A Bayesian Hierarchical Topic Model for Political Texts: Supplemental Appendix

Justin Grimmer

March 12, 2009

In this supplemental appendix, I derive in detail the variational algorithm used to estimate the expressed agenda model, provide the posterior for a version of the model with multinomial mixture components and the algorithm to estimate that model. The basic approach to the derivation outlined here is fairly standard in the mean-field literature, but I highlight the major steps in the derivation because variational inference is non-standard in political science.

1 Variational Estimation of the Expressed Agenda Model

As stated in the Appendix, variational inference approximates the full posterior distribution with a simpler distribution that allows Bayesian inference to be tractable. Then, we select the member of this distributional family that minimizes a measure of divergence between the true and approximating distribution: the *Kullback-Leibler* divergence (Jordan et al., 1999). Variational methods are regularly used in the machine-learning literature, but I am unaware of another application within political science.

1.1 Full Posterior

We restate the full posterior and data generation process here for convenience.

$$\begin{aligned}
\alpha_k|\delta,\lambda &\sim \text{Gamma}(\lambda,\delta) \text{ for all } k = 1,\ldots,K \\
\boldsymbol{\pi}_i|\boldsymbol{\alpha} &\sim \text{Dirichlet}(\boldsymbol{\alpha}) \text{ for all } i = 1,\ldots,n \\
\boldsymbol{\tau}_{ij}|\boldsymbol{\pi}_i &\sim \text{Multinom}(\boldsymbol{\pi_i}) \text{ for all } j = 1,\ldots,D_i; i = 1,\ldots,n \\
\boldsymbol{\mu}_k|\boldsymbol{\eta}_k,\kappa &\sim \text{vMF}_w(\boldsymbol{\eta},\kappa) \text{ for all } k = 1,\ldots,K \\
\boldsymbol{y}^*_{ij}|\boldsymbol{\mu},\kappa,\tau_{d_i,j}=1 &\sim \text{vMF}_w(\boldsymbol{\mu}_j,\kappa) \text{ for all } j = 1,\ldots,D_i; i = 1,\ldots,n
\end{aligned}$$

with parametric form

$$\begin{aligned}
p(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\tau}|\boldsymbol{Y}) &\propto \prod_{k=1}^{K}\exp(-\alpha_k)\exp(\kappa\boldsymbol{\eta}'\boldsymbol{\mu}_k) \\
\times\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} &\times \prod_{i=1}^{100}\left[\prod_{k=1}^{K}(\pi_{ik})^{\alpha_k-1}\prod_{j=1}^{D_i}\prod_{k=1}^{K}\left[\pi_{ik}\exp(\kappa\boldsymbol{\mu}'\boldsymbol{y}^*_{ij})\right]^{\tau_{ijk}=1}\right] \qquad (1.1)
\end{aligned}$$

1.2 Approximating Distribution

We approximate Equation 1.1 with an approximating distribution that has additional independence, *but we make no assumptions about the parametric form.* Rather, the parametric family (specific distribution) of the approximating distribution emerges as part of the estimation procedure. We will approximate the posterior with $q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})$, which will assume can be factored as $q(\boldsymbol{\pi})q(\boldsymbol{\tau})q(\boldsymbol{\mu})q(\boldsymbol{\alpha})$. This is a standard step in mean-field research and has exhibited quite useful properties in a wide-variety of applications. We will further only estimate *Maximum a posteriori* estimates for both the topic centers $\boldsymbol{\mu}_k$ and the Dirichlet shape parameters $\boldsymbol{\alpha}$: placing all point mass on the MAP estimates. Therefore, we will represent these distributions using the Dirac delta function, which places all point mass on its argument. Represent the MAP parameters for the topic centers as $\boldsymbol{\mu}_k^*$ and for the shape parameters $\boldsymbol{\alpha}^*$ This assumption and the structure of the model allows us to write the approximating distribution as

$$q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha}) \;=\; \prod_{i=1}^{N} q(\boldsymbol{\pi})_i \prod_{i=1}^{N} \prod_{j=1}^{D_i} q(\boldsymbol{\tau})_{ij} \prod_{k=1}^{K} \delta_{\boldsymbol{\mu}_k^*} \delta_{\boldsymbol{\alpha}^*} \tag{1.2}$$

1.3 Minimizing the KL-Divergence

To minimize the KL-divergence between Equation 1.2 and Equation 1.1 we maximize a lower-bound on the *evidence*: the marginal probability of the data (Bishop, 2006). To derive the lower-bound, first write the log-evidence as (Bishop, 2006),

$$\log p(\boldsymbol{Y}) \;=\; \log \sum_{\boldsymbol{\tau}} \iiint p(\boldsymbol{Y}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\alpha}.$$

Insert the approximating distribution $q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})$ by multiplying by 1 (Blei et al., 2003),

$$\log p(\boldsymbol{Y}) \;=\; \log \sum_{\boldsymbol{\tau}} \iiint \frac{q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})}{q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})} p(\boldsymbol{\tau}, \boldsymbol{Y}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\alpha}.$$

Applying Jensen's inequality yields the lower bound (Bishop, 2006)

$$\log p(\boldsymbol{Y}) \;\geq\; \sum_{\boldsymbol{\tau}} \iiint q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha}) \log \left\{ \frac{p(\boldsymbol{\tau}, \boldsymbol{Y}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\mu})}{q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})} \right\} d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\alpha} \tag{1.3}$$

We will define the right hand side of Inequality 1.3 as $\mathcal{L}(q)$. Following Bishop (2006), if we calculate $\log p(\boldsymbol{Y}) - \mathcal{L}(q)$, we find

$$\log p(\boldsymbol{Y}) - \mathcal{L}(q) \;=\; -\sum_{\boldsymbol{\tau}} \iiint q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha}) \frac{p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\tau} | \boldsymbol{Y})}{q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})} d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\alpha}$$

and $-\sum_{\boldsymbol{\tau}} \iiint q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha}) \frac{p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\tau} | \boldsymbol{Y})}{q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\alpha})} d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\alpha}$ is the KL-divergence between the approximating distribution and the true posterior, which we call $\mathrm{KL}(q||p)$. The following identity always holds,

$$\log p(\boldsymbol{Y}) - \mathcal{L}(q) \;=\; \mathrm{KL}(q||p)$$

Notice, that $\log p(\boldsymbol{Y})$ is a fixed number and that $\mathrm{KL}(q||p)$ is always greater than zero. Therefore, as we change q to increase $\mathcal{L}(q)$, the KL-divergence between q and p must decrease. By this same argument, if we choose q to maximize $\mathcal{L}(q)$, then $\mathrm{KL}(q||p)$ must be at a minimum (Bishop, 2006).

Therefore, to minimize the KL-divergence between the approximating distribution and the true-posterior, we will maximize the lower-bound with respect to the approximating distribution– restricted of course to the class of distributions we have assumed. Others (see Bishop (2006)) have shown that $\mathcal{L}(q)$ is convex in the independent components of the approximating distribution, which suggests an iterative algorithm (like EM) can be used to maximize the lower-bound reliably.

1.4 Update Steps

Bishop (2006) shows that the update steps for each independent component of the approximating distributions are easily obtained expected averages. Specifically, the algorithm will iteratively update,

$$\begin{aligned}
q(\boldsymbol{\pi})_i &\propto \exp\left(\mathrm{E}_{\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\alpha}}[\log p(\boldsymbol{Y},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\pi})]\right) \\
q(\boldsymbol{\tau})_{ij} &\propto \exp\left(\mathrm{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\alpha}}[\log p(\boldsymbol{Y},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\pi})]\right) \\
q(\boldsymbol{\mu})_k &\propto \exp\left(\mathrm{E}_{\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\alpha}}[\log p(\boldsymbol{Y},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\pi})]\right) \\
q(\boldsymbol{\alpha}) &\propto \exp\left(\mathrm{E}_{\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu}}[\log p(\boldsymbol{Y},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\pi})]\right)
\end{aligned}$$

or, we iteratively take the expected value of $\log p(\boldsymbol{Y},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\pi})$ over all the parameters other than approximating distribution parameter, where the expectation is taken over the approximating distribution. In this section, I show how to take the update steps.

1.4.1 $q(\boldsymbol{\tau})_{ij}$

Writing out the terms in $\log p(\boldsymbol{Y},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\pi})$ that depend on $\boldsymbol{\tau}_{ij}$,

$$\log q(\boldsymbol{\tau})_{ij} = \sum_{k=1}^{K} \tau_{ijk}\mathrm{E}[\log \pi_{ik}] + \tau_{ijk}\kappa\boldsymbol{\mu}_k^*\boldsymbol{y}_{ij} + \text{constants}$$

