Text Mining Homework - Week 2

Aimee Barciauskas, Felix Gutmann, Guglielmo Pelino, Thomas Vicente

Exercise 2

(a) The parameters are:

- $\{\rho_k\}_{k=1,\dots,K}$ for the latent variables;
- $\{\beta_k^1\}_{k=1,\dots,K}$ for the first distribution, where each β_k^1 is a V_1 dimensional probability vector (i.e. belonging to the (V_1-1)-simplex);
- $\{\beta_k^2\}_{k=1,\dots,K}$ for the second distribution, where each β_k^2 is a V_2 dimensional probability vector.

The observed data are the two vector of counts matrices which we will denote by $\mathbf{X}^1$ and $\mathbf{X}^2$; finally, the latent variables are the z_i's.

(b) Denoting the complete likelihood as $L(\mathbf{X}^1, \mathbf{X}^2, \mathbf{z} \,|\, \boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2)$, we observe that the joint distribution for a single observation with $z_i = k$ (where x_i^1 is the i-th row of $\mathbf{X}^1$) can be written as

$$\begin{aligned}
&P(x_i^1, x_i^2, z_i = k \,|\, \boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2) \\
&\qquad = P(x_i^1, x_i^2 \,|\, \boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2, z_i = k) P(z_i = k \,|\, \boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2) = [\text{cond. independence of demands}] \\
&\qquad = P(x_i^1 \,|\, \boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2, z_i = k) P(x_i^2 \,|\, \boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2, z_i = k) P(z_i = k \,|\, \boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2) \\
&\qquad = \prod_{v_1=1}^{V_1} (\beta_{k,v_1}^1)^{x_{i,v_1}^1} \prod_{v_2=1}^{V_2} (\beta_{k,v_2}^2)^{x_{i,v_2}^2} \rho_k.
\end{aligned}$$

Using this expression we can write in general,

$$P(x_i^1, x_i^2, z_i \,|\, \boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2) = \prod_k \left[\rho_k \prod_{v_1} (\beta_{k,v_1}^1)^{x_{i,v_1}^1} \prod_{v_2} (\beta_{k,v_2}^2)^{x_{i,v_2}^2} \right]^{\mathbb{1}_{(z_i=k)}}.$$

Thus, for the independence between the different observations, the complete likelihood is:

$$L(\boldsymbol{X}^1, \boldsymbol{X}^2, \boldsymbol{z}) = \prod_i \prod_k \left[\rho_k \prod_{v_1}^{V_1} (\beta_{k,v_1}^1)^{x_{i,v_1}^1} \prod_{v_2}^{V_2} (\beta_{k,v_2}^2)^{x_{i,v_2}^2} \right]^{\mathbb{1}_{(z_i=k)}}.$$

Taking the log we get the complete data log-likelihood:

$$l(\boldsymbol{X}^1, \boldsymbol{X}^2, \boldsymbol{z}) = \sum_i \sum_k \mathbb{1}_{(z_i=k)} \left[\log(\rho_k) + \sum_{v_1}^{V_1} x_{i,v_1}^1 \log(\beta_{k,v_1}^1) + \sum_{v_2}^{V_2} x_{i,v_2}^2 \log(\beta_{k,v_2}^2) \right].$$

(c) In the n-th E-step of the algorithm we compute the expected value of the complete log-likelihood w.r.t. the conditional distribution of $\boldsymbol{z} \,|\, \boldsymbol{X}^1, \boldsymbol{X}^2, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2$, where $\boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2$ are the parameter values in the current iteration.
Given this conditional distribution, all we have to compute is

$$\mathbb{E}(\mathbb{1}_{(z_i=k)} \,|\, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2, \boldsymbol{X}^1, \boldsymbol{X}^2) = P(z_i = k \,|\, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2, \boldsymbol{X}^1, \boldsymbol{X}^2) \equiv \hat{z}_{i,k}^n \,,$$

because the other terms in l are not functions of $\boldsymbol{z}$.
By Bayes formula we can compute

$$\begin{aligned}
\hat{z}_{i,k}^n &= P(z_i = k \,|\, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2, x_i^1, x_i^2\,) \\
&\propto P(x_i^1, x_i^2 \,|\, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2, z_i = k\,) P(z_i = k \,|\, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2\,) = [\text{cond. independence of demands}] \\
&= \rho_k^n P(x_i^1 \,|\, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2, z_i = k\,) P(x_i^2 \,|\, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2, z_i = k\,) \\
&= \rho_k^n \prod_{v_1}^{V_1} (\beta_{k,v_1}^{(n,1)})^{x_{i,v_1}^1} \prod_{v_2}^{V_2} (\beta_{k,v_2}^{(n,2)})^{x_{i,v_2}^2}.
\end{aligned}$$

Thus, we can write the Q function as

$$Q(\boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2) = \sum_i \sum_k \hat{z}_{i,k}^n [\log(\rho_k) + \sum_{v_1}^{V_1} x_{i,v_1}^1 \log(\beta_{k,v_1}^1) + \sum_{v_2}^{V_2} x_{i,v_2}^2 \log(\beta_{k,v_2}^2)].$$

We note that Q depends on both the current iteration values $\boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2$ because of $\hat{z}_{i,k}^n$ and on $\boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2$ because of the second part of the expression.

