Assignment 3 - DD2434 Machine Learning, Advanced Course

August Regnell 970712-9491

22 January, 2021

Contents

1 3.1 Easier EM for Advertisements

Already given credits for.

NOTE: An attempted solution to the previous question can be found in the Appendix. However, the solution is incorrect sine it has an error with the graphical model.

2 3.2 Hard EM for advertisements

Removed.

NOTE: An attempted solution to the removed question can be found in the Appendix. However, the solution is incorrect sine it has an error with the graphical model.

3 3.3 Complicated likelihood for leaky units on a tree

Definitions

Consider a binary tree T where the vertices $u \in V(T)$ have an observable variable X_u and a latent class variable Z_u. The class variables are i.i.d., taking the values $c \in [C]$, following the categorical distribution π. X_u, conditional on its neighbors (v_1, v_2, v_3) and its own latent variable has a normal distribution:

$$p(X_u \mid Z_u = c, Z_{v_1} = c_1, Z_{v_2} = c_2, Z_{v_3} = c_3) \sim N(X_u \mid (1-\alpha)\mu_c + \sum_{i \in [3]} \frac{1}{3}\alpha\mu_{c_i}, \sigma^2) \tag{1}$$

The task is to find a linear time algorithm (w.r.t. $N = \mid V(T) \mid$) that computes $P(\mathbf{X} \mid T, M, \sigma, \alpha, \pi)$ where T is the tree, $\mathbf{X} = \{X_v : v \in V(T)\}$, $M = \{\mu_c : c \in [C]\}$, σ, α and π.

Define $\mathbf{Z} := \{Z_v : v \in V(T)\}$, $pa(u)$ as the parent node of u and $de_i(u)$ as the i:th descendant node of u where $i \in \{1, 2\}$.

Note that according to Eq (1) the X_u:s are normally distributed, meaning that we can extend the probability of a node as follows:

$$\begin{aligned} &p\big(X_u \mid Z_u = c_1, Z_{pa(u)} = c_2, Z_{de_1(u)} = c_3, Z_{de_2(u)} = c_4\big) \\ =&\frac{1}{\sigma\sqrt{2\pi}} \exp\Big\{\frac{1}{2\sigma^2}\big(X_u - (1-\alpha)\mu_{c_1} - \sum_{i=2}^{4}\frac{1}{3}\alpha\mu_{c_i}\big)^2\Big\} \end{aligned} \tag{2}$$

Furthermore, according to Jens on slack, I will set the equivalent probability of the root and the leaves (having less neighbours than the nodes "inside" the tree) to

$$\begin{aligned} &p\big(X_r \mid Z_r = c_1, Z_{de_1(r)} = c_2, Z_{de_2(r)} = c_3\big) \\ =&\frac{1}{\sigma\sqrt{2\pi}} \exp\Big\{\frac{1}{2\sigma^2}\big(X_r - (1-\alpha)\mu_{c_1} - \sum_{i=2}^{3}\frac{2}{3}\alpha\mu_{c_i}\big)^2\Big\} \end{aligned} \tag{3}$$

and

$$p\big(X_u \mid Z_u = c_1, Z_{pa(u)} = c_2\big) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big\{\frac{1}{2\sigma^2}\big(X_u - (1-\alpha)\mu_{c_1} - \frac{1}{3}\alpha\mu_{c_2}\big)^2\Big\} \tag{4}$$

respectively. Notice that the other parameters have been omitted for clarity.

Solution

Looking at Eq (1) we see that the X_u:s are conditionally dependent given their own latent variable and the latent variable of their three (for inner nodes, less for the root and the leaves) neighbouring nodes, their parent and their descendant nodes. Mathematically this can be expressed as:

$$p(X_u \mid \mathbf{Z}) = p(X_u \mid Z_u, Z_{pa(u)}, Z_{de_1(u)}, Z_{de_2(u)}) \tag{5}$$

That is, $p(\mathbf{X} \mid \mathbf{Z})$ can be expressed as

$$p(\mathbf{X} \mid \mathbf{Z}) = \prod_{u \in V(T)} p(X_u \mid Z_u, Z_{pa(u)}, Z_{de_1(u)}, Z_{de_2(u)}) \tag{6}$$

where $p(X_u \mid Z_u, Z_{pa(u)}, Z_{de_1(u)}, Z_{de_2(u)})$ becomes $p(X_r \mid Z_r, Z_{de_1(r)}, Z_{de_2(r)})$ for the root and $p(X_u \mid Z_u, Z_{pa(u)})$ for the leaves, where r denotes the root node.

We also know that the latent variables are i.i.d. and follows a categorical distribution π, so $p(Z_u = c) = \pi_c$ for all $u \in V(T)$.

Combining these we can calculate the wanted probability. Since only $\mathbf{X}$ is observed (and the parameters), and their values are dependent on the latent variables, we must marginalize out the latent variables.

Using Bayes' rule and marginalizing over $\mathbf{Z}$ we get

$$p(\mathbf{X}) = \sum_{\mathbf{Z}} p(\mathbf{X} \mid \mathbf{Z}) p(\mathbf{Z}) \tag{7}$$

However, this implementation is extremely naive. Since there are N latent variables with C states, there are C^N values that needs to be evaluated when the expression is in the above form. This is clearly not linear in N.

By applying a dynamic programming approach and by exploiting the conditional indecencies, we can move the sums into the product and break the problem down into smaller problems.

Applying Eq (6) and using the Z_u:s i.i.d. categorical distribution, Eq (7) can be extended to

$$\begin{aligned} p(\mathbf{X}) &= \sum_{\mathbf{Z}} \prod_{u \in V(T)} p(X_u \mid Z_u, Z_{pa(u)}, Z_{de_1(u)}, Z_{de_2(u)}) p(Z_u) \\ &= \sum_{\mathbf{Z}} \prod_{u \in V(T)} p(X_u \mid Z_u, Z_{pa(u)}, Z_{de_1(u)}, Z_{de_2(u)}) \pi_{Z_u} \end{aligned} \tag{8}$$

We now define

$$\begin{aligned} \mu_u(c_1, c_2) = \sum_{c_3 \in [C]} \sum_{c_4 \in [C]} \Big[& p\big(X_u \mid Z_u = c_1, Z_{pa(u)} = c_2, Z_{de_1(u)} = c_3, Z_{de_2(u)} = c_4\big) \cdot \\ & \cdot \pi_{c_1} \mu_{de_1(u)}(c_1, c_3) \mu_{de_2(u)}(c_1, c_4) \Big] \end{aligned} \tag{9}$$

where $\mu_u(c_1, c_2) = \sum_{c_3 \in [C]} \sum_{c_4 \in [C]} p\big(X_u \mid Z_u = c_1, Z_{pa(u)} = c_2\big) \pi_{c_1}$ if $u \in L(T)$, where $L(T)$ are the leaves of the tree T.

The wanted likelihood (with the other parameters still omitted for clarity) can now be expressed as

$$\begin{aligned} p(\mathbf{X}) = \sum_{c_1 \in [C]} \sum_{c_2 \in [C]} \sum_{c_3 \in [C]} \Big[& p\big(X_u \mid Z_r = c_1, Z_{de_1(r)} = c_2, Z_{de_2(r)} = c_3\big) \cdot \\ & \cdot \pi_{c_1} \mu_{de_1(r)}(c_1, c_2) \mu_{de_2(r)}(c_1, c_3) \Big] \end{aligned} \tag{10}$$

where the index r denotes the root of the tree.

We will now calculate the complexity of this algorithm. Looking at Eq (9) we see that we "visit" every node once (but sum over them several times). Furthermore, all inner nodes are dependent on the values of four latent variables, which we sum over their C values. This gives a cost of $\mathcal{O}(NC^4)$, which clearly is linear in N. We have thus found an algorithm that satisfies the requirements.

4 3.4 Easier VI for Covid-19

Definitions

In the description, Z_d^k is the latent variable of worker k's status during day d, where the worker can be non-infected, n, infected, i, or have antibodies a. When a worker is infected, it will be infected for 9 days before getting antibodies or returning to the non-infected state.

I will define Z_d^k in such a way that one can keep track of the number of days an individual has been infected.

$$Z_d^k = \begin{cases} 0, & \text{if the worker is non-infected} \\ i, & \text{if the worker is infected, } i \in \{1,\dots,9\} \\ 10, & \text{if the worker has antibodies} \end{cases}$$

where i is on which day the worker is on during the infection, starting at 1.
Now let $\mathbf{Z}^k = Z_1^k,\dots,Z_D^k$ and $\mathbf{Z} = \mathbf{Z}^1,\dots,\mathbf{Z}^K$. We also define $\mathbf{Z}_d = Z_d^1,\dots,Z_d^K$ to make notation simpler.

We are given the contact graph G_d and the absence table A_d. The probability of getting infected when one have met **at least one**[1] infected individual when one is non-infected is ι. The probability of getting antibodies after being infected for 9 days is α. The probability of staying home (for each day during the infection) when one is infected is σ. These parameters are Bernoulli distributed with Beta priors and are kept in $\Omega = (\iota, \alpha, \sigma)$. Define also $A = A_1,\dots,A_D$ and $G = G_1,\dots,G_D$. Let $A_d^k = 1$ if worker k is home on day d and 0 otherwise. Let $G_d^{k,k'} = 1$ if worker k and k' met during day d and 0 otherwise. Let also $G_d^{k,k'} := 0$ for $k' = k$.

The goal is to design a Variational Inference algorithm to approximate the posterior probability over Z by using the VI distribution

$$q(Z) = \prod_{d,k} q(Z_d^k) \tag{11}$$

Solution

Starting of with the log-probability $\log p(A, \Omega \mid G)$ we can split it into two parts, the lower-bound and the KL divergence between the wanted posterior and $q(\mathbf{Z})$.

$$\log p(A, \Omega \mid G) = L(q) + D(q(\mathbf{Z}) \mid\mid p(\mathbf{Z} \mid A, \Omega, G)) \tag{12}$$

We know see that if we maximize the lower-bound the KL divergence between the wanted posterior and $q(\mathbf{Z})$ will be minimized, and we will get the best approximation (w.r.t. to our variational distribution) of the posterior over $\mathbf{Z}$.

Using the characteristics of mean field theory[2] we get that $q^*(Z_d^k)$ is given by

$$q^*(Z_d^k) \propto \exp\left(\mathbb{E}_{-Z_d^k}[\log p(A, \mathbf{Z}, \Omega \mid G)]\right) \tag{13}$$

where $-Z_d^k$ denotes over all $Z_{d'}^{k'}$ except $\{k' = k \cap d' = d\}$. The expectation can easily be split up, but first we need to define all probabilities.