where we have used the fact that $\mathrm{E}[\boldsymbol{\mu}] = \boldsymbol{\mu}_k^*$ because we are placing mass only the MAP estimates. Exponentiating both sides we find that this is a Multinomial distribution. Call the probability vector that parameterizes this distribution $\boldsymbol{r}_{ij}$. Some simple mathematics yields that typical element of the vector r_{ijk} is equal to,

$$r_{ijk} = \frac{\exp\left(\mathrm{E}[\log \pi_{ik}] + \kappa\boldsymbol{\mu}_k^*\boldsymbol{y}_{ij}\right)}{\sum_{z=1}^{K}\exp\left(\mathrm{E}[\log \pi_{iz}] + \kappa\boldsymbol{\mu}_z^*\boldsymbol{y}_{ij}\right)}$$

and define $\boldsymbol{r}_{ij} = (r_{ij1},\ldots,r_{ijK})$. Notice, that we need to return to finish this to finish the definition of $\boldsymbol{r}_{ij}$ once we know the distributional form for $q(\boldsymbol{\pi})_i$.

1.4.2 $q(\boldsymbol{\pi})_i$

Writing out the terms that depend upon $\boldsymbol{\pi}_i$,

$$\begin{aligned}\log q(\boldsymbol{\pi})_i &= \sum_{k=1}^{K}(\alpha_k - 1)\log \pi_{ik} + \sum_{j=1}^{D_i}\sum_{k=1}^{K}\mathrm{E}[\tau_{ijk}]\log \pi_{ik} + \text{constants} \\ &= \sum_{k=1}^{K}(\alpha_k - 1)\log \pi_{ik} + \sum_{j=1}^{D_i}\sum_{k=1}^{K} r_{ijk}\log \pi_{ik} + \text{constants} \qquad (1.4)\end{aligned}$$

Looking at Equation 1.4 we see that this is the kernel of the Dirichlet distribution. Therefore, $q(\boldsymbol{\pi})_i$ is a Dirichlet distribution, with shape parameters $\boldsymbol{\theta}_i$, with typical parameter,

$$\theta_{ik} = \alpha_k + \sum_{j=1}^{D_i} r_{ijk}$$

1.4.3 $\delta_{\boldsymbol{\mu}_k}$

First, calculating the expected values

$$\begin{aligned}\log \delta_{\boldsymbol{\mu}_k} &= \sum_{i=1}^{N}\sum_{j=1}^{D_i}\mathrm{E}[\tau_{ijk}]\kappa\boldsymbol{\mu}_k^{'}\boldsymbol{y}_{ij} + \kappa\boldsymbol{\eta} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}\kappa\boldsymbol{\mu}_k^{'}\boldsymbol{y}_{ij} + \kappa\boldsymbol{\eta}\end{aligned}$$

Now, we obtain the maximized values of $\boldsymbol{\mu}_k$, which is a straightforward constrained optimization problem outlined in Banerjee et al. (2005), which we follow closely here. We introduce the Lagrangian λ, with the constraint that $\boldsymbol{\mu}_k^{'}\boldsymbol{\mu}_k = 1$,

$$\log \delta_{\boldsymbol{\mu}_k} = \sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}\kappa\boldsymbol{\mu}_k^{'}\boldsymbol{y}_{ij} + \kappa\boldsymbol{\eta} - \lambda(\boldsymbol{\mu}_k^{'}\boldsymbol{\mu}_k - 1)$$

Differentiating with respect to $\boldsymbol{\mu}_k$, setting equal to zero, and solving yields (Banerjee et al., 2005),

$$\frac{\kappa\left(\sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}\boldsymbol{y}_{ij}^{*} + \boldsymbol{\eta}\right)}{2\lambda} = \boldsymbol{\mu}_k \qquad (1.5)$$

And differentiating with respect to λ we see that $\boldsymbol{\mu}_k^{'}\boldsymbol{\mu}_k = 1$, or $||\boldsymbol{\mu}_k^{'}\boldsymbol{\mu}_k|| = 1$. Substituting in Equation 1.5 we find

$$\begin{aligned}\frac{\kappa}{2\lambda}\left(\left(\sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}\boldsymbol{y}_{ij}^{*} + \boldsymbol{\eta}\right)^{'}\left(\sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}\boldsymbol{y}_{ij}^{*} + \boldsymbol{\eta}\right)\right)^{1/2} &= 1 \\ \frac{k||\sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}\boldsymbol{y}_{ij}^{*} + \boldsymbol{\eta}||}{2} &= \lambda\end{aligned}$$

Substituting in we have,

$$\boldsymbol{\mu}_k^* = \frac{\sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}\boldsymbol{y}_{ij}^* + \boldsymbol{\eta}}{||\sum_{i=1}^{N}\sum_{j=1}^{D_i} r_{ijk}\boldsymbol{y}_{ij}^* + \boldsymbol{\eta}||}$$

1.4.4 δ_{α^*}

A closed form update for the $\boldsymbol{\alpha}$ parameters does not exist. Therefore, we use a straightforward and computationally efficient Newton-Raphson algorithm developed in Blei et al. (2003).