(d) For the M step we have to maximize Q w.r.t. the parameter values, with the constraints on the probability vectors $\boldsymbol{\rho}, \boldsymbol{\beta}^1, \boldsymbol{\beta}^2$.
The associated Lagrangian is the following:

$$Q(\boldsymbol{\rho}, \boldsymbol{B}^1, \boldsymbol{B}^2, \boldsymbol{\rho}^n, \boldsymbol{B}_n^1, \boldsymbol{B}_n^2) + \nu(1 - \sum_k \rho_k) + \sum_k \lambda_{k,1}(1 - \sum_{v_1} \beta_{k,v_1}^1) + \sum_k \lambda_{k,2}(1 - \sum_{v_2} \beta_{k,v_2}^2).$$

Taking the derivative w.r.t. ρ_j and setting it to 0 we find

$$\frac{\partial}{\partial \rho_j} = \sum_i \hat{z}_{i,j}^n \frac{1}{\rho_j} - \nu = 0,$$

and thus

$$\rho_j = \sum_i \hat{z}_{i,j}^n \frac{1}{\nu}.$$

By summing over j in the last expression, and recalling our constraint on $\boldsymbol{\rho}$ which is a probability vector, we obtain:

$$1 = \sum_{i,j} \hat{z}_{i,j}^n \frac{1}{\nu},$$

which implies $\nu = \sum_{i,j} \hat{z}_{i,j}^n$.
This finally gives us the expression for the updated parameter for ρ_k^{n+1}, $k = 1, \ldots, K$ in the $n+1$-th iteration of the algorithm:

$$\rho_k^{n+1} = \frac{\sum_i \hat{z}_{i,k}^n}{\sum_{i,k} \hat{z}_{i,k}^n}.$$

Moreover, maximizing over β's we obtain

$$\frac{\partial}{\partial \beta_{j,v_1}^1} = \sum_i \left[\hat{z}_{i,j}^n x_{i,v_1}^1 \frac{1}{\beta_{j,v_1}^1} \right] - \lambda_{j,1} = 0$$

which in turns can be rewritten as

$$\sum_i \hat{z}_{i,j}^n x_{i,v_1}^1 - \beta_{j,v_1}^1 \lambda_{j,1} = 0.$$

Summing over v_1 and recalling that by definition $\sum_{v_1} \beta_{j,v_1}^1 = 1$ for all j's, we find

$$\sum_{v_1} \sum_i \hat{z}_{i,j}^n x_{i,v_1}^1 = \lambda_{j,1},$$

which finally gives for each $k = 1, \dots, K$

$$\beta_{k,v_1}^{1,(n+1)} = \frac{\sum_i \hat{z}_{i,k}^n x_{i,v_1}^1}{\sum_i \hat{z}_{i,k}^n \sum_{v_1} x_{i,v_1}^1},$$

and repeating the same steps for β^2 we find

$$\beta_{k,v_2}^{2,(n+1)} = \frac{\sum_i \hat{z}_{i,k}^n x_{i,v_2}^2}{\sum_i \hat{z}_{i,k}^n \sum_{v_2} x_{i,v_2}^2}.$$

(e) Noting that in all updates formulae $\hat{z}_{i,k}^n$ is present both in the numerator and denominator, and given that we know its value up to a normalization constant (see above when we first computed it), we can actually use the un-normalized version, which we denote by $\xi_{i,k}^n$, to compute the updated parameters in each iteration. The result is the following algorithm:

Algorithm 1 EM algorithm pseudo-code

Initialize parameters $\boldsymbol{\rho}^0, \mathbf{B}_1^0, \mathbf{B}_2^0$
FOR $n > 0,$ while a stopping condition is not met, **DO**
Compute $\xi_{i,k}^n = \rho_k^n \prod_{v_1} (\beta_{k,v_1}^{n,1})^{x_{i,v_1}} \prod_{v_2} (\beta_{k,v_2}^{n,2})^{x_{i,v_2}}$
Update:
$\rho_k^{n+1} = \frac{\sum_i \xi_{i,k}^n}{\sum_i \sum_k \xi_{i,k}^n}$
$\beta_{k,v_1}^{(n+1,1)} = \frac{\sum_i \xi_{i,k}^n x_{i,v_1}^1}{\sum_i \xi_{i,k}^n \sum_{v_1} x_{i,v_1}^1}$
$\beta_{k,v_2}^{(n+1,2)} = \frac{\sum_i \xi_{i,k}^n x_{i,v_2}^2}{\sum_i \xi_{i,k}^n \sum_{v_2} x_{i,v_2}^2}$

As a stopping condition, one can typically fix a threshold and stop as soon as the difference between the updated parameter and the previous value is smaller than that threshold.