Starting of with the priors of the parameters in Ω we know that, $\iota \sim \text{Beta}(\alpha_\iota, \beta_\iota)$, $\alpha \sim \text{Beta}(\alpha_\alpha, \beta_\alpha)$ and $\sigma \sim \text{Beta}(\alpha_\sigma, \beta_\sigma)$. That is

[1] Simplifying assumption permitted by the TAs
[2] Bishop, *Pattern Recognition and Machine Learning*, p. 466

$$p(\iota) = \frac{\iota^{\alpha_\iota - 1}(1-\iota)^{\beta_\iota - 1}}{B(\alpha_\iota, \beta_\iota)} \tag{14}$$

$$p(\alpha) = \frac{\alpha^{\alpha_\alpha - 1}(1-\alpha)^{\beta_\alpha - 1}}{B(\alpha_\alpha, \beta_\alpha)} \tag{15}$$

$$p(\sigma) = \frac{\sigma^{\alpha_\sigma - 1}(1-\sigma)^{\beta_\sigma - 1}}{B(\alpha_\sigma, \beta_\sigma)} \tag{16}$$

Now to the absence tables. A_d^k is only dependent on Z_d^k and σ, and it follows a Bernoulli distribution w.r.t. σ. We get that

$$p(A_d^k \mid Z_d^k \in \{1, \dots, 9\}, \sigma) = \sigma^{A_d^k}(1-\sigma)^{1-A_d^k} \tag{17}$$

For other values of Z_d^k the $A_d^k = 0$ with probability 1. Extending this we get

$$\begin{aligned} p(A \mid \mathbf{Z}, \sigma) &= \prod_{k,d} p(A_d^k \mid Z_d^k, \sigma) \\ &= \prod_{k,d} \left(\sigma^{A_d^k}(1-\sigma)^{1-A_d^k}\right)^{I\{Z_d^k \in \{1,\dots,9\}\}} \end{aligned} \tag{18}$$

Now to $\mathbf{Z}$. We start by noticing the markov property of Z_d^k given $\mathbf{Z}_{d-1}$. Namely, it is only dependent on $\mathbf{Z}_{d-1}$, G_{d-1} and Ω. With this we can write

$$p(\mathbf{Z} \mid \iota, \alpha, G) = \prod_k \prod_d p(Z_d^k \mid \mathbf{Z}_{d-1}, G_{d-1}, \iota, \alpha) \tag{19}$$

To make the notation simpler we define M_d^k. $M_d^k = 1$ if worker k was in contact with at least one infected individual during day d and 0 otherwise. That is $M_d^k := I\{\exists k' \mid G_d^{k,k'} = 1 \cap Z_d^{k'} \in \{1, \dots, 9\}\}$. Using the negation technique, that is $M_d^k = 1 - I\{\text{worker } k \text{ did not meet an infected worker during day } d\}$, and the disjoint properties of $q(\mathbf{Z})$ it can be further rewritten as:

$$M_d^k = 1 - \prod_{i : G_d^{i,k} = 1} \left(I\{Z_d^i = 0\} + I\{Z_d^i = 10\}\right) \tag{20}$$

This expression will be used later. Now, using M_d^k, we can write the probability (where impossible cases have been omitted, that is cases with zero probability) as

$$p(Z_d^k \mid Z_{d-1}, G_{d-1}, \iota, \alpha) = \begin{cases} (1-\iota)^{M_{d-1}^k}(1-\alpha)^{I\{Z_{d-1}^k = 9\}}, & Z_d^k = 0 \\ \iota^{M_{d-1}^k}, & Z_d^k = 1 \\ I\{Z_{d-1}^k = i-1\}, & Z_d^k = i, i \in \{2, \dots, 9\} \\ \alpha^{I\{Z_{d-1}^k = 9\}}, & Z_d^k = 10 \end{cases} \tag{21}$$

This pragmatic approach can be justified by assuming that the combination of values for the random variables that has probability zero under p also has probability zero under our VI distribution q. A more thorough justification can be given by studying the KL of $q(\mathbf{Z})$ and the wanted posterior. If $q(\mathbf{Z}) = 0$ for any positive value of the true posterior the KL will diverge towards infinity.

Notice that if a worker has antibodies they will always have antibodies the next day (probability 1), so that case is redundant. Using even more indicator functions we extend Eq (19) like follows

$$\begin{aligned} p(\mathbf{Z} \mid \iota, \alpha, G) = \prod_k \prod_{d=2}^{D} \Bigg[&\left((1-\iota)^{M_{d-1}^k} (1-\alpha)^{I\{Z_{d-1}^k = 9\}} \right)^{I\{Z_d^k = 0\}} \\ &\cdot \left(\iota^{M_{d-1}^k} \right)^{I\{Z_d^k = 1\}} \\ &\cdot \prod_{i=2}^{9} I\{Z_{d-1}^k = i-1\}^{I\{Z_d^k = i\}} \\ &\cdot \left(\alpha^{I\{Z_{d-1}^k = 9\}} \right)^{I\{Z_d^k = 10\}} \Bigg] \end{aligned} \tag{22}$$

where we can start on day 2 since Z_1 is known. We now see that the third factor inside the outer product always is one since the indicator functions are all equal to 1 or all equal to 0. We can thus remove it and get the somewhat simpler expression

$$p(\mathbf{Z} \mid \iota, \alpha, G) = \prod_k \prod_{d=2}^{D} \left[\left((1-\iota)^{M_{d-1}^k} (1-\alpha)^{I\{Z_{d-1}^k = 9\}} \right)^{I\{Z_d^k = 0\}} \left(\iota^{M_{d-1}^k} \right)^{I\{Z_d^k = 1\}} \left(\alpha^{I\{Z_{d-1}^k = 9\}} \right)^{I\{Z_d^k = 10\}} \right] \tag{23}$$

We are now almost ready to compute Eq (13). First we need to re-write $p(A, \mathbf{Z}, \Omega \mid G)$ using Bayes rule.

$$\begin{aligned} p(\mathbf{Z}, A, \Omega \mid G) =& p(\mathbf{Z}, A \mid \Omega, G) p(\Omega \mid G) \\ =& p(A \mid \mathbf{Z}, \Omega, G) p(\mathbf{Z} \mid \Omega, G) p(\Omega \mid G) \\ =& p(A \mid \mathbf{Z}, \Omega) p(\mathbf{Z} \mid \Omega, G) p(\Omega) \end{aligned} \tag{24}$$

Using this form of the joint probability we extend the expectation in Eq (13) like follows

$$\mathbb{E}_{-Z_d^k}[\log p(A, \mathbf{Z}, \Omega \mid G)] = \mathbb{E}_{-Z_d^k}[\log p(A \mid \mathbf{Z}, \Omega)] + \mathbb{E}_{-Z_d^k}[\log p(\mathbf{Z} \mid \Omega, G)] + \mathbb{E}_{-Z_d^k}[\log p(\Omega)] \tag{25}$$

Since $q^*(Z_d^k)$ is a function of Z_d^k we are only interested in the terms containing it (since constants appearing everywhere will be "normalized out"), which is the first two terms in Eq (25). These can be further expanded into terms only containing Z_d^k.

$$\begin{aligned} \log p(A \mid \mathbf{Z}, \Omega) = &\sum_{k' \neq k} \sum_{d' \neq d} I\{Z_{d'}^{k'} \in \{1, \dots, 9\}\} \big(A_{d'}^{k'} \log \sigma (1 - A_{d'}^{k'}) \log(1-\sigma) \big) \\ &+ I\{Z_d^k \in \{1, \dots, 9\}\} \big(A_d^k \log \sigma (1 - A_d^k) \log(1-\sigma) \big) \end{aligned} \tag{26}$$

where we see that the second term is the only one containing Z_d^k. Moving on,

$$\begin{aligned} \log p(\mathbf{Z} \mid \Omega, G) = &\sum_{k' \neq k} \sum_{d' \notin \{d, d+1\}} \log p(Z_{d'}^{k'} \mid Z_{d'-1}, \Omega, G) \\ &+ \log p(Z_d^k \mid Z_{d-1}, \Omega, G) + \sum_{k'} \log p(Z_{d+1}^{k'} \mid Z_d^{-k}, Z_d^k, \Omega, G) \end{aligned} \tag{27}$$

where we see that the only the two last terms contain Z_d^k. We are now ready to take the expectation.

$$\begin{aligned} \mathbb{E}_{-Z_d^k}[\log p(A, \mathbf{Z}, \Omega \mid G)] \stackrel{+}{=} &\mathbb{E}_{-Z_d^k} \Big[I\{Z_d^k \in \{1, \dots, 9\}\} \big(A_d^k \log \sigma + (1 - A_d^k) \log(1-\sigma) \big) \Big] \\ &+ \mathbb{E}_{-Z_d^k} \Big[\log p(Z_d^k \mid Z_{d-1}, \Omega, G) + \sum_{k'} \log p(Z_{d+1}^{k'} \mid Z_d^{-k}, Z_d^k, \Omega, G) \Big] \end{aligned} \tag{28}$$

The first expectation is clearly not dependent on $-Z_d^k$ and can be taken out from the expectation. The same is not true for the second term, which we will split into two. Taking the first part

$$\begin{aligned}\mathbb{E}_{-Z_d^k}\Big[\log p(Z_d^k \mid Z_{d-1}, \Omega, G)\Big] =& \mathbb{E}_{-Z_d^k}\Bigg[I\{Z_d^k = 0\}\Big((M_{d-1}^k)\log(1-\iota) + I\{Z_{d-1}^k = 9\}\log(1-\alpha)\Big) \\ &+ M_{d-1}^k I\{Z_d^k = 1\}\log\iota + I\{Z_d^k = 10\}I\{Z_{d-1}^k = 9\}\log\alpha\Bigg] \\ =& I\{Z_d^k = 0\}\Big(\mathbb{E}_{-Z_d^k}[M_{d-1}^k]\log(1-\iota) + \mathbb{E}_{-Z_d^k}[I\{Z_{d-1}^k = 9\}]\log(1-\alpha)\Big) \\ &+ I\{Z_d^k = 1\}\,\mathbb{E}_{-Z_d^k}[M_{d-1}^k]\log\iota + I\{Z_d^k = 10\}\log\alpha\,\mathbb{E}_{-Z_d^k}[I\{Z_{d-1}^k = 9\}]\end{aligned} \tag{29}$$

The expectations in the expression are, due to $q(\mathbf{Z})$ being separable

$$\begin{aligned}\mathbb{E}_{-Z_d^k}[I\{Z_{d-1}^k = 9\}] &= q(Z_{d-1}^k = 9) \\ \mathbb{E}_{-Z_d^k}[M_{d-1}^k] &= 1 - \mathbb{E}_{-Z_d^k}\Bigg[\prod_{i:G_{d-1}^{i,k}=1}\Big(I\{Z_{d-1}^i = 0\} + I\{Z_{d-1}^i = 10\}\Big)\Bigg] \\ &= 1 - \prod_{i:G_{d-1}^{i,k}=1}\Big(q(Z_{d-1}^i = 0) + q(Z_{d-1}^i = 10)\Big)\end{aligned} \tag{30}$$

Now to the second part. It has a sum in the expectation and they consist of two classes, $k' = k$ and $k' \neq k$. For the first case we get

$$\begin{aligned}&\mathbb{E}_{-Z_d^k}[\log p(Z_{d+1}^k \mid Z_d^{-k}, Z_d^k, \Omega, G)] = \mathbb{E}_{-Z_d^k}\Bigg[I\{Z_{d+1}^k = 0\}\Big((M_d^k)\log(1-\iota) + I\{Z_d^k = 9\}\log(1-\alpha)\Big) \\ &+ M_d^k I\{Z_{d+1}^k = 1\}\log\iota + I\{Z_{d+1}^k = 10\}I\{Z_d^k = 9\}\log\alpha\Bigg] \\ &\overset{+}{=} \mathbb{E}_{-Z_d^k}\Big[I\{Z_{d+1}^k = 0\}I\{Z_d^k = 9\}\log(1-\alpha) + I\{Z_{d+1}^k = 10\}I\{Z_d^k = 9\}\log\alpha\Big] \\ &= I\{Z_d^k = 9\}\Big(q(Z_{d+1}^k = 0)\log(1-\alpha) + q(Z_{d+1}^k = 10)\log\alpha\Big)\end{aligned} \tag{31}$$

The parts containing M_d^k could be removed since M_d^k only contains information about the workers worker w_k met on day d, not worker w_k herself.

In a similar fashion for the second case, however two terms disappears since it's no longer contains Z_d^k. The parts containing $M_d^{k'}$ now remains since information about worker w_k is in them.

$$\mathbb{E}_{-Z_d^k}[\log p(Z_{d+1}^{k'} \mid Z_d, \Omega, G)] \stackrel{\pm}{=} \mathbb{E}_{-Z_d^k}\left[I\{Z_{d+1}^{k'} = 0\} M_d^{k'} \log(1-\iota) + M_d^{k'} I\{Z_{d+1}^{k'} = 1\} \log \iota \right]$$

$$= \Big(q(Z_{d+1}^k = 0) \log(1-\iota) + q(Z_{d+1}^k = 1) \log \iota \Big) \, \mathbb{E}_{-Z_d^k}\left[M_d^{k'} \right]$$

$$= \Big(q(Z_{d+1}^k = 0) \log(1-\iota) + q(Z_{d+1}^k = 1) \log \iota \Big) \Big(1 - (I\{Z_d^k = 0\} + I\{Z_d^k = 10\})$$
$$\cdot \prod_{i: G_d^{i,k'} = 1, i \neq k} \big(q(Z_d^i = 0) + q(Z_d^i = 10) \big) \Big)$$

$$\stackrel{\pm}{=} - \Big(q(Z_{d+1}^k = 0) \log(1-\iota) + q(Z_{d+1}^k = 1) \log \iota \Big) \big(I\{Z_d^k = 0\} + I\{Z_d^k = 10\} \big) \prod_{i: G_d^{i,k'} = 1, i \neq k} \big(q(Z_d^i = 0) + q(Z_d^i = 10) \big) \tag{32}$$

We have finally calculated all the necessary parts, and we can now start allocating the terms to the different distributions. But first we notice that in no expression are there any function of Z_d^k (except of course the indicator functions, but these only take values 0 and 1) which indicates that $q(Z_d^k)$ must be a categorical distribution. This could also be understood intuitively, since the sample space is disjoint, discrete and sum to 1 for each k and d. We denote this categorical distribution $\pi^{k,d} \in \mathbb{R}^{11}$ (there are eleven cases).

Thankfully, using this notation, $q(Z_d^k = i) = \pi_i^{k,d}$, we can simplify the expressions a lot. Eq (29) becomes

$$\begin{aligned} &\mathbb{E}_{-Z_d^k}\Big[\log p(Z_d^k \mid Z_{d-1}, \Omega, G) \Big] \\ &= I\{Z_d^k = 0\} \Big[\Big(1 - \prod_{i: G_{d-1}^{i,k} = 1} \big(\pi_0^{i,d-1} + \pi_{10}^{i,d-1} \big) \Big) \log(1-\iota) + \pi_9^{k,d-1} \log(1-\alpha) \Big] \\ &+ I\{Z_d^k = 1\} \Big(1 - \prod_{i: G_{d-1}^{i,k} = 1} \big(\pi_0^{i,d-1} + \pi_{10}^{i,d-1} \big) \Big) \log \iota + I\{Z_d^k = 10\} \pi_9^{k,d-1} \log \alpha \end{aligned} \tag{33}$$

Eq (31) becomes

$$\mathbb{E}_{-Z_d^k}[\log p(Z_{d+1}^k \mid Z_d^{-k}, Z_d^k, \Omega, G)] = I\{Z_d^k = 9\} \Big(\pi_0^{k,d+1} \log(1-\alpha) + \pi_{10}^{k,d+1} \log \alpha \Big) \tag{34}$$

and Eq (32) becomes

$$\begin{aligned} &\mathbb{E}_{-Z_d^k}[\log p(Z_{d+1}^{k'} \mid Z_d, \Omega, G)] \\ &= - \Big(\pi_0^{k',d+1} \log(1-\iota) + \pi_1^{k',d+1} \log \iota \Big) \big(I\{Z_d^k = 0\} + I\{Z_d^k = 10\} \big) \prod_{i: G_d^{i,k'} = 1, i \neq k} \big(\pi_0^{i,d} + \pi_{10}^{i,d} \big) \end{aligned} \tag{35}$$

Putting all these into Eq (28) we get the following massive expression

$$
\begin{aligned}
\mathbb{E}_{-Z_d^k}[\log p(A, \mathbf{Z}, \Omega \mid G)] \overset{\pm}{=} & \mathbb{E}_{-Z_d^k}\Big[I\{Z_d^k \in \{1,\dots,9\}\}\big(A_d^k \log\sigma + (1-A_d^k)\log(1-\sigma)\big)\Big] \\
& + \mathbb{E}_{-Z_d^k}\Big[\log p(Z_d^k \mid Z_{d-1}, \Omega, G) + \sum_{k'} \log p(Z_{d+1}^{k'} \mid Z_d^{-k}, Z_d^k, \Omega, G)\Big] \\
& = I\{Z_d^k \in \{1,\dots,9\}\}\big(A_d^k \log\sigma + (1-A_d^k)\log(1-\sigma)\big) \\
& + I\{Z_d^k = 0\}\Big[\Big(1 - \prod_{i:G_{d-1}^{i,k}=1} \big(\pi_0^{i,d-1} + \pi_{10}^{i,d-1}\big)\Big)\log(1-\iota) + \pi_9^{k,d-1}\log(1-\alpha)\Big] \\
& + I\{Z_d^k = 1\}\Big(1 - \prod_{i:G_{d-1}^{i,k}=1} \big(\pi_0^{i,d-1} + \pi_{10}^{i,d-1}\big)\Big)\log\iota + I\{Z_d^k = 10\}\pi_9^{k,d-1}\log\alpha \\
& + I\{Z_d^k = 9\}\Big(\pi_0^{k,d+1}\log(1-\alpha) + \pi_{10}^{k,d+1}\log\alpha\Big) \\
& - \big(I\{Z_d^k = 0\} + I\{Z_d^k = 10\}\big)\sum_{k'\neq k}\Big[\Big(\pi_0^{k',d+1}\log(1-\iota) + \pi_1^{k',d+1}\log\iota\Big) \\
& \cdot \prod_{i:G_d^{i,k'}=1, i\neq k} \big(\pi_0^{i,d} + \pi_{10}^{i,d}\big)\Big]
\end{aligned}
\tag{36}
$$

We can now use the indicator functions to divide these up. For a non-infected worker:

$$
\begin{aligned}
\log q^*(Z_d^k = 0) = & \Big(1 - \prod_{i:G_{d-1}^{i,k}=1} \big(\pi_0^{i,d-1} + \pi_{10}^{i,d-1}\big)\Big)\log(1-\iota) + \pi_9^{k,d-1}\log(1-\alpha) \\
& - \sum_{k'\neq k}\Big[\Big(\pi_0^{k',d+1}\log(1-\iota) + \pi_1^{k',d+1}\log\iota\Big)\prod_{i:G_d^{i,k'}=1, i\neq k} \big(\pi_0^{i,d} + \pi_{10}^{i,d}\big)\Big]
\end{aligned}
\tag{37}
$$

For a worker that was infected the day before:

$$
\log q^*(Z_d^k = 1) = \Big(1 - \prod_{i:G_{d-1}^{i,k}=1} \big(\pi_0^{i,d-1} + \pi_{10}^{i,d-1}\big)\Big)\log\iota + A_d^k\log\sigma + (1-A_d^k)\log(1-\sigma) \tag{38}
$$

For a worker that is on their i:th infected day, $i = 2,\dots,8$:

$$
\log q^*(Z_d^k = i) = A_d^k\log\sigma + (1-A_d^k)\log(1-\sigma) \tag{39}
$$

For a worker that is on their last day of infection:

$$
\log q^*(Z_d^k = 9) = A_d^k\log\sigma + (1-A_d^k)\log(1-\sigma) + \pi_0^{k,d+1}\log(1-\alpha) + \pi_{10}^{k,d+1}\log\alpha \tag{40}
$$

And finally, for a worker with antibodies

$$
\begin{aligned}
\log q^*(Z_d^k = 10) = & \pi_9^{k,d-1}\log\alpha \\
& - \sum_{k'\neq k}\Big[\Big(\pi_0^{k',d+1}\log(1-\iota) + \pi_1^{k',d+1}\log\iota\Big)\prod_{i:G_d^{i,k'}=1, i\neq k} \big(\pi_0^{i,d} + \pi_{10}^{i,d}\big)\Big]
\end{aligned}
\tag{41}
$$

To make these updates equations more tangible, we will take the expectation of ι, α and σ. Since these all have beta priors we know that (proof can be found in the Appendix)

$$\mathbb{E}[\log \iota] = \psi(\alpha_\iota) - \psi(\alpha_\iota + \beta_\iota) \tag{42}$$

In a similar fashion we can show that

$$\mathbb{E}[\log(1-\iota)] = \psi(\beta_\iota) - \psi(\alpha_\iota + \beta_\iota) \tag{43}$$

The expectations for the other parameters are similar but with their respective prior parameters. Taking the exponential of the log-update equations and putting in these expectations we get for a non-infected worker:

$$\begin{aligned} q^*(Z_d^k = 0) \propto \exp\Bigg\{ & \Big(1 - \prod_{i:G_{d-1}^{i,k}=1} \big(\pi_0^{i,d-1} + \pi_{10}^{i,d-1}\big)\Big)\big(\psi(\beta_\iota) - \psi(\alpha_\iota + \beta_\iota)\big) + \pi_9^{k,d-1}\big(\psi(\beta_\alpha) - \psi(\alpha_\alpha + \beta_\alpha)\big) \\ & - \sum_{k' \neq k} \Bigg[\Big(\pi_0^{k',d+1}\big(\psi(\beta_\iota) - \psi(\alpha_\iota + \beta_\iota)\big) + \pi_1^{k',d+1}\big(\psi(\alpha_\iota) - \psi(\alpha_\iota + \beta_\iota)\big)\Big) \prod_{i:G_d^{i,k'}=1, i \neq k} \big(\pi_0^{i,d} + \pi_{10}^{i,d}\big) \Bigg] \Bigg\} \end{aligned} \tag{44}$$

For a worker that was infected the day before:

$$\begin{aligned} q^*(Z_d^k = 1) \propto \exp\Bigg\{ & \Big(1 - \prod_{i:G_{d-1}^{i,k}=1} \big(\pi_0^{i,d-1} + \pi_{10}^{i,d-1}\big)\Big)\big(\psi(\alpha_\iota) - \psi(\alpha_\iota + \beta_\iota)\big) \\ & + A_d^k\big(\psi(\alpha_\sigma) - \psi(\alpha_\sigma + \beta_\sigma)\big) + (1 - A_d^k)\big(\psi(\beta_\sigma) - \psi(\alpha_\sigma + \beta_\sigma)\big) \Bigg\} \end{aligned} \tag{45}$$

For a worker that is on their i:th infected day, $i = 2, \dots, 8$:

$$q^*(Z_d^k = i) \propto \exp\Big\{ A_d^k\big(\psi(\alpha_\sigma) - \psi(\alpha_\sigma + \beta_\sigma)\big) + (1 - A_d^k)\big(\psi(\beta_\sigma) - \psi(\alpha_\sigma + \beta_\sigma)\big) \Big\} \tag{46}$$

For a worker that is on their last day of infection:

$$\begin{aligned} q^*(Z_d^k = 9) \propto \exp\Big\{ & A_d^k\big(\psi(\alpha_\sigma) - \psi(\alpha_\sigma + \beta_\sigma)\big) + (1 - A_d^k)\big(\psi(\beta_\sigma) - \psi(\alpha_\sigma + \beta_\sigma)\big) \\ & + \pi_0^{k,d+1}\big(\psi(\beta_\alpha) - \psi(\alpha_\alpha + \beta_\alpha)\big) + \pi_{10}^{k,d+1}\big(\psi(\alpha_\alpha) - \psi(\alpha_\alpha + \beta_\alpha)\big) \Big\} \end{aligned} \tag{47}$$

And finally, for a worker with antibodies

$$\begin{aligned} q^*(Z_d^k = 10) \propto \exp\Big\{ & \pi_9^{k,d-1}\big(\psi(\alpha_\alpha) - \psi(\alpha_\alpha + \beta_\alpha)\big) \\ & - \sum_{k' \neq k} \Bigg[\Big(\pi_0^{k',d+1}\big(\psi(\beta_\iota) - \psi(\alpha_\iota + \beta_\iota)\big) + \pi_1^{k',d+1}\big(\psi(\alpha_\iota) - \psi(\alpha_\iota + \beta_\iota)\big)\Big) \prod_{i:G_d^{i,k'}=1, i \neq k} \big(\pi_0^{i,d} + \pi_{10}^{i,d}\big) \Bigg] \Big\} \end{aligned} \tag{48}$$

And to get the final expressions we just normalize them. Say $\tilde{q}^*(Z_d^k = i)$ are the expressions above (with $=$ instead of $\propto$), then the final update equations are

$$q^*(Z_d^k = i) = \frac{\tilde{q}^*(Z_d^k = i)}{\sum_{j=0}^{10} \tilde{q}^*(Z_d^k = j)} = (\pi_i^{k,d})^* \tag{49}$$

Discussion and Initialization

We can now iterate between the workers and the days and calculate $(\pi_i^{k,d})^*$ for $k = 1, \ldots, K$ and $d = 2, \ldots, D-1$. We start on the second day since we know $\mathbf{Z}_1$ (and since we don't consider day $d = 0$, which the first day would consider in its calculations), worker w_1 just got infected and all other workers are non-infected. Furthermore, we do not have to update the probabilities for worker w_1 during the first 9 nine days since we know that she will be infected during that time. But why stop on day $D-1$? The algorithm uses the probabilities for day $d+1$, and that day doesn't "exist" for day $d = D$. We thus set $\pi_{10}^{k,D} = 1$ for all k, which is deemed reasonable if D is large enough, since sooner or later all workers will have antibodies.

But how should the algorithm be initialized? Using the same argument as in the previous graph we can put $\pi_1^{1,1} = 1$ and $\pi_0^{j,1} = 1$ for $j \neq 1$. The only other thing we know for sure is that $\pi_d^{1,d} = 1$ for days $d \in \{2, \ldots, 9\}$. We can initialize all other $\pi_j^{k,d}$'s as $\frac{1}{11}$, that is an equal probability of being in every respective case.

5 3.5 Hard VI for Covid-19

Not attempted.

6 3.6 Spectral Graph Analysis

Definitions

Let $G = (V, E)$ be an undirected d-regular graph, let $\mathbf{A}$ bet the adjency matrix of G, and let $\mathbf{L}$ be the normalized Laplacian of G where $\mathbf{L} = \mathbf{I} - \frac{1}{d}\mathbf{A}$. We thus have that
$\deg(v_i) = d \ \forall v_i \in V$.

Note that the order of the proofs have been changed in order to not have to define variables multiple times.

Proof 2

Let $\mathbf{B}$ be the incidence matrix of G. It is defined as follows: its rows are labelled by the edges of G, its columns laballed by the vertices of G, and the entry (i, j) position is 1 if vertex j is incident with edge i, and 0 if not. Thus, every row of $\mathbf{B}$ has exactly two 1s, and every column j has as many 1s as the degree of vertex j. We now change $\mathbf{B}$ in to and oriented incidence matrix, $\mathbf{B}_o$, by changing the first 1 of every row to -1.

We now create the square matrix $\mathbf{B}_o^T\mathbf{B}_o$, which has rows and columns labelled by the vertices $v_i \in V$. The entry in the (i, j) position is now the scalar product of columns i and j of $\mathbf{B}_o$. If $i = j$ this is $deg(v_i) = d$ (a sum of d 1s) and 0 or -1 if $i \neq j$. The entry takes the value -1 if there is a row where both column i and column j have nonzero entries, which only happens for one row, when $(v_i, v_j) \in E$. We thus get:

$$(\mathbf{B}_o^T\mathbf{B}_o)_{ij} = \begin{cases} d, & \text{if } i = j \\ -1, & \text{if } (v_i, v_j) \in E \\ 0, & \text{if } (v_i, v_j) \notin E \end{cases}$$

We now see that

$$\mathbf{B}_o^T\mathbf{B}_o = d\mathbf{I} - \mathbf{A} = d(\mathbf{I} - \frac{1}{d}\mathbf{A}) = d\mathbf{L}$$

This means that $d\mathbf{L}$ is positive semidefinite. This is due to for any $\mathbf{x} \in \mathbb{R}^{|V|}$ we have that

$$\mathbf{x}^T(d\mathbf{L})\mathbf{x} = \mathbf{x}^T(\mathbf{B}_o^T\mathbf{B}_o)\mathbf{x} = (\mathbf{B}_o\mathbf{x})^T\mathbf{B}_o\mathbf{x} \geq 0$$

And since $d \geq 1$ this also implies that the normalized Laplacian is positive semidefinite. ■

Proof 1

Using the oriented incidence matrix $\mathbf{B}_o$ that was defined in the previous proof we recall that $\mathbf{L} = \frac{1}{d}\mathbf{B}_o^T\mathbf{B}_o$. Now,

$$\mathbf{x}^T\mathbf{L}\mathbf{x} = \frac{1}{d}\mathbf{x}^T\mathbf{B}_o^T\mathbf{B}_o\mathbf{x} = \frac{1}{d}(\mathbf{B}_o\mathbf{x})^T(\mathbf{B}_o\mathbf{x})$$

We now have that the entries of $\mathbf{B}_o\mathbf{x}$ are labelled by the edges of G, since the rows of $\mathbf{B}_o$ are. Now, if $e = (u, v)$ is an edge of G, then the entry of $\mathbf{B}_o\mathbf{x}$ in the position corresponding to e is either $(x_u - x_v)$ or $-(x_u - x_v)$, due to the characteristics of $\mathbf{B}_o$ defined in the previous proof. That is the nodes connected by an edge having either a value of 1 or -1. Thus, we get

$$\mathbf{x}^T\mathbf{L}\mathbf{x} = \frac{1}{d}(\mathbf{B}_o\mathbf{x})^T(\mathbf{B}_o\mathbf{x}) = \frac{1}{d}\sum_{(u,v)\in\mathrm{E(G)}}(\pm(x_u - x_v))^2 = \frac{1}{d}\sum_{(u,v)\in\mathrm{E(G)}}(x_u - x_v)^2 \tag{50}$$

■

Proof 3

Since we have shown that $\mathbf{L}$ is positive semidefinite, we know by definition that $\mathbf{x}^T\mathbf{L}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^{|V|}$. This means that the smallest value equation Eq (50) can take is 0. This can easily be achieved by choosing $\mathbf{x}$ to be a constant vector ($c \cdot \mathbf{1}$ where $c \in \mathbb{R}$). This solution is called the trivial solution. All non constant vectors that minimizes Eq (50) are called non-trivial solutions.

The coordinate x_u^* can be seen as the value assigned to vertex u. We can thus see $\mathbf{x}^*$ as a one-dimensional embedding of the graph G which assigns a value to every vertex $v \in V$. Defining $\mathbf{x}^*$ as

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \not\propto \mathbf{1}} \left(\mathbf{x}^T\mathbf{L}\mathbf{x}\right) = \arg\min_{\mathbf{x} \not\propto \mathbf{1}} \left(\frac{1}{d} \sum_{(u,v)\in \mathrm{E(G)}} (x_u - x_v)^2\right) \tag{51}$$

one sees that the summation can be seen as a potential function that is to be minimized. In fact, one sees that minimizing the summation make the values of connected vertices close to each other. This is a meaningful embedding in the sense that related vertices are kept close in the embedded space. This is illustrated by figure 1.

∎

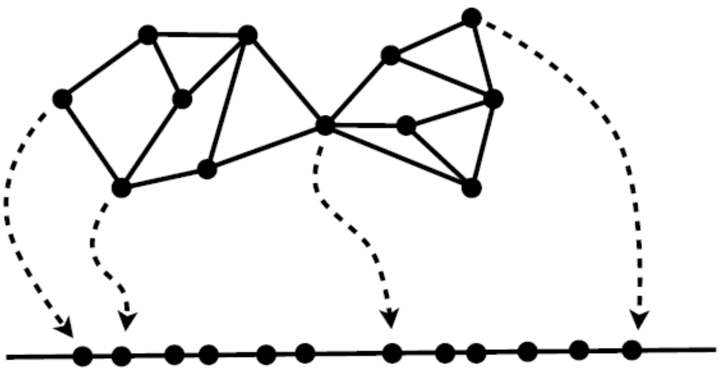

Figure 1: A picture showing how a graph can be embedded into the real line.
Taken from Aristides Gionis' slides for module 4

A Proof of expectation of log of beta distribution

Expectation of $\log X$

Let X be $Beta(\alpha, \beta)$ distributed. Then

$$
\begin{aligned}
\mathbb{E}[\log X] &= \int_0^1 \log x \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} dx \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 \frac{\partial (x^{\alpha-1}(1-x)^{\beta-1})}{\partial \alpha} dx \\
&= \frac{1}{B(\alpha,\beta)} \frac{\partial}{\partial \alpha} \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx \\
&= \frac{1}{B(\alpha,\beta)} \frac{\partial B(\alpha,\beta)}{\partial \alpha} \\
&= \frac{\partial \log B(\alpha,\beta)}{\partial \alpha} \\
&= \frac{\partial \log \Gamma(\alpha)}{\partial \alpha} - \frac{\partial \log \Gamma(\alpha+\beta)}{\partial \alpha} \\
&= \psi(\alpha) - \psi(\alpha+\beta)
\end{aligned}
\tag{52}
$$

Expectation of $\log(1-X)$

In a similar fashion we can derive $\mathbb{E}[\log(1-X)]$ by changing the integration variable, $x = 1-u$

$$
\begin{aligned}
\mathbb{E}[\log(1-X)] &= \int_0^1 \log(1-x) \frac{x^{\alpha-1}(x-1)^{\beta-1}}{B(\alpha,\beta)} dx \\
&= \int_0^1 \log u \frac{(1-u)^{\alpha-1} u^{\beta-1}}{B(\alpha,\beta)} du \\
&= \cdots = \psi(\beta) - \psi(\alpha+\beta)
\end{aligned}
\tag{53}
$$

B Previous version - 3.1 Easier EM for Advertisements

Let $\mathbf{X} = \{X_{n,l,m} \mid n \in [N], l \in [L], m \in [M]\}$ be if a reader has read a certain advertisement or not. More specifically,

$$X_{n,l,m} = \begin{cases} 1, & \text{if reader } n \text{ has read advertisement } m \text{ in edition } l \\ 0, & \text{if reader } n \text{ has } \textbf{not} \text{ read advertisement } m \text{ in edition } l \end{cases}$$

The readers, editions and advertisement can be grouped into C, D and F groups respectively. These class variables have categorical distribution $(\theta^r, \theta^a, \theta^e)$, which can be expressed has follows:

$$p(Z_n^r) = \prod_{c=1}^{C} (\theta_c^r)^{I\{Z_n^r = c\}}$$

$$p(Z_l^e) = \prod_{d=1}^{D} (\theta_d^r)^{I\{Z_l^e = d\}}$$

$$p(Z_m^a) = \prod_{f=1}^{F} (\theta_f^a)^{I\{Z_m^a = f\}}$$

where I is the indicator function.

$\mathbf{Z}$ will be used to express these latent variables, where

$$\mathbf{Z} = \{(Z_n^r, Z_l^e, Z_m^a) \mid n \in [N], l \in [L], m \in [M]\}$$

The probability of a reader clicking on an advertisement is thought to be Bernoulli distributed and only dependent on the latent class variables. The distribution can be expressed as follows:

$$p(X_{n,l,m} \mid Z_n^r = c, Z_l^e = d, Z_m^a = f) = \psi_{c,d,f}^{X_{n,l,m}} (1 - \psi_{c,d,f})^{1 - X_{n,l,m}}$$

The model parameters will be kept as follows

$$\Omega = (\theta^r, \theta^a, \theta^e, \{\psi_{c,d,f} : c \in [C], d \in [D], f \in [F]\})$$

In the **E-step** $p(\mathbf{Z} \mid \mathbf{X}, \Omega^{old})$ should be evaluated. In the following expressions the superscript ^{old} for all parameters in Ω has been omitted for clarity's sake. Using Bayes' rule we get

$$p(\mathbf{Z} \mid \mathbf{X}, \Omega) = \frac{p(\mathbf{Z}, \mathbf{X} \mid \Omega)}{p(\mathbf{X} \mid \Omega)} = \frac{p(\mathbf{X} \mid \mathbf{Z}, \Omega) p(\mathbf{Z} \mid \Omega)}{\sum_{\mathbf{Z}} p(\mathbf{X} \mid \mathbf{Z}, \Omega) p(\mathbf{Z} \mid \Omega)} \tag{54}$$

$$p(\mathbf{X} \mid \mathbf{Z}, \Omega) p(\mathbf{Z} \mid \Omega) = \prod_{n=1}^{N} \prod_{l=1}^{L} \prod_{m \in A(l)} \prod_{c=1}^{C} \prod_{d=1}^{D} \prod_{f=1}^{F} (\theta_c^r \theta_d^e \theta_f^a \psi_{c,d,f}^{X_{n,l,m}} (1 - \psi_{c,d,f})^{1 - X_{n,l,m}})^{I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\}}$$

We now go over to the log-space to simplify the expressions (and avoid numerical underflow if this algorithm would have been implemented).

$$\begin{aligned} \log(p(\mathbf{X} \mid \mathbf{Z}, \Omega) p(\mathbf{Z} \mid \Omega)) = & \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{f=1}^{F} I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\} \cdot \\ & \cdot \Big[\log \theta_c^r + \log \theta_d^e + \log \theta_f^a + X_{n,l,m} \log \psi_{c,d,f} + (1 - X_{n,l,m}) \log(1 - \psi_{c,d,f}) \Big] \end{aligned}$$

With this, we can now calculate $p(\mathbf{Z} \mid \mathbf{X}, \Omega^{old})$ using equation (54). Continuing with the **Q-step** we now want to calculate

$$\Omega^{new} = \arg\max_{\Omega} Q(\Omega, \Omega^{old})$$

where

$$Q(\Omega, \Omega^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \Omega^{old}) \log p(\mathbf{X}, \mathbf{Z} \mid \Omega) \tag{55}$$

We now re-write equation (55) to get

$$\begin{aligned}
Q(\Omega, \Omega^{old}) =& \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \Omega^{old}) \log p(\mathbf{X}, \mathbf{Z} \mid \Omega) \\
=& \sum_{\mathbf{Z}} \mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[\log p(\mathbf{X}, \mathbf{Z} \mid \Omega)] \\
=& \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{f=1}^{F} \mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\}] \\
& \Big(\log \theta_c^r + \log \theta_d^e + \log \theta_f^a + X_{n,l,m} \log \psi_{c,d,f} + (1 - X_{n,l,m}) \log(1 - \psi_{c,d,f})\Big)
\end{aligned} \tag{56}$$

Now, we will re-write $\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\}]$.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\}] \\
&= \sum_{c'=1}^{C} \sum_{d'=1}^{D} \sum_{f'=1}^{F} I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\} p(Z_n^r = c', Z_l^e = d', Z_m^a = f' \mid X_{n,l,m}, \Omega^{old}) \\
&= p(Z_n^r = c, Z_l^e = d, Z_m^a = f \mid X_{n,l,m}, \Omega^{old}) \\
&= \frac{p(Z_n^r = c, Z_l^e = d, Z_m^a = f, X_{n,l,m} \mid \Omega^{old})}{p(X_{n,l,m} \mid \Omega^{old})} \\
&= \frac{p(X_{n,l,m} \mid Z_n^r = c, Z_l^e = d, Z_m^a = f, \Omega^{old}) p(Z_n^r = c, Z_l^e = d, Z_m^a = f \mid \Omega^{old})}{p(X_{n,l,m} \mid \Omega^{old})}
\end{aligned} \tag{57}$$

Putting the numerator as

$$\begin{aligned}
\phi_{c,d,f}^{n,l,m} :=& p(X_{n,l,m} \mid Z_n^r = c, Z_l^e = d, Z_m^a = f, \Omega^{old}) p(Z_n^r = c, Z_l^e = d, Z_m^a = f \mid \Omega^{old}) \\
=& (\psi^{old})_{c,d,f}^{X_{n,l,m}} (1 - (\psi^{old})_{c,d,f})^{1 - X_{n,l,m}} \theta_c^{r,\text{old}} \theta_d^{e,\text{old}} \theta_f^{a,\text{old}}
\end{aligned} \tag{58}$$

we can define the expected value as

$$\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\}] := \gamma_{c,d,f}^{n,l,m} = \frac{\phi_{c,d,f}^{n,l,m}}{\sum_{c',d',f'} \phi_{c',d',f'}^{n,l,m}}$$

Q can now be expressed as

$$\begin{aligned}
Q(\Omega, \Omega^{old}) =& \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{f=1}^{F} \gamma_{c,d,f}^{n,l,m} \\
& \Big(\log \theta_c^r + \log \theta_d^e + \log \theta_f^a + X_{n,l,m} \log \psi_{c,d,f} + (1 - X_{n,l,m}) \log(1 - \psi_{c,d,f})\Big)
\end{aligned} \tag{59}$$

which will be maximized in terms of Ω. This will be done one parameter at a time, which is sometimes referred to as generalized EM.

Starting with $(\theta^r, \theta^a, \theta^e)$ we know that they must satisfy $\sum_{c=1}^{C} \theta_c^r = 1$, $\sum_{d=1}^{D} \theta_d^e = 1$ and $\sum_{f=1}^{F} \theta_f^a = 1$ since they are probabilities.

Using Lagrange relaxation on the first θ we get the Lagrangian

$$\begin{aligned} L(\theta_c^r, \lambda) =& Q(\Omega, \Omega^{old}) - \lambda\Big(-1 + \sum_{c'=1}^{C} \theta_{c'}^r \Big) \\ \stackrel{+}{=}& \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{d=1}^{D} \sum_{f=1}^{F} \gamma_{c,d,f}^{n,l,m} \log \theta_c^r - \lambda\Big(-1 + \sum_{c'=1}^{C} \theta_{c'}^r \Big) \end{aligned} \tag{60}$$

Firstly, the partial derivative w.r.t. θ_c^r gives

$$\frac{\partial L}{\partial \theta_c^r} = \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{d=1}^{D} \sum_{f=1}^{F} \gamma_{c,d,f}^{n,l,m} \frac{1}{\theta_c^r} - \lambda = 0$$

$$\theta_c^r = \frac{\sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{d=1}^{D} \sum_{f=1}^{F} \gamma_{c,d,f}^{n,l,m}}{\lambda}$$

Secondly, the partial derivative w.r.t. λ gives

$$\frac{\partial L}{\partial \lambda} = 1 - \sum_{c'=1}^{C} \theta_{c'}^r = 0$$

$$\sum_{c'=1}^{C} \theta_{c'}^r = 1$$

Now,

$$\lambda = \lambda \cdot 1 = \lambda \sum_{c'=1}^{C} \theta_{c'}^r = \sum_{c'=1}^{C} \theta_{c'}^r \lambda = \sum_{c'=1}^{C} \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{d=1}^{D} \sum_{f=1}^{F} \gamma_{c',d,f}^{n,l,m}$$

And thus,

$$\theta_c^{r,new} = \frac{\sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{d=1}^{D} \sum_{f=1}^{F} \gamma_{c,d,f}^{n,l,m}}{\sum_{c'=1}^{C} \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{d=1}^{D} \sum_{f=1}^{F} \gamma_{c',d,f}^{n,l,m}} \tag{61}$$

Repeating these steps in the same manner for θ_d^e and θ_f^a gives

$$\theta_d^{e,new} = \frac{\sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{c=1}^{C} \sum_{f=1}^{F} \gamma_{c,d,f}^{n,l,m}}{\sum_{d'=1}^{D} \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{c=1}^{C} \sum_{f=1}^{F} \gamma_{c,d',f}^{n,l,m}} \tag{62}$$

and

$$\theta_f^{a,new} = \frac{\sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{c=1}^{C} \sum_{d=1}^{D} \gamma_{c,d,f}^{n,l,m}}{\sum_{f'=1}^{F} \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m \in A(l)} \sum_{c=1}^{C} \sum_{d=1}^{D} \gamma_{c,d,f'}^{n,l,m}} \tag{63}$$

We will now optimize w.r.t to $\psi_{c,d,f}$. Setting the first derivative to zero and solving we get

$$\frac{\partial Q}{\partial \psi_{c,d,f}} = \sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \gamma_{c,d,f}^{n,l,m}\Big(X_{n,l,m}\frac{1}{\psi_{c,d,f}} + (X_{n,l,m}-1)\frac{1}{1-\psi_{c,d,f}}\Big) = 0$$

$$\frac{1}{\psi_{c,d,f}}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \gamma_{c,d,f}^{n,l,m}X_{n,l,m} = \frac{1}{1-\psi_{c,d,f}}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \gamma_{c,d,f}^{n,l,m}(1-X_{n,l,m})$$

$$(1-\psi_{c,d,f})\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \gamma_{c,d,f}^{n,l,m}X_{n,l,m} = \psi_{c,d,f}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \gamma_{c,d,f}^{n,l,m}(1-X_{n,l,m})$$

$$\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \gamma_{c,d,f}^{n,l,m}X_{n,l,m} = \psi_{c,d,f}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \gamma_{c,d,f}^{n,l,m}$$

$$\psi_{c,d,f}^{new} = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \gamma_{c,d,f}^{n,l,m}X_{n,l,m}}{\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \gamma_{c,d,f}^{n,l,m}} \tag{64}$$

Now to the EM algorithm itself.

1. Initialize by choosing a suitable Ω^{old}, e.g.
   - $\theta_c^{r,old} = \frac{1}{C}$ for all c
   - $\theta_d^{e,old} = \frac{1}{D}$ for all d
   - $\theta_f^{a,old} = \frac{1}{F}$ for all f
   - $\psi_{c,d,f}^{old} = 0.1$ for all c, d, f.
2. Evaluate Ω^{new} by calculating
   - $\theta_c^{r,new}$ according to equation (61)
   - $\theta_d^{e,new}$ according to equation (62)
   - $\theta_f^{a,new}$ according to equation (63)
   - $\psi_{c,d,f}^{new}$ according to equation (64).
3. Check if the parameters of Ω have converged. If not, let

$$\Omega^{old} \leftarrow \Omega^{new}$$

   and return to step 2.

When the algorithm has converged the MLE of the model parameters Ω have been obtained.

C Removed question - 3.2 Hard EM for Advertisements

We now have the same setting as in 3.1 but all readers have not necessarily read all editions. A reader n having read an edition l is given by the latent variable $Z_{nl}^{re} \in \{0, 1\}$ which has a Bernoulli distribution with parameter θ. So, now

$$\Omega = (\theta, \theta^r, \theta^a, \theta^e, \{\psi_{c,d,f} : c \in [C], d \in [D], f \in [F]\})$$

where

$$p(Z_{nl}^{re}) = \theta^{I\{Z_{nl}^{re}=1\}}(1-\theta)^{1-I\{Z_{nl}^{re}=1\}}$$

and the other parameters keep the distribution they had in the previous question. $\mathbf{Z}$ is now extended to

$$\mathbf{Z} = (\mathbf{Z}^{re}, \mathbf{Z}^r, \mathbf{Z}^e, \mathbf{Z}^a)$$

where

- $\mathbf{Z}^{re} = \{Z_{nl}^{re} \mid n \in [N], l \in [L]\}$
- $\mathbf{Z}^r = \{Z_l^r \mid n \in [N]\}$
- $\mathbf{Z}^e = \{Z_l^e \mid l \in [L]\}$
- $\mathbf{Z}^a = \{Z_m^a \mid m \in [M]\}$.

We now get

$$p(\mathbf{X}, \mathbf{Z} \mid \Omega) = p(\mathbf{X} \mid \mathbf{Z}, \Omega)p(\mathbf{Z} \mid \Omega) = p(\mathbf{X} \mid \mathbf{Z}^r, \mathbf{Z}^e, \mathbf{Z}^a, \mathbf{Z}^{re}, \Omega)p(\mathbf{Z}^r, \mathbf{Z}^e, \mathbf{Z}^a, \mathbf{Z}^{re} \mid \Omega)$$

$$= p(\mathbf{X} \mid \mathbf{Z}^r, \mathbf{Z}^e, \mathbf{Z}^a, \mathbf{Z}^{re}, \Omega)p(\mathbf{Z}^r, \mathbf{Z}^e, \mathbf{Z}^a \mid \mathbf{Z}^{re}, \Omega)p(\mathbf{Z}^{re} \mid \Omega) = \{\textit{conditional indepence}\}$$

$$= p(\mathbf{X} \mid \mathbf{Z}^r, \mathbf{Z}^e, \mathbf{Z}^a, \mathbf{Z}^{re}, \Omega)p(\mathbf{Z}^r, \mathbf{Z}^e, \mathbf{Z}^a \mid \Omega)p(\mathbf{Z}^{re} \mid \Omega)$$

$$= \left(\prod_{n=1}^{N} \prod_{l=1}^{L} \prod_{m \in A(l)} \prod_{c=1}^{C} \prod_{d=1}^{D} \prod_{f=1}^{F} \left(\psi_{c,d,f}^{X_{n,l,m}} (1-\psi_{c,d,f})^{1-X_{n,l,m}} \right)^{I\{Z_n^r=c, Z_l^e=d, Z_m^a=f\} I\{Z_{nl}^{re}=1\}} \right)$$
$$\cdot \left(\prod_{n=1}^{N} \prod_{l=1}^{L} \prod_{m \in A(l)} \prod_{c=1}^{C} \prod_{d=1}^{D} \prod_{f=1}^{F} \left(\theta_c^r \theta_d^e \theta_f^a \right)^{I\{Z_n^r=c, Z_l^e=d, Z_m^a=f\}} \right) \left(\prod_{n=1}^{N} \prod_{l=1}^{L} \theta^{I\{Z_{nl}^{re}=1\}} (1-\theta)^{1-I\{Z_{nl}^{re}=1\}} \right)$$

$$= \prod_{n=1}^{N} \prod_{l=1}^{L} \left[\theta^{I\{Z_{nl}^{re}=1\}} \prod_{m \in A(l)} \prod_{c=1}^{C} \prod_{d=1}^{D} \prod_{f=1}^{F} \left(\theta_c^r \theta_d^e \theta_f^a \left(\psi_{c,d,f}^{X_{n,l,m}} (1-\psi_{c,d,f})^{1-X_{n,l,m}} \right)^{I\{Z_{nl}^{re}=1\}} \right)^{I\{Z_n^r=c, Z_l^e=d, Z_m^a=f\}} \right]$$
$$\cdot \left[(1-\theta)^{1-I\{Z_{nl}^{re}=1\}} \right]$$

We now go over to the log-space to simplify the expressions (and avoid numerical underflow if this algorithm would have been implemented).

$$
\begin{aligned}
&\log(p(\mathbf{X},\mathbf{Z}\mid\Omega))\\
&=\sum_{n=1}^{N}\sum_{l=1}^{L}\big[I\{Z_{nl}^{re}=1\}\log(\theta)+(1-I\{Z_{nl}^{re}=1\})\log(1-\theta)\big]\\
&+\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{d=1}^{D}\sum_{f=1}^{F}I\{Z_n^r=c,Z_l^e=d,Z_m^a=f\}\big[\log\theta_c^r+\log\theta_d^e+\log\theta_f^a\big]\\
&+\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{d=1}^{D}\sum_{f=1}^{F}I\{Z_n^r=c,Z_l^e=d,Z_m^a=f\}I\{Z_{nl}^{re}=1\}\\
&\cdot\big[X_{n,l,m}\log\psi_{c,d,f}+(1-X_{n,l,m})\log(1-\psi_{c,d,f})\big]
\end{aligned}
$$

Note that the sum has been split into three separate sums. We will now calculate $Q(\Omega,\Omega^{old})$ where

$$
Q(\Omega,\Omega^{old})=\sum_{\mathbf{Z}}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[\log p(\mathbf{X},\mathbf{Z}\mid\Omega)]
$$

Due to the linearity of the expected value we can split the sums directly

$$
\begin{aligned}
&Q(\Omega,\Omega^{old})=\sum_{\mathbf{Z}}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[\log p(\mathbf{X},\mathbf{Z}\mid\Omega)]\\
&=\sum_{n=1}^{N}\sum_{l=1}^{L}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}\left[I\{Z_{nl}^{re}=1\}\log\theta+(1-I\{Z_{nl}^{re}=1\})\log(1-\theta)\right]\\
&+\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{d=1}^{D}\sum_{f=1}^{F}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}\Big[I\{Z_n^r=c,Z_l^e=d,Z_m^a=f\}\big[\log\theta_c^r+\log\theta_d^e+\log\theta_f^a\big]\Big]\\
&+\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{d=1}^{D}\sum_{f=1}^{F}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}\Big[I\{Z_n^r=c,Z_l^e=d,Z_m^a=f\}I\{Z_{nl}^{re}=1\}\\
&\cdot\big[X_{n,l,m}\log\psi_{c,d,f}+(1-X_{n,l,m})\log(1-\psi_{c,d,f})\big]\Big]
\end{aligned}
\tag{65}
$$

$$
\begin{aligned}
&=\sum_{n=1}^{N}\sum_{l=1}^{L}\Big(\log(1-\theta)+\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}\left[I\{Z_{nl}^{re}=1\}\right]\big(\log(\theta)-\log(1-\theta)\big)\Big)\\
&+\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{d=1}^{D}\sum_{f=1}^{F}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}\Big[I\{Z_n^r=c,Z_l^e=d,Z_m^a=f\}\Big]\big[\log\theta_c^r+\log\theta_d^e+\log\theta_f^a\big]\\
&+\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{d=1}^{D}\sum_{f=1}^{F}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}\Big[I\{Z_n^r=c,Z_l^e=d,Z_m^a=f,Z_{nl}^{re}=1\}\Big]\\
&\cdot\big[X_{n,l,m}\log\psi_{c,d,f}+(1-X_{n,l,m})\log(1-\psi_{c,d,f})\big]
\end{aligned}
$$

We can now calculate the three expected values separately.

The first expectation

$$\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[I\{Z_{nl}^{re}=1\}]$$

$$= \sum_{a\in\{0,1\}}\sum_{c=1}^{C}\sum_{d=1}^{D} I\{Z_{nl}^{re}=1\}p(Z_{nl}^{re}=a, Z_n^r=c, Z_l^e=d \mid \mathbf{X}_{n,l},\Omega^{old})$$

$$= \sum_{c=1}^{C}\sum_{d=1}^{D} p(Z_{nl}^{re}=1, Z_n^r=c, Z_l^e=d \mid \mathbf{X}_{n,l},\Omega^{old})$$

$$= \sum_{c=1}^{C}\sum_{d=1}^{D} \frac{p(Z_{nl}^{re}=1, Z_n^r=c, Z_l^e=d, \mathbf{X}_{n,l} \mid \Omega^{old})}{p(\mathbf{X}_{n,l} \mid \Omega^{old})}$$

$$= \frac{\sum_{c=1}^{C}\sum_{d=1}^{D} p(\mathbf{X}_{n,l} \mid Z_{nl}^{re}=1, Z_n^r, Z_l^e, \Omega^{old})p(Z_n^r=c, Z_l^e=d \mid \Omega^{old})p(Z_{nl}^{re}=1 \mid \Omega^{old})}{p(\mathbf{X}_{n,l} \mid \Omega^{old})}$$

The numerator becomes

$$\sum_{c=1}^{C}\sum_{d=1}^{D} p(\mathbf{X}_{n,l} \mid Z_{nl}^{re}=1, Z_n^r=c, Z_l^e=d, \Omega^{old})p(Z_n^r=c, Z_l^e=d \mid \Omega^{old})p(Z_{nl}^{re}=1 \mid \Omega^{old})$$

$$= \sum_{c=1}^{C}\sum_{d=1}^{D} \theta^{old}\theta_c^{r,old}\theta_d^{e,old} p(\mathbf{X}_{n,l} \mid Z_{nl}^{re}=1, Z_n^r=c, Z_l^e=d, \Omega^{old})$$

$$= \sum_{c=1}^{C}\sum_{d=1}^{D} \theta^{old}\theta_c^{r,old}\theta_d^{e,old} \prod_{m\in A(l)} p(X_{n,l,m} \mid Z_{nl}^{re}=1, Z_n^r=c, Z_l^e=d\Omega^{old})$$

$$= \sum_{c=1}^{C}\sum_{d=1}^{D} \theta^{old}\theta_c^{r,old}\theta_d^{e,old} \prod_{m\in A(l)} \sum_{f=1}^{F} p(X_{n,l,m}, Z_m^a=f \mid Z_{nl}^{re}=1, Z_n^r=c, Z_l^e=d, \Omega^{old})$$

$$= \sum_{c=1}^{C}\sum_{d=1}^{D} \theta^{old}\theta_c^{r,old}\theta_d^{e,old} \prod_{m\in A(l)} \sum_{f=1}^{F} p(X_{n,l,m} \mid Z_{nl}^{re}=1, Z_n^r=c, Z_l^e=d, Z_m^a=f, \Omega^{old})p(Z_m^a=f \mid \Omega^{old})$$

$$= \theta^{old}\sum_{c=1}^{C}\sum_{d=1}^{D} \theta_c^{r,old}\theta_d^{e,old} \prod_{m\in A(l)} \sum_{f=1}^{F} \theta_f^{a,old}\big(\psi^{old})_{c,d,f}^{X_{n,l,m}}(1-(\psi^{old})_{c,d,f})^{1-X_{n,l,m}}\big)$$

and the denominator becomes

$$p(\mathbf{X}_{n,l} \mid \Omega^{old})$$

$$= \sum_{a\in\{0,1\}} \sum_{c=1}^{C} \sum_{d=1}^{D} p(\mathbf{X}_{n,l} \mid Z_{nl}^{re} = a, Z_n^r = c, Z_l^e = d, \Omega^{old}) p(Z_n^r = c, Z_l^e = d \mid \Omega^{old}) p(Z_{nl}^{re} = a \mid \Omega^{old})$$

$$= \sum_{c=1}^{C} \sum_{d=1}^{D} p(\mathbf{X}_{n,l} \mid Z_{nl}^{re} = 0, Z_n^r = c, Z_l^e = d, \Omega^{old}) p(Z_n^r = c, Z_l^e = d \mid \Omega^{old}) p(Z_{nl}^{re} = 0 \mid \Omega^{old})$$

$$+ \sum_{c=1}^{C} \sum_{d=1}^{D} p(\mathbf{X}_{n,l} \mid Z_{nl}^{re} = 1, Z_n^r = c, Z_l^e = d, \Omega^{old}) p(Z_n^r = c, Z_l^e = d \mid \Omega^{old}) p(Z_{nl}^{re} = 1 \mid \Omega^{old})$$

Here we recognize that the second term is the same expression as the numerator. We thus only have to calculate the first term.

Using the same reasoning as in the numerator but now with $Z_{nl}^{re} = 0$ we get:

$$\sum_{c=1}^{C} \sum_{d=1}^{D} p(\mathbf{X}_{n,l} \mid Z_{nl}^{re} = 0, Z_n^r = c, Z_l^e = d, \Omega^{old}) p(Z_n^r = c, Z_l^e = d \mid \Omega^{old}) p(Z_{nl}^{re} = 0 \mid \Omega^{old})$$

$$= \sum_{c=1}^{C} \sum_{d=1}^{D} (1-\theta^{old}) \theta_c^{r,old} \theta_d^{e,old} p(\mathbf{X}_{n,l} \mid Z_{nl}^{re} = 0, Z_n^r = c, Z_l^e = d, \Omega^{old})$$

$$= \sum_{c=1}^{C} \sum_{d=1}^{D} (1-\theta^{old}) \theta_c^{r,old} \theta_d^{e,old} \prod_{m\in A(l)} p(X_{n,l,m} \mid Z_{nl}^{re} = 0, Z_n^r = c, Z_l^e = d, \Omega^{old})$$

$$= \sum_{c=1}^{C} \sum_{d=1}^{D} (1-\theta^{old}) \theta_c^{r,old} \theta_d^{e,old} \prod_{m\in A(l)} \sum_{f=1}^{F} \underbrace{p(X_{n,l,m} \mid Z_{nl}^{re} = 0, Z_n^r = c, Z_l^e = d, Z_m^a = f, \Omega^{old})}_{=1} p(Z_m^a = f \mid \Omega^{old})$$

$$= (1-\theta^{old}) \sum_{c=1}^{C} \sum_{d=1}^{D} \theta_c^{r,old} \theta_d^{e,old} \prod_{m\in A(l)} \underbrace{\sum_{f=1}^{F} \theta_f^{a,old}}_{=1} = (1-\theta^{old}) \sum_{c=1}^{C} \sum_{d=1}^{D} \theta_c^{r,old} \theta_d^{e,old}$$

$$= \left\{ \sum_{c=1}^{C} \sum_{d=1}^{D} \theta_c^{r,old} \theta_d^{e,old} = \sum_{c=1}^{C} \theta_c^{r,old} \sum_{d=1}^{D} \theta_d^{e,old} = 1 \right\} = (1-\theta^{old})$$

So

$$p(\mathbf{X}_{n,l} \mid \Omega^{old}) = (1-\theta^{old}) + \theta^{old} \sum_{c=1}^{C} \sum_{d=1}^{D} \theta_c^{r,old} \theta_d^{e,old} \prod_{m \in A(l)} \sum_{f=1}^{F} \theta_f^{a,old} \big(\psi^{old}\big)_{c,d,f}^{X_{n,l,m}} (1-(\psi^{old})_{c,d,f})^{1-X_{n,l,m}} \big)$$

Finally

$$\gamma_{n,l} := \mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[I\{Z_{nl}^{re} = 1\}]$$

$$= \frac{\theta^{old} \sum_{c=1}^{C} \sum_{d=1}^{D} \theta_c^{r,old} \theta_d^{e,old} \prod_{m \in A(l)} \sum_{f=1}^{F} \theta_f^{a,old} \big(\psi^{old}\big)_{c,d,f}^{X_{n,l,m}} (1 - (\psi^{old})_{c,d,f})^{1-X_{n,l,m}} \big)}{(1-\theta^{old}) + \theta^{old} \sum_{c=1}^{C} \sum_{d=1}^{D} \theta_c^{r,old} \theta_d^{e,old} \prod_{m \in A(l)} \sum_{f=1}^{F} \theta_f^{a,old} \big(\psi^{old}\big)_{c,d,f}^{X_{n,l,m}} (1 - (\psi^{old})_{c,d,f})^{1-X_{n,l,m}} \big)} \tag{66}$$

The second expectation

Now to the second expectation

$$\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\}]$$

$$= \sum_{a \in \{0,1\}} \sum_{c'=1}^{C} \sum_{d'=1}^{D} \sum_{f'=1}^{F} I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\} p(Z_{nl}^{re} = a, Z_n^r = c', Z_l^e = d', Z_m^a = f' \mid X_{n,l,m}, \Omega^{old})$$

$$= \sum_{a \in \{0,1\}} p(Z_{nl}^{re} = a, Z_n^r = c, Z_l^e = d, Z_m^a = f \mid X_{n,l,m}, \Omega^{old})$$

$$= \sum_{a \in \{0,1\}} \frac{p(Z_{nl}^{re} = a, Z_n^r = c, Z_l^e = d, Z_m^a = f, X_{n,l,m} \mid \Omega^{old})}{p(X_{n,l,m} \mid \Omega^{old})}$$

$$= \frac{\sum_{Z_{nl}^{re}} p(X_{n,l,m} \mid Z_{nl}^{re}, Z_n^r = c, Z_l^e = d, Z_m^a = f, \Omega^{old}) p(Z_n^r = c, Z_l^e = d, Z_m^a = f \mid \Omega^{old}) p(Z_{nl}^{re} \mid \Omega^{old})}{p(X_{n,l,m} \mid \Omega^{old})}$$

The denominator becomes

$$p(X_{n,l,m} \mid \Omega^{old})$$

$$= \sum_{Z_{nl}^{re}} \sum_{Z_n^r} \sum_{Z_l^e} \sum_{Z_m^a} p(X_{n,l,m} \mid Z_{nl}^{re}, Z_n^r, Z_l^e, Z_m^a, \Omega^{old}) p(Z_n^r, Z_l^e, Z_m^a \mid \Omega^{old}) p(Z_{nl}^{re} \mid \Omega^{old})$$

$$= \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{f=1}^{F} (1-\theta^{old}) \theta_c^{r,old} \theta_d^{e,old} \theta_f^{a,old}$$
$$+ \theta^{old} \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{f=1}^{F} \theta_c^{r,old} \theta_d^{e,old} \theta_f^{a,old} (\psi^{old})_{c,d,f}^{X_{n,l,m}} (1 - (\psi^{old})_{c,d,f})^{1-X_{n,l,m}}$$

$$= \left\{ \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{f=1}^{F} (1-\theta^{old}) \theta_c^{r,old} \theta_d^{e,old} \theta_f^{a,old} = (1-\theta^{old}) \sum_{c=1}^{C} \theta_c^{r,old} \sum_{d=1}^{D} \theta_d^{e,old} \sum_{f=1}^{F} \theta_f^{a,old} = (1-\theta^{old}) \right\}$$
$$= (1-\theta^{old}) + \theta^{old} \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{f=1}^{F} \theta_n^{r,old} \theta_d^{e,old} \theta_f^{a,old} (\psi^{old})_{c,d,f}^{X_{n,l,m}} (1 - (\psi^{old})_{c,d,f})^{1-X_{n,l,m}}$$

and the numerator now becomes

$$\sum_{Z_{nl}^{re}} p(X_{n,l,m} \mid Z_{nl}^{re}, Z_n^r = c, Z_l^e = d, Z_m^a = f, \Omega^{old}) p(Z_n^r = c, Z_l^e = d, Z_m^a = f \mid \Omega^{old}) p(Z_{nl}^{re} \mid \Omega^{old})$$

$$= (\psi^{old})_{c,d,f}^{X_{n,l,m}} (1 - (\psi^{old})_{c,d,f})^{1-X_{n,l,m}} \theta_n^{r,old} \theta_d^{e,old} \theta_f^{a,old} \theta^{old} + \theta_n^{r,old} \theta_d^{e,old} \theta_f^{a,old} (1 - \theta^{old})$$

$$= \theta_n^{r,old} \theta_d^{e,old} \theta_f^{a,old} \bigg(\theta^{old} (\psi^{old})_{c,d,f}^{X_{n,l,m}} (1 - (\psi^{old})_{c,d,f})^{1-X_{n,l,m}} + (1 - \theta^{old}) \bigg)$$

So

$$\phi_{c,d,f}^{n,l,m} := \mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[I\{Z_n^r = c, Z_l^e = d, Z_m^a = f\}]$$

$$= \frac{\theta_c^{r,old}\theta_d^{e,old}\theta_f^{a,old}\Big(\theta^{old}(\psi^{old})_{c,d,f}^{X_{n,l,m}}(1-(\psi^{old})_{c,d,f})^{1-X_{n,l,m}} + (1-\theta^{old})\Big)}{(1-\theta^{old}) + \theta^{old}\sum_{c=1}^{C}\sum_{d=1}^{D}\sum_{f=1}^{F}\theta_n^{r,old}\theta_d^{e,old}\theta_f^{a,old}(\psi^{old})_{c,d,f}^{X_{n,l,m}}(1-(\psi^{old})_{c,d,f})^{1-X_{n,l,m}}} \tag{67}$$

The third expectation

Now to the final expectation.

$$\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}\Big[I\{Z_n^r = c, Z_l^e = d, Z_m^a = f, Z_{nl}^{re} = 1\}\Big]$$

$$= \sum_{a\in\{0,1\}}\sum_{c'=1}^{C}\sum_{d'=1}^{D}\sum_{f'=1}^{F} I\{Z_n^r = c, Z_l^e = d, Z_m^a = f, Z_{nl}^{re} = 1\}p(Z_{nl}^{re} = a, Z_n^r = c', Z_l^e = d', Z_m^a = f' \mid X_{n,l,m}, \Omega^{old})$$

$$= p(Z_{nl}^{re} = 1, Z_n^r = c, Z_l^e = d, Z_m^a = f \mid X_{n,l,m}, \Omega^{old})$$

$$= \frac{p(Z_{nl}^{re} = 1, Z_n^r = c, Z_l^e = d, Z_m^a = f, X_{n,l,m} \mid \Omega^{old})}{p(X_{n,l,m} \mid \Omega^{old})}$$

$$= \frac{p(X_{n,l,m} \mid Z_{nl}^{re} = 1, Z_n^r = c, Z_l^e = d, Z_m^a = f, \Omega^{old})p(Z_n^r = c, Z_l^e = d, Z_m^a = f \mid \Omega^{old})p(Z_{nl}^{re} = 1 \mid \Omega^{old})}{p(X_{n,l,m} \mid \Omega^{old})}$$

We notice that the denominator is the same as for the second expectation. Now the numerator is one part of the sum of the previous numerator where $Z_{nl}^{re} = 1$ which gives,

$$\xi_{c,d,f}^{n,l,m} := \mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}\Big[I\{Z_n^r = c, Z_l^e = d, Z_m^a = f, Z_{nl}^{re} = 1\}\Big]$$

$$= \frac{\theta_c^{r,old}\theta_d^{e,old}\theta_f^{a,old}\theta^{old}(\psi^{old})_{c,d,f}^{X_{n,l,m}}(1-(\psi^{old})_{c,d,f})^{1-X_{n,l,m}}}{(1-\theta^{old}) + \theta^{old}\sum_{c=1}^{C}\sum_{d=1}^{D}\sum_{f=1}^{F}\theta_n^{r,old}\theta_d^{e,old}\theta_f^{a,old}(\psi^{old})_{c,d,f}^{X_{n,l,m}}(1-(\psi^{old})_{c,d,f})^{1-X_{n,l,m}}} \tag{68}$$

Optimization of the parameters

We can now re-write (69) like follows

$$
\begin{aligned}
Q(\Omega, \Omega^{old}) &= \sum_{\mathbf{Z}} \mathbb{E}_{\mathbf{Z}|\mathbf{X},\Omega^{old}}[\log p(\mathbf{X}, \mathbf{Z} \mid \Omega)] \\
&= \sum_{n=1}^{N} \sum_{l=1}^{L} \Big(\log(1-\theta) + \gamma_{n,l}\big(\log\theta - \log(1-\theta)\big)\Big) \\
&+ \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m\in A(l)} \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{f=1}^{F} \phi_{c,d,f}^{n,l,m} \big[\log\theta_c^r + \log\theta_d^e + \log\theta_f^a\big] \\
&+ \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m\in A(l)} \sum_{c=1}^{C} \sum_{d=1}^{D} \sum_{f=1}^{F} \xi_{c,d,f}^{n,l,m} \big[X_{n,l,m}\log\psi_{c,d,f} + (1-X_{n,l,m})\log(1-\psi_{c,d,f})\big]
\end{aligned} \tag{69}
$$

Starting of by optimizing for θ we take the first partial derivative of Q w.r.t to θ and set it to zero.

$$
\frac{\partial Q}{\partial \theta} = \sum_{n=1}^{N} \sum_{l=1}^{L} \Big(-\frac{1}{1-\theta} + \gamma_{n,l}\Big(\frac{1}{\theta} + \frac{1}{1-\theta}\Big)\Big) = 0
$$

$$
\sum_{n=1}^{N} \sum_{l=1}^{L} \frac{1}{1-\theta} = \sum_{n=1}^{N} \sum_{l=1}^{L} \Big(\gamma_{n,l}\Big(\frac{1}{\theta} + \frac{1}{1-\theta}\Big)\Big)
$$

$$
NL\frac{1}{1-\theta} = \Big(\frac{1}{\theta} + \frac{1}{1-\theta}\Big) \sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{n,l} = \frac{1}{\theta(1-\theta)} \sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{n,l}
$$

And thus we get

$$
\theta^{new} = \frac{\sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{n,l}}{NL} \tag{70}
$$

The derivations of the other parameters are quite similar in comparison to the previous question, but the expected values are different.

We start of with $(\theta^r, \theta^a, \theta^e)$ which we know that they must satisfy $\sum_{c=1}^{C} \theta_c^r = 1$, $\sum_{d=1}^{D} \theta_d^e = 1$ and $\sum_{f=1}^{F} \theta_f^a = 1$ since they are probabilities.

Using Lagrange relaxation on the θ_c^r we get the Lagrangian

$$\begin{aligned} L(\theta_c^r, \lambda) =& Q(\Omega, \Omega^{old}) - \lambda\Big(-1 + \sum_{c'=1}^{C} \theta_{c'}^r\Big) \\ \overset{+}{=}& \sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{d=1}^{D}\sum_{f=1}^{F} \phi_{c,d,f}^{n,l,m} \log \theta_c^r - \lambda\Big(-1 + \sum_{c'=1}^{C} \theta_{c'}^r\Big) \end{aligned}$$

Firstly, the partial derivative w.r.t. θ_c^r gives

$$\frac{\partial L}{\partial \theta_c^r} = \sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{d=1}^{D}\sum_{f=1}^{F} \phi_{c,d,f}^{n,l,m} \frac{1}{\theta_c^r} - \lambda = 0$$

$$\theta_c^r = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{d=1}^{D}\sum_{f=1}^{F} \phi_{c,d,f}^{n,l,m}}{\lambda}$$

Secondly, the partial derivative w.r.t. λ gives

$$\frac{\partial L}{\partial \lambda} = 1 - \sum_{c'=1}^{C} \theta_{c'}^r = 0$$

$$\sum_{c'=1}^{C} \theta_{c'}^r = 1$$

Now,

$$\lambda = \lambda \cdot 1 = \lambda \sum_{c'=1}^{C} \theta_{c'}^r = \sum_{c'=1}^{C} \theta_{c'}^r \lambda = \sum_{c'=1}^{C}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{d=1}^{D}\sum_{f=1}^{F} \phi_{c',d,f}^{n,l,m}$$

And thus,

$$\theta_c^{r,new} = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{d=1}^{D}\sum_{f=1}^{F} \phi_{c,d,f}^{n,l,m}}{\sum_{c'=1}^{C}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{d=1}^{D}\sum_{f=1}^{F} \phi_{c',d,f}^{n,l,m}} \tag{71}$$

Repeating these steps in the same manner for θ_d^e and θ_f^a gives

$$\theta_d^{e,new} = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{f=1}^{F} \phi_{c,d,f}^{n,l,m}}{\sum_{d'=1}^{D}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{f=1}^{F} \phi_{c,d',f}^{n,l,m}} \tag{72}$$

and

$$\theta_f^{a,new} = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{d=1}^{D} \phi_{c,d,f}^{n,l,m}}{\sum_{f'=1}^{F}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\sum_{c=1}^{C}\sum_{d=1}^{D} \phi_{c,d,f'}^{n,l,m}} \tag{73}$$

Finally, we will now optimize w.r.t to $\psi_{c,d,f}$. Setting the first derivative to zero and solving we get

$$\frac{\partial Q}{\partial \psi_{c,d,f}} = \sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)} \xi_{c,d,f}^{n,l,m}\Big(X_{n,l,m}\frac{1}{\psi_{c,d,f}} + (X_{n,l,m}-1)\frac{1}{1-\psi_{c,d,f}}\Big) = 0$$

$$\frac{1}{\psi_{c,d,f}}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\xi_{c,d,f}^{n,l,m}X_{n,l,m}=\frac{1}{1-\psi_{c,d,f}}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\xi_{c,d,f}^{n,l,m}(1-X_{n,l,m})$$

$$(1-\psi_{c,d,f})\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\xi_{c,d,f}^{n,l,m}X_{n,l,m}=\psi_{c,d,f}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\xi_{c,d,f}^{n,l,m}(1-X_{n,l,m})$$

$$\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\xi_{c,d,f}^{n,l,m}X_{n,l,m}=\psi_{c,d,f}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\xi_{c,d,f}^{n,l,m}$$

$$\psi_{c,d,f}^{new}=\frac{\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\xi_{c,d,f}^{n,l,m}X_{n,l,m}}{\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m\in A(l)}\xi_{c,d,f}^{n,l,m}} \tag{74}$$

The EM algorithm

Now to the EM algorithm itself.

1. Initialize by choosing a suitable Ω^{old}, e.g.
   - $\theta^{old} = 0.5$
   - $\theta_c^{r,old} = \frac{1}{C}$ for all c
   - $\theta_d^{e,old} = \frac{1}{D}$ for all d
   - $\theta_f^{a,old} = \frac{1}{F}$ for all f
   - $\psi_{c,d,f}^{old} = 0.1$ for all c, d, f.
2. Evaluate Ω^{new} by calculating
   - θ^{new} according to equation (70)
   - $\theta_c^{r,new}$ according to equation (71)
   - $\theta_d^{e,new}$ according to equation (72)
   - $\theta_f^{a,new}$ according to equation (73)
   - $\psi_{c,d,f}^{new}$ according to equation (74).
3. Check if the parameters of Ω have converged. If not, let

$$\Omega^{old} \leftarrow \Omega^{new}$$

   and return to step 2.