1.4.5 Completing $q(\boldsymbol{r})_{ij}$

To finish the updates, we need to calculate $\mathrm{E}[\log \pi_{ik}]$. Blei et al. (2003) show that if $\boldsymbol{\pi}_i$ is Dirichlet$(\boldsymbol{\theta}_i)$ then, $\mathrm{E}[\log \pi_{ik}] = \Psi(\theta_{ik}) - \Psi(\sum_{z=1}^{K}\theta_{iz})$ where $\Psi(\cdot)$ is the digamma function, $\Psi(x) = \frac{\partial \Gamma(x)}{\partial x}$. Therefore,

$$r_{ijk} \propto \exp\left(\Psi(\theta_{ik}) - \Psi(\sum_{z=1}^{K}\theta_{iz}) + \kappa\boldsymbol{\mu}_k^*\boldsymbol{y}_{ij}\right)$$

1.5 Calculating Lower Bound

Convergence of the algorithm is calculated with the lower-bound,

$$\begin{aligned}
\mathcal{L}(q) = & \sum_{i=1}^{N}(\sum_{j=1}^{D_i}\sum_{k=1}^{K} r_{ijk}\kappa\boldsymbol{\mu}_k^{'}\boldsymbol{y}_{ij}^* + \sum_{j=1}^{D_i}\sum_{k=1}^{K} r_{ijk}\left(\Psi(\theta_{ik}) - \Psi(\sum_{j=1}^{K}\theta_{ij})\right) + \log\Gamma(\sum_{k=1}^{K}\alpha_k) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) \\
& + \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\theta_{ik}) - \Psi(\sum_{j=1}^{K}\theta_{ij})\right) - \log\Gamma(\sum_{k=1}^{K}\theta_{ik}) + \sum_{k=1}^{K}\log\Gamma(\theta_{ik}) \\
& - \sum_{k=1}^{K}(\theta_{ik} - 1)\left(\Psi(\theta_{ik}) - \Psi(\sum_{j=1}^{K}\theta_{ij})\right) - \sum_{j=1}^{D_i}\sum_{k=1}^{K} r_{ijk}\log r_{ijk} + \sum_{k=1}^{K}\kappa\boldsymbol{\eta}^{'}\boldsymbol{\mu}_k + \sum_{k=1}^{K} -\alpha_k.) \qquad (1.6)
\end{aligned}$$

2 Expressed Agenda Model with Multinomial Topic Components

We can write out the full hierarchical model for the expressed agenda model with multinomial components,

$$\begin{aligned}
\alpha_k &\sim \text{Exponential}(1) \text{ for } k = 1, \ldots, K \\
\boldsymbol{\pi}_i|\boldsymbol{\alpha} &\sim \text{Dirichlet } (\boldsymbol{\alpha}) \text{ for } i = 1, \ldots, N \\
\boldsymbol{\theta}_k|\boldsymbol{\lambda} &\sim \text{Dirichlet } (\boldsymbol{\lambda}) \text{ for } k = 1, \ldots, K \\
\boldsymbol{\tau}_{ij}|\boldsymbol{\pi}_i &\sim \text{Multinomial } (1, \boldsymbol{\pi}_i) \text{ for } j = 1, \ldots, D_i; i = 1, \ldots, N \\
\boldsymbol{y}_{ij}|\tau_{ijk} = 1, \boldsymbol{\theta} &\sim \text{Multinomial } (n_{ij}, \boldsymbol{\theta}_k) \text{ for } j = 1, \ldots, D_i \text{ for } i = 1, \ldots, N
\end{aligned}$$

where n_{ij} counts the number of words in the j^{th} document from the i^{th} senator. This implies the following posterior distribution,

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\theta} | \boldsymbol{Y}, \boldsymbol{\lambda}) \quad \propto \quad & \prod_{k=1}^{K} \exp(-\alpha_k) \times \frac{\Gamma(\sum_{w=1}^{W} \lambda_w)}{\prod_{w=1}^{W} \Gamma(\lambda_w)} \prod_{w=1}^{W} \theta_{k,w}^{\lambda_w - 1} \times \\ & \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{i=1}^{N} \left[\prod_{k=1}^{K} (\pi_{ik})^{\alpha_k - 1} \prod_{j=1}^{D_i} \prod_{k=1}^{K} \left[\pi_{ik} \prod_{w=1}^{W} \theta_{w,k}^{y_{ijw}} \right]^{\tau_{ijk}} \right] \end{aligned} \tag{2.1}$$

2.1 Approximating Distribution

As is standard in the application of variational inference to mixture problems, I approximate the true posterior with a factorized distribution,

$$q(\boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\theta}) \quad = \quad q(\boldsymbol{\alpha}) \prod_{i=1}^{N} q(\boldsymbol{\pi}^i) \prod_{k=1}^{K} q(\boldsymbol{\theta}_k) \prod_{i=1}^{N} \prod_{j=1}^{D_i} q(\boldsymbol{\tau}_{ij}) \tag{2.2}$$

but we make no assumptions about the functional form of the approximating distribution. But, we do restrict $q(\boldsymbol{\alpha})$ to place all of its point-mass on the MAP solution $\boldsymbol{\alpha}^*$.

2.2 Estimation Steps

As in the previous section with the vMF components, we will estimate an iterative algorithm that will sequentially update the following equations to maximize a lower-bound on the evidence of the data (equivalently, minimize the KL-divergence between the true posterior and approximating distribution) (Bishop, 2006),

$$\begin{aligned} q(\boldsymbol{\tau})_{ij} \quad &\propto \quad \exp\left(\mathrm{E}_{\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\theta}} \left[\log p(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\tau})\right]\right) \\ q(\boldsymbol{\pi})_{i} \quad &\propto \quad \exp\left(\mathrm{E}_{\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\theta}} \left[\log p(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\tau})\right]\right) \\ q(\boldsymbol{\theta})_{k} \quad &\propto \quad \exp\left(\mathrm{E}_{\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\pi}} \left[\log p(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\tau})\right]\right) \\ q(\boldsymbol{\alpha}) \quad &\propto \quad \exp\left(\mathrm{E}_{\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\pi}} \left[\log p(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\tau})\right]\right) \end{aligned}$$

2.2.1 $q(\boldsymbol{\tau})_{ij}$

As in the previous section, we write out the components dependent upon $\boldsymbol{\tau}_{ij}$ and then uncover the remaining distribution. Carrying this out,

$$\log q(\boldsymbol{\tau})_{ij} \quad = \quad \sum_{k=1}^{K} \left[\tau_{ijk} \mathrm{E}[\log \pi_{ik}] + \tau_{ijk} \sum_{w=1}^{W} y_{ijw} \mathrm{E}[\log \theta_{kw}] \right] + \text{constants}$$

This is the kernel of the Multinomial distribution, which again we will call the parameters of the distribution $\boldsymbol{r}_{ij}$ with typical element equal to

$$r_{ijk} \quad = \quad \frac{\exp\left(\mathrm{E}[\log \pi_{ik}] + \sum_{w=1}^{W} y_{ijw} \mathrm{E}[\log \theta_{kw}]\right)}{\sum_{z=1}^{K} \exp\left(\mathrm{E}[\log \pi_{iz}] + \sum_{w=1}^{W} y_{ijw} \mathrm{E}[\log \theta_{zw}]\right)}$$

Once again, we will return to complete this update steps once we have obtained the other distributional forms.

2.2.2 $q(\boldsymbol{\pi})_i$

Writing out the terms that depend upon $\boldsymbol{\pi}_i$,

$$\log q(\boldsymbol{\pi})_i \quad = \quad \sum_{k=1}^{K} \log \pi_{ik} + \sum_{j=1}^{D_i} \sum_{k=1}^{K} \mathrm{E}[\tau_{ijk} \log \pi_{ik} + \text{constants}$$

which is the kernel of a Dirichlet($\boldsymbol{\gamma}_i$) distribution, where $\boldsymbol{\gamma}_i$ is the vector of shape parameters for the Dirichlet. $\boldsymbol{\gamma}_i$ has typical element γ_{ik},

$$\gamma_{ik} \quad = \quad \alpha_k + \sum_{j=1}^{D_i} r_{ijk}$$

2.2.3 $q(\boldsymbol{\theta})_k$

Writing out the terms that depend upon $\boldsymbol{\theta}_k$,

$$\log q(\boldsymbol{\theta})_k \quad = \quad \sum_{w=1}^{W} (\lambda_w - 1) \log \theta_{kw} + \sum_{j=1}^{D_i} \sum_{k=1}^{K} \mathrm{E}[\tau_{ijk}] \sum_{w=1}^{W} y_{ijw} \log \theta_{kw} + \text{constants}$$

which is another Dirichlet($\boldsymbol{\eta}_k$) distribution, where $\boldsymbol{\eta}_k$ parameterizes the Dirichlet distribution, with typical element,

$$\eta_{kw} \quad = \quad \lambda_w + \sum_{i=1}^{N} \sum_{j=1}^{D_i} r_{ijk} y_{ijw}$$

2.2.4 $q(\boldsymbol{\alpha})$

A closed form update does not exist, so we use the Newton-Raphson algorithm from Blei et al. (2003) to efficiently optimize the α parameters.

2.2.5 Finishing $q(\boldsymbol{\tau})_{ij}$

In light of the update steps, we find that

$$\begin{aligned} \mathrm{E}[\log \pi_{ik}] \quad &= \quad \Psi(\gamma_{ik}) - \Psi(\sum_{z=1}^{K} \gamma_{iz}) \\ \mathrm{E}[\log \theta_{kw}] \quad &= \quad \Psi(\eta_{kw}) - \Psi(\sum_{z=1}^{W} \eta_{kz}) \end{aligned}$$

which shows that,

$$r_{ijk} \quad \propto \quad \exp\left(\Psi(\gamma_{ik}) - \Psi(\sum_{z=1}^{K} \gamma_{iz}) + \sum_{w=1}^{W} y_{ijw} \left[\Psi(\eta_{kw}) - \Psi(\sum_{z=1}^{W} \eta_{kz}) \right] \right).$$

2.3 Lower Bound

The lower bound used to assess convergence is,

$$
\begin{aligned}
\mathcal{L}(q) \;=\; & \sum_{i=1}^{N}\Bigg[\log\Gamma(\sum_{k=1}^{K}\alpha_k)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)+\sum_{k=1}^{K}(\alpha_k-1)\left[\Psi(\gamma_{ik})-\Psi(\sum_{j=1}^{K}\gamma_{ik})\right] \\
& +\sum_{j=1}^{D_i}\sum_{k=1}^{K}r_{ijk}\left(\Psi(\gamma_{ik})-\Psi(\sum_{j=1}^{K}\gamma_{ij})\right) \\
& +\sum_{j=1}^{D_i}\sum_{k=1}^{K}r_{ijk}\sum_{w=1}^{W}y_{ijw}\left[\Psi(\eta_{kw})-\Psi(\sum_{j}^{w}\eta_{kj})\right]\Bigg] \\
& +\log\Gamma(\sum_{w=1}^{W}\lambda_w)-\sum_{w=1}^{W}\log\Gamma(\lambda_w)+\sum_{k=1}^{K}\sum_{w=1}^{W}(\lambda_w-1)\left[\Psi(\eta_{kw})-\Psi(\sum_{w=1}^{W}\eta_{k,w})\right] \\
& -\sum_{i=1}^{N}\Bigg[\log\Gamma(\sum_{j=1}^{K}\gamma_{ik})-\sum_{k=1}^{K}\log\Gamma(\gamma_{ik})+\sum_{k=1}^{K}(\gamma_{ik}-1)\left[\Psi(\gamma_{ik})-\Psi(\sum_{j=1}^{K}\gamma_{ij})\right] \\
& +\sum_{j=1}^{D_i}\sum_{k=1}^{K}r_{ijk}\log r_{ijk}\Bigg] \\
& -\sum_{k=1}^{K}\left[\log\Gamma(\sum_{w=1}^{W}\eta_{k,w})-\sum_{w=1}^{W}\log\Gamma(\eta_{k,w})+\sum_{w=1}^{w}(\eta_{k,w}-1)\left(\Psi(\eta_{k,w})-\Psi(\sum_{j=1}^{w}\eta_{k,w})\right)\right]
\end{aligned}
$$

References

Banerjee, Arindam, Inderjit S Dhillon, Joydeep Ghosh and Suvrit Sra. 2005. "Clustering on the Unit Hypersphere using von Mises-Fisher Distributions." *Journal of Machine Learning Research* 6:1345–1382.

Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning*. Springer.

Blei, David et al. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning and Research* 3:993–1022.

Jordan, Michael et al. 1999. "An Introduction to Variational Methods for Graphical Models." *Machine Learning* 37:183–233.