psi|\zeta)}\left[\log p(\psi, X, \text{Data})\right] - \log q(X \mid \delta)\right] \\ &= -\text{KL}\left(q(X \mid \delta) \middle|\middle| \frac{\exp(\mathbb{E}_{q(\psi|\zeta)}\left[\log p(\psi, X, \text{Data})\right])}{\int \exp(\mathbb{E}_{q(\psi|\zeta)}\left[\log p(\psi, x, \text{Data})\right])dx}\right) + \text{constant} \\ \Rightarrow & q(X \mid \delta^{(s)}) \propto e^{\mathbb{E}_{q(\psi|\zeta^{(s-1)})}[\log p(\psi, X, \text{Data})]} \end{aligned}$$
  2. Maximization w.r.t. ζ: $q(\psi \mid \zeta^{(s)}) \propto e^{\mathbb{E}_{q(X|\delta^{(s)})}[\log p(\psi, X, \text{Data})]}$

Collapsed Variational Inference for LDA

- Log posterior density:

$$\log p(\mathbf{Z}, \mathbf{W}) = \log p(\mathbf{Z} \mid \mathbf{W}) - \underbrace{\log p(\mathbf{W})}_{\text{constant}}$$

$$= \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{m=1}^{\sum_{i=1}^{N_j} Z_{ijk}} \log(a_k + m)$$

$$+ \sum_{k=1}^{K} \left(\sum_{l=1}^{L} \sum_{m=1}^{\sum_{j=1}^{J} \sum_{i=1}^{N_j} Z_{ijk} W_{ijl}} \log(\beta_l + m) \right.$$

$$\left. - \sum_{m=1}^{\sum_{l=1}^{L} \sum_{j=1}^{J} \sum_{i=1}^{N_j} Z_{ijk} W_{ijl}} \log\left(\sum_{l=1}^{L} \beta_l + m \right) \right)$$

$$+ \text{constant}$$

- Independent multinomial variational distributions:

$$q(\mathbf{Z} \mid \kappa) = \prod_{j=1}^{J} \prod_{i=1}^{N_j} q(Z_{ij} \mid \kappa_{ij}) = \prod_{j=1}^{J} \prod_{i=1}^{N_j} \prod_{k=1}^{K} \kappa_{ijk}^{Z_{ijk}}$$

- Variational update of $q(Z_{i^*j^*} \mid \kappa_{i^*j^*})$ given all κ_{ij} s.t. $(i,j) \neq (i^*,j^*)$

$$\widehat{\kappa}_{i^*j^*k^*} = \frac{\exp(\mathbb{E}_{\prod_{(i,j)\neq(i^*,j^*)} q(Z_{ij}|\widehat{\kappa}_{ij})}[\log p(\mathbf{Z}_{-(i^*,j^*)}, Z_{i^*j^*k^*} = 1, \mathbf{W})])}{\sum_{k=1}^{K} \exp(\mathbb{E}_{\prod_{(i,j)\neq(i^*,j^*)} q(Z_{ij}|\widehat{\kappa}_{ij})}[\log p(\mathbf{Z}_{-(i^*,j^*)}, Z_{i^*j^*k} = 1, \mathbf{W})])}$$

$$\propto e^{\mathbb{E}_{\prod_{(i,j)\neq(i^*,j^*)} q(Z_{ij}|\gamma_{ij})}\left[\log(a_{k^*} + \sum_{i\neq i^*} Z_{ijk^*}) + \log(\beta_{l^*} + \sum_{(i,j)\neq(i^*,j^*)} Z_{ijk^*} W_{ijl^*})\right]}$$

$$\times e^{\mathbb{E}_{\prod_{(i,j)\neq(i^*,j^*)} q(Z_{ij}|\gamma_{ij})}\left[-\log(\sum_{l=1}^{L} \beta_l + \sum_{l=1}^{L} \sum_{(i,j)\neq(i^*,j^*)} Z_{ijk^*} W_{ijl^*})\right]}$$

where l^* is such that $W_{i^*j^*l^*} = 1$

- 0th order approximation ("CVB0")

$$\mathbb{E}_q\left[\log\left(a_{k^*} + \sum_{i\neq i^*} Z_{ijk^*}\right)\right] \approx \log\left(a_{k^*} + \underbrace{\mathbb{E}_q\left[\sum_{i\neq i^*} Z_{ijk^*}\right]}_{=\sum_{i\neq i^*} \widehat{\kappa}_{ijk^*}}\right)$$

Summary

- Latent Dirichlet Allocation
  - Workhorse model for statistical text analysis
  - Measurement of "topics"
  - Bag-of-words assumption
- Computational challenges and various estimation algorithms
  - Collapsed Gibbs sampler
  - EM algorithm
  - Variational inference
- Readings for review
  1. LDA
     - Blei et. al. (2003) "Latent Dirichlet Allocation"
  2. Collapsed Gibbs sampler:
     - Griffiths and Steyvers (2004) "Finding Scientific Topics"
  3. Variational inference
     - **BDA3** Sections 13.7
     - **Bishop** Ch. 10
     - Teh et. al. (2007) "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation"