Assignment 1: EM for Categorical Data Advanced Signal Processing

Daniel Barrejón Moreno

January 28, 2019

1 Problem formulation

In this project we want to study a set of documents with a model following a mixture of categorical distributions. For a document $\mathbf{x}$ belonging to a set of documents $\mathbf{X}$, the marginal li,kelihood is expressed as follows

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{D} \text{Cat}(x_j|\boldsymbol{\theta}_k), \tag{1}$$

where K is the number of mixtures (in our case the topics), D is the number of words that appear in a certain document and $\text{Cat}(x_j|\boldsymbol{\theta}_k)$ is the categorical distribution with I categories and parameter $\boldsymbol{\theta}_k$ which represents the probability that a certain category appears. Bear in mind that the model parameters are defined as $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k, \boldsymbol{\pi}\}$. Since $\boldsymbol{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,m}, \ldots, \theta_{k,I})$ is a vector with the probabilities of topic k probability, it must satisfy two constraints:

$$0 < \theta_{k,m} < 1 \tag{2}$$

$$\text{and} \quad \sum_{m=1}^{I} \theta_{k,m} = 1. \tag{3}$$

The expression for the log li,kelihood of the observed data $\mathcal{D}$ looks li,ke this

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{D} \text{Cat}(x_j|\boldsymbol{\theta}_k). \tag{4}$$

However this function is hard to optimize due to the sum inside the log. In order to solve this we introduce the latent variable $\mathcal{Z}$. The marginal distribution for the hidden variable z_i is defined by the mixing coefficient of the mixture π_k as follows

$$p(z_i = k) = \pi_k. \tag{5}$$

Since $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ is a vector with the probabilities for the mixtures, certain restrictions must be satisfied:

$$0 < \pi_k < 1 \tag{6}$$

$$\text{and} \quad \sum_{k=1}^{K} \pi_k = 1. \tag{7}$$

Now we need to find the expression for the complete data log li,kelihood $p(\mathcal{D}, \mathcal{Z})$. First of all, the marginal of z_i is given by

$$p(z_i|\boldsymbol{\theta}) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}\{z_i=k\}}, \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The probability density function for x_i given z_i is given by the following expression

$$p(\mathbf{x}_i|z_i, \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k)^{\mathbb{I}\{z_i=k\}}. \tag{9}$$

Since we are interested in the joint probability, applying Bayes' rule

$$p\left(\mathbf{x}_i, z_i|\boldsymbol{\theta}\right) = p\left(\mathbf{x}_i|z_i, \boldsymbol{\theta}\right) p(z_i|\boldsymbol{\theta}) \tag{10}$$

yields the resulting expression

$$p\left(\mathbf{x}_i, z_i|\boldsymbol{\theta}\right) = \prod_{k=1}^{K} \left(\pi_k \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k) \right)^{\mathbb{I}\{z_i=k\}}. \tag{11}$$

2 Complete data log likelihood $l_c(\boldsymbol{\theta})$

Assuming that our observed data is independent and identically distributed (i.i.d) and taking natural logarithm we can find expression of the complete data log li,kelihood

$$l_c(\boldsymbol{\theta}) = \log p(\mathcal{D}, \mathcal{Z}|\boldsymbol{\theta}) = \log \prod_{n=1}^{N} p\left(\mathbf{x}_i, z_i|\boldsymbol{\theta}\right) \tag{12}$$

$$= \sum_{i=1}^{N} \log p\left(\mathbf{x}_i, z_i|\boldsymbol{\theta}\right) \tag{13}$$

$$= \sum_{i=1}^{N} \log \prod_{k=1}^{K} \left(\pi_k \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k) \right)^{\mathbb{I}\{z_i=k\}} \tag{14}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \left(\pi_k \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k) \right). \tag{15}$$

We can simplify the expression for the marginal of $\mathbf{x}_i$ as

$$p(\mathbf{x}_i) = \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}) = \prod_{j=1}^{D}\prod_{m=1}^{I} \theta_m^{\mathbb{I}\{x_{i,j}=m\}} = \prod_{m=1}^{I}\prod_{j=1}^{D} \theta_m^{\mathbb{I}\{x_{i,j}=m\}} \tag{16}$$

$$= \prod_{m=1}^{I} \theta_m^{\sum_{j=1}^{D}\mathbb{I}\{x_{i,j}=m\}} = \prod_{m=1}^{I} \theta_m^{\mu_{i,m}}, \tag{17}$$

where we have defined a new metric $\mu_{i,m}$ that represents the number of times the word associated to the category m appears at document i, *i.e.* ,

$$\mu_{i,m} = \sum_{j=1}^{D} \mathbb{I}(x_{i,j} = m). \tag{18}$$

Therefore, the final expression will be

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \left(\pi_k \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k) \right) \tag{19}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \prod_{m=1}^{I} \theta_{k,m}^{\mu_{i,m}} \tag{20}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m}. \tag{21}$$

3 ML Inference

In order to solve the maximum li,kelihood problem, we define an auxiliary function Q which will be the expected complete data log li,kelihood over the hidden variable $\mathcal{Z}$. This function Q will be evaluated in the E-step and maximized for the model parameters $\boldsymbol{\theta}$ to update the model parameters in each iteration of the algorithm.

3.1 E-step: $Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1})$ for ML inference

Taking the expectation with respect to the latent variables z we get the following expected complete data log-li,kelihood.

$$Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}) = \mathbb{E}_Z\{l_c(\boldsymbol{\theta})|\mathcal{D}, \boldsymbol{\theta}^{t-1}\} \tag{22}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_Z\{\mathbb{I}(z_i = k)\} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_Z\{\mathbb{I}(z_i = k)\} \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m} \tag{23}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m}, \tag{24}$$

where we have defined the metric $r_{i,k} \triangleq p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{t-1})$ [1] as the responsibility that mixture k, *i.e.* topic k, takes at explaining the document $\mathbf{x}_i$. It is defined as follows

$$r_{i,k} = \mathbb{E}_Z\{\mathbb{I}(z_i = k)\} = p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{t-1}) \tag{25}$$

$$= \frac{p(z_i = k, \mathbf{x}_i|\boldsymbol{\theta}^{t-1})}{p(\mathbf{x}_i|\boldsymbol{\theta}^{t-1})} = \frac{p(z_i = k, \mathbf{x}_i|\boldsymbol{\theta}^{t-1})}{\sum_{k\prime}^{K} p(z_i = k\prime, \mathbf{x}_i|\boldsymbol{\theta}^{t-1})} \tag{26}$$

$$= \frac{\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k)}{\sum_{k\prime}^{K} \pi_{k\prime} p(\mathbf{x}_i|\boldsymbol{\theta}_{k\prime})} = \frac{\pi_k \prod_{j=1}^{D} \text{Cat}(x_{i,j}|\boldsymbol{\theta}_k)}{\sum_{k\prime}^{K} \pi_{k\prime} \prod_{j=1}^{D} \text{Cat}(x_{i,j}|\boldsymbol{\theta}_{k\prime})}. \tag{27}$$

This quantity must satisfy the following constrains, given that z it is a probability

$$0 \leq r_{i,k} \leq 1, \tag{28}$$

$$\text{and} \quad \sum_{k=1}^{K} r_{i,k} = 1. \tag{29}$$

3.1.1 Arithmetic underflow solved by log-sum-exp trick

In our problem we have I categories which correspond to the words from a dictionary. When we perform the product in Equation 25 we obtain really small values that cannot be represented by the computer. This problem is known as **arithmetic underflow**. However, in order to maintain the range in probability if categories $\theta_{k,m}$ that influence the value of $r_{i,k}$ we use the **log-sum-exp trick** [2].
Taking the logarithm from Equation 25 and using Equation 16 for $p(\mathbf{x}_i|\boldsymbol{\theta}_k)$ we obtain the following expression

$$\log r_{i,k} = \log \pi_k \prod_{j=1}^{D} \text{Cat}(x_{i,j}|\boldsymbol{\theta}_k) - \log \sum_{k\prime=1}^{K} \pi_{k\prime} \prod_{j=1}^{D} \text{Cat}(x_{i,j}|\boldsymbol{\theta}_{k\prime}) \tag{30}$$

$$= \log \pi_k + \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m} - \log \sum_{k\prime=1}^{K} \pi_{k\prime} \prod_{m=1}^{I} \theta_{k\prime,m}^{\mu_{i,m}}. \tag{31}$$

The log-sum-exp trick will be applied on the last term from the above equation. The trick states that

$$\log \sum_{v=1}^{V} e^{g_v} = a + \log \sum_{v=1}^{V} e^{g_v - a}, \tag{32}$$

where $a = \max_v g_v$ and g_v is defined as

$$g_v = \log \pi_{k\prime} \prod_{m=1}^{I} \theta_{k\prime,m}^{\mu_{i,m}} = \log \pi_{k\prime} + \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k\prime,m} \tag{33}$$

Once $\log r_{i,k}$ from Equation 30 is known, we can find $r_{i,k}$ with an exponential. With this trick we do not loose any information by forcing values assigned to be 0 by the computer using clipping and we also increase computational speed.

3.2 M-step for ML inference

Now we need to find the closed-form formulas to update the model parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k, \boldsymbol{\pi}\}$. Since we must tackle a maximization problem with constraints we will apply Lagrange Multipliers to solve equation

$$\boldsymbol{\theta}^t = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}\right). \tag{34}$$

3.2.1 Maximization of π_k

Using the constraints on π from Equations 6 we propose as Lagrangian the function

$$L\left(Q(\pi_k), \lambda\right) = Q(\pi_k) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right), \tag{35}$$

which will be optimized in this way

$$\min_{\lambda} \max_{\pi_k} \{L\left(Q(\pi_k), \lambda\right)\}. \tag{36}$$

First, we take the derivative of Equation 35 w.r.t π_k and equate it to 0, which yields

$$\frac{\partial L}{\partial \pi_k} = 0 = \sum_{i=1}^{N} \frac{r_{i,k}}{\pi_k} - \lambda, \tag{37}$$

$$\pi_k = \frac{1}{\lambda} \sum_{i=1}^{N} r_{i,k}. \tag{38}$$

Now, we take the derivate w.r.t λ which yields

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 = 0, \tag{39}$$

$$\sum_{k=1}^{K} \pi_k = 1. \tag{40}$$

If we sum over k at both sides of Equation 38

$$\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} \frac{1}{\lambda} \sum_{i=1}^{N} r_{i,k} \tag{41}$$

$$1 = \frac{1}{\lambda} \sum_{i=1}^{N} \sum_{k=1}^{K} r_{i,k}, \tag{42}$$

and using the constraints on $r_{i,k}$ from Equations 28 we get the value of λ

$$1 = \frac{1}{\lambda} \sum_{i=1}^{N} 1. \tag{43}$$

$$\lambda = N. \tag{44}$$

Knowing the value of λ the estimated value of $\hat{\pi}_k$ can be expressed as follows

$$\hat{\pi}_k = \frac{1}{N}\sum_{i=1}^{N} r_{i,k}s = \frac{N_k}{N}, \tag{45}$$

where

$$N_k = \sum_{i=1}^{N} r_{i,k}. \tag{46}$$

This result is actually intuitive since N_k represents the 'weight' of topic k at explaining the documents, and therefore π_k is just the percentage of topic k at explaining the data.

3.2.2 Maximization of $\boldsymbol{\theta}_k$

We follow a similar approach. Now, using the constraints on $\boldsymbol{\theta}_k$ from Equations 2 we propose as Lagrangian the following function

$$L\left(Q(\theta_{k,m}), \lambda\right) = Q(\theta_{k,m}) + \lambda\left(\sum_{m=1}^{I}\theta_{k,m} - 1\right), \tag{47}$$

which will be optimized as a min-max problem

$$\min_{\lambda}\max_{\theta_{km}}\{L\left(Q(\theta_{km}), \lambda\right)\}. \tag{48}$$

Firstly, we take the derivative of Equation 47 w.r.t $\theta_{k,m}$ and equate it to 0, which yields

$$\frac{\partial L}{\partial \theta_{k,m}} = \sum_{i=1}^{N}\frac{r_{i,k}\mu_{i,m}}{\theta_{k,m}} - \lambda = 0, \tag{49}$$

$$\theta_{k,m} = \frac{1}{\lambda}\sum_{i=1}^{N} r_{i,k}\mu_{i,m}. \tag{50}$$

Secondly, we take the derivative w.r.t λ, which yields

$$\frac{\partial L}{\partial \lambda} = \sum_{m=1}^{I}\theta_{k,m} - 1 = 0, \tag{51}$$

$$\sum_{m=1}^{I}\theta_{k,m} = 1. \tag{52}$$

If now we sum over I at both sides of Equation 50, the value of λ is obtained

$$\sum_{m=1}^{I}\theta_{k,m} = \sum_{m=1}^{I}\frac{1}{\lambda}\sum_{i=1}^{N} r_{i,k}\mu_{i,m} \tag{53}$$

$$1 = \frac{1}{\lambda}\sum_{i=1}^{N}\sum_{m=1}^{I} r_{i,k}\mu_{i,m}, \tag{54}$$

$$\lambda = \sum_{i=1}^{N}\sum_{m=1}^{I} r_{i,k}\mu_{i,m}. \tag{55}$$

Using the value of λ in Equation 50, $\theta_{k,m}$ is obtained

$$\hat{\theta}_{k,m} = \frac{\sum_{i=1}^{N} r_{i,k}\mu_{i,m}}{\sum_{i=1}^{N}\sum_{m=1}^{I} r_{i,k}\mu_{i,m}}. \tag{56}$$

Again, the result is quite intuitive. $\theta_{k,m}$ is just the average of the category m weighted by the responsibility $r_{i,k}$.

4 MAP inference

Maximum li,kelihood estimations is an estimation that tends to overfit [1]. A solution to such problem is applying maximum a posteriori estimation (MAP). In this case, we do not only consider the li,kelihood, but also some prior information on the model parameters $\boldsymbol{\theta}$. From Bayes' rule we know that

$$p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{Z}) \propto p(\mathcal{D}, \mathcal{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{57}$$

If we take logarithms at both sides of the equation we get

$$\log p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{Z}) \propto \log p(\mathcal{D}, \mathcal{Z}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}). \tag{58}$$

Notice that the first term is just the complete data log li,kelihood from Equation 21 and the second term corresponds to the prior information. The goal now is to define some prior over the model parameters $\boldsymbol{\theta}$ and use it on the function Q from Equation 24 to work with the posterior instead of the li,kelihood. Afterwards, we need to reformulate Equation 45 and 56 for π_k and $\theta_{k,m}$ to take into account the priors.

4.1 E-step: $Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1})$ for MAP inference

From [3] and [1] we know it is natural that the prior on the mixture weights π and the category probabilities $\boldsymbol{\theta}_k$ follow a Dirichlet distribution, *i.e.* ,

$$\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\beta}), \quad \text{s.t.} \quad p(\boldsymbol{\pi}|\boldsymbol{\beta}) = \frac{1}{\text{B}(\boldsymbol{\beta})}\prod_{k=1}^{K}\pi_k^{\beta_k - 1}, \tag{59}$$

$$\boldsymbol{\theta}_k \sim \text{Dir}(\boldsymbol{\alpha}), \quad \text{s.t.} \quad p(\boldsymbol{\theta}_k|\boldsymbol{\alpha}) = \frac{1}{\text{B}(\boldsymbol{\alpha})}\prod_{m=1}^{I}\theta_{k,m}^{\alpha_m - 1}, \tag{60}$$

where the function $\text{B}(\cdot)$ stands for the Beta function, and the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the parameters or hyperpriors of the Dirichlet distributions for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_k$ respectively.

The expression for the new Q function is as follows

$$Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m} + \log p(\boldsymbol{\pi}|\boldsymbol{\beta}) + \sum_{k=1}^{K} \log p(\boldsymbol{\theta}_k|\boldsymbol{\alpha}), \tag{61}$$

where $r_{i,k}$ remains the same. Taking logarithm on the probability of the priors we get

$$\log p(\boldsymbol{\pi}|\boldsymbol{\beta}) = \log\left(\frac{1}{\mathrm{B}(\boldsymbol{\beta})}\prod_{k=1}^{K}\pi_k^{\beta_k-1}\right) = \log\frac{1}{\mathrm{B}(\boldsymbol{\beta})} + \sum_{k=1}^{K}(\beta_k-1)\log(\pi_k), \tag{62}$$

$$\log p(\boldsymbol{\theta}_k|\boldsymbol{\alpha}) = \log\left(\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{m=1}^{I}\theta_{k,m}^{\alpha_m-1}\right) = \log\frac{1}{\mathrm{B}(\boldsymbol{\alpha})} + \sum_{m=1}^{I}(\alpha_m-1)\log(\theta_{k,m}). \tag{63}$$

Using these results in Equation 61 we get the expression for $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$ using MAP estimation

$$Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m} \tag{64}$$

$$+ \log\left(\frac{1}{\mathrm{B}(\boldsymbol{\beta})}\right) + \sum_{k=1}^{K}(\beta_k-1)\log(\pi_k) \tag{65}$$

$$+ \sum_{k=1}^{K}\log\frac{1}{\mathrm{B}(\boldsymbol{\alpha})} + \sum_{k=1}^{K}\sum_{m=1}^{I}(\alpha_m-1)\log(\theta_{k,m}). \tag{66}$$

Notice that $\mathrm{B}(\boldsymbol{\alpha})$ can be decomposed as

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)},$$

where $\Gamma(\cdot)$ is the Gamma function. The same applies for $\mathrm{B}(\boldsymbol{\beta})$

4.2 M step for MAP inference

We will follow the same procedure as for the ML case; but now we must use function 64 instead to take into account the priors.

4.2.1 MAP estimation of π_k

Again, we have the same maximization problem and hence we propose the same Lagrangian as in Equation 35 and we use the same constraints on π_k

$$L\left(Q(\pi_k), \lambda\right) = Q(\pi_k) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right). \tag{67}$$

Again, we derivative Equation 64 w.r.t π_k and equate it to 0

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{N} \frac{r_{i,k}}{\pi_k} + \frac{\beta_k - 1}{\pi_k} - \lambda = 0, \tag{68}$$

$$\pi_k = \frac{1}{\lambda} \left(\sum_{i=1}^{N} r_{i,k} + \beta_k - 1 \right). \tag{69}$$

Now we derivate w.r.t λ

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 = 0, \tag{70}$$

$$\sum_{k=1}^{K} \pi_k = 1. \tag{71}$$

Summing over K at both sides of Equation 69

$$\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} \frac{1}{\lambda} \left(\sum_{i=1}^{N} r_{i,k} + \beta_k - 1 \right) \tag{72}$$

$$1 = \frac{1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{N} r_{i,k} + \frac{1}{\lambda} \sum_{k=1}^{K} \beta_k - \frac{1}{\lambda} \sum_{k=1}^{K} 1, \tag{73}$$

we get the value of λ

$$\lambda = N + \sum_{k=1}^{K} \beta_k - K. \tag{74}$$

Once λ is known the estimated value of π_k is as follows

$$\hat{\pi}_k = \frac{\sum_{i=1}^{N} r_{i,k} + \beta_k - 1}{N + \sum_{k=1}^{K} \beta_k - K}. \tag{75}$$

4.2.2 MAP estimation of $\theta_{k,m}$

For the estimation of $\theta_{k,m}$ we use the same Lagrangian from Equation 47 and the same restrictions from 2

$$L\left(Q(\theta_{k,m}), \lambda\right) = Q(\theta_{k,m}) + \lambda \left(\sum_{m=1}^{I} \theta_{k,m} - 1 \right). \tag{76}$$

As before, we first derivate w.r.t $\theta_{k,m}$, which yields

$$\frac{\partial L}{\partial \theta_{k,m}} = \sum_{i=1}^{N} \frac{r_{i,k}\mu_{im}}{\theta_{k,m}} + \frac{\alpha_m - 1}{\theta_{km}} - \lambda = 0, \tag{77}$$

$$\theta_{k,m} = \frac{1}{\lambda} \left(\sum_{i=1}^{N} r_{i,k}\mu_{i,m} + \alpha_m - 1 \right). \tag{78}$$

And later with respect to λ

$$\frac{\partial L}{\partial \lambda} = 0 = \sum_{m=1}^{I} \theta_{k,m} - 1, \tag{79}$$

$$\sum_{m=1}^{I} \theta_{k,m} = 1. \tag{80}$$

Summing at both sides over I Equation 78 we can obtain the value of λ

$$\sum_{m=1}^{I} \theta_{k,m} = \sum_{m=1}^{I} \frac{1}{\lambda} \left(\sum_{i=1}^{N} r_{i,k} \mu_{i,m} + \alpha_m - 1 \right) \tag{81}$$

$$1 = \frac{1}{\lambda} \sum_{m=1}^{I} \left(\sum_{i=1}^{N} \frac{r_{i,k} \mu_{i,m}}{\theta_{k,m}} + \frac{\alpha_m}{\theta_{k,m}} - 1 \right), \tag{82}$$

$$\lambda = \sum_{m=1}^{I} \sum_{i=1}^{N} r_{i,k} \mu_{i,m} + \sum_{m=1}^{I} \alpha_m - \sum_{m=1}^{I} 1, \tag{83}$$

$$\lambda = \sum_{m=1}^{I} \sum_{i=1}^{N} r_{i,k} \mu_{i,m} + \sum_{m=1}^{I} \alpha_m - I. \tag{84}$$

And with that value of λ the estimated value of $\theta_{k,m}$ is

$$\hat{\theta}_{k,m} = \frac{\sum_{i=1}^{N} r_{i,k} \mu_{im} + \alpha_m - 1}{\sum_{m=1}^{I} \sum_{i=1}^{N} r_{i,k} \mu_{i,m} + \sum_{m=1}^{I} \alpha_m - I}. \tag{85}$$

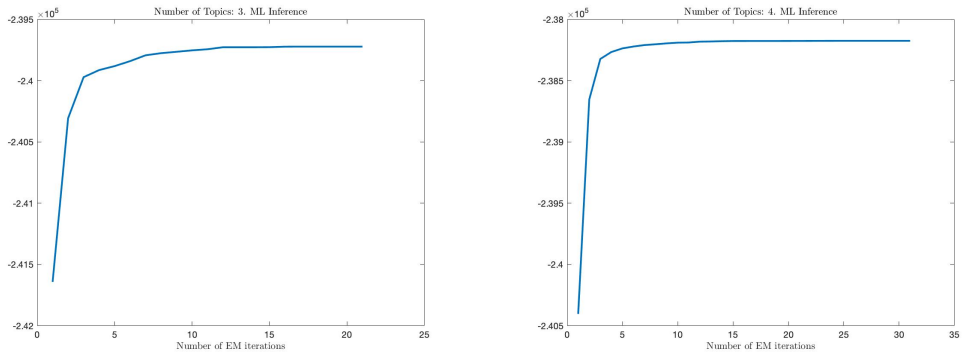
5 Experiments

The experiments have been divided into two sets: a set of experiment using ML estimation and another set using MAP estimation. For both sets of experiments the maximum number of iterations of the EM algorithm has been set to 100 iterations, the tolerance value has been set to 10^{-3} and a total of 5 different initializations have been done.

In order to assess the performance of the algorithm, for the two approaches the log-likelihood will be shown, as well as the responsibilities for each document $r_{i,k}$. In order to obtain a better visualization of the results, a word cloud for the different found topic will be shown, so that the most common words for each topic can be displayed and they could be associated to a certain abstract theme.

5.1 ML Experiments

From the setup described above, it can be seen from Figure 1 that the log-likelihood is always increasing, and it actually converges reasonably fast. As it is known from the literature, the

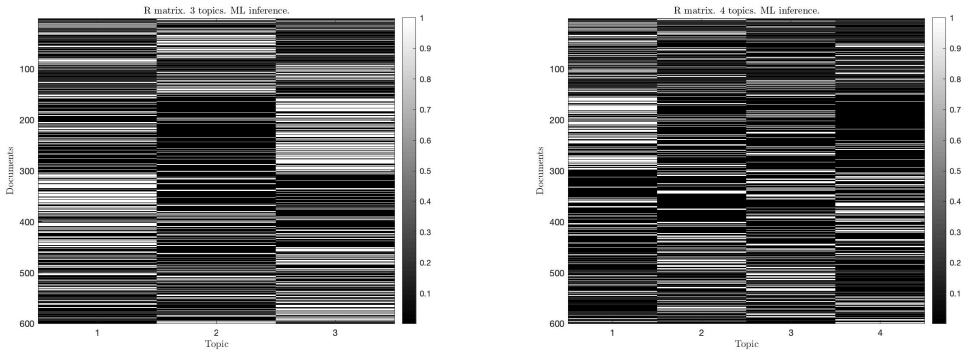


(a) (b)

Figure 1. The log-likelihood is increasing for both $K = 3$ in Figure 1a and for $K = 4$ in Figure 1b

EM algorithm is quite likely to achieve a good solution rather fast, and then it tends to continue increasing in a plateau-fashion way.

In Figure 2 it can bee seen the probabilities for each document for belonging to a certain topic. This result is obtained from the matrix of $r_{i,k}$. Although it is clear that each document is associated to a certain topic, still the results is quite scattered. With detailed observation it can be seen that there are some segments of following documents that might belong to a same topic, and hence we could differentiate different segments of documents.



(a) (b)

Figure 2. Figures 2a and 2b show the probability of a certain document to be in a specific topic. As it can be seen, the result is quite scattered. This will be solved with the prior information in the MAP estimation.

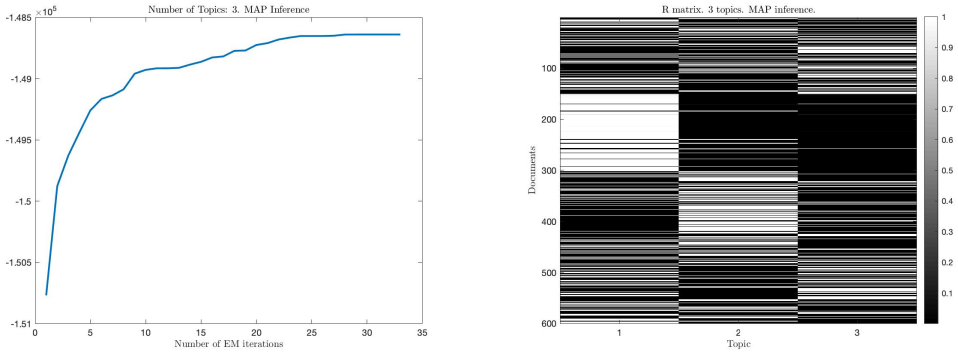
5.2 MAP Experiments

From Figure 2 it could be seen that with the ML inference approach it is not possible to achieve the desired result of segments of documents with the same topic. To obtain this we need to incorporate some prior information about how we believe the topics and the categories

of each document are distributed. To do so, we show three different experiments. For all of them the hyperparameter $\boldsymbol{\beta}$ has been set as $\boldsymbol{\beta} = 2\mathbf{I}_{Kx1}$ and $\boldsymbol{\alpha}$ was set to $\boldsymbol{\alpha} = 1.1\mathbf{I}_{Ix1}$, $\boldsymbol{\alpha} = 5\mathbf{I}_{Ix1}$ and finally $\boldsymbol{\alpha} = 2\mathbf{I}_{Ix1}$. For smaller values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the order of 0.5 the algorithm breaks since the results for Equations 85 and 75 are close to 0 and therefore the evaluation of the posterior in Equation 64 returns complex values.

5.2.1 Experiment 1: $\boldsymbol{\alpha} = 1.1\mathbf{I}_{Ix1}$

In this case, choosing $\boldsymbol{\alpha} = 1.1\mathbf{I}_{Ix1}$ is almost the same as choosing it uniformly, meaning that the prior probability for each different category of the categorical distribution is the same. As it was expected, the results is quite similar to the result obtained with the ML approach. The log-likelihood and the $r_{i,k}$ results are depicted in Figure 3 for value $K = 3$.



(a) (b)

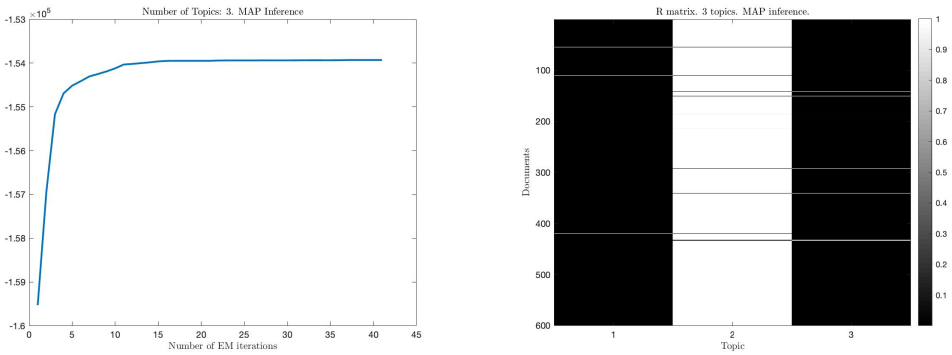
Figure 3. The increase in the posterior in Figure 3a is not as fast as the ML approach. In Figure 3b it can bee seen that thanks to the prior information the different segments of documents are starting to appear.

5.2.2 Experiment 2: $\boldsymbol{\alpha} = 5\mathbf{I}_{Ix1}$

Now, instead of using a rather uniform distribution for the hyperparameters of the prior we choose a more concentrated distribution by selecting $\boldsymbol{\alpha} = 5\mathbf{I}_{Ix1}$. However, the results are not good as it can be seen in Figure 4. Although the posterior is increasing quite smoothly, when we check the responsibilities $r_{i,k}$ it is clear that the algorithm is just assigning almost all the documents to one topic, and this is clearly not the case.

5.2.3 Experiment 3: $\boldsymbol{\alpha} = 2\mathbf{I}_{Ix1}$

The best solution was actually found for values $\boldsymbol{\alpha} = 2\mathbf{I}_{Ix1}$. In this case the posterior does not increase as fast and smooth as it should, since sometimes some saddle points appear. However, when we check the responsibilities associated to each document, it can be seen that the algorithm is performing rather good. Besides, we will compare the results for different values of $K = 2, 3, 4, 5$.



(a) (b)

Figure 4. Posterior and responsibilities values for the document using $K = 3$ and $\boldsymbol{\alpha} = 5\mathbf{I}_{Ix1}$.

In Figure 5 the different increments of the posterior can be seen. As mentioned above, the increasing for the MAP approach is not as smooth as the ML approach since it depends on the choose of the hyperparameters of the prior. However, the results for the $r_{i,k}$ are quite reasonable, as it can be check in Figure 6. For $K = 2$ we start to see that the document collection is divided into different segments belonging to the same topic. However, two topics is too general. The best results is found for $K = 3$ because we can differentiate three different segments of documents very clearly. For $K = 4$ notice that the 4th topic could actually be merged with the first one to form the same topic as in the case with $K = 3$. Finally, $K = 5$ could be simplified again to three since the 4th topic could be merged with the first topic, and the third topic could be merged with the fifth topic to form $K = 3$.

This has been our model selection criterion for selecting $K = 3$. At the beginning we tried two Bayesian model criteria, the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC), which are defined with the following formulas

$$\text{AIC} = 2\text{dof}(\boldsymbol{\theta}) - 2\ln(\hat{L}) \tag{86}$$

$$\text{BIC} = \ln(n)\text{dof}(\boldsymbol{\theta}) - 2\ln(\hat{L}), \tag{87}$$

where $(\hat{L})$ is the maximized value of the likelihood function and $\text{dof}(\boldsymbol{\theta})$ is the number of parameters or degrees of freedom of the model. In our case, the number of parameters of the model is

$$\text{dof}(\boldsymbol{\theta}) = K(I-1) + (K-1) = KI - 1, \tag{88}$$

given by $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. Since the number of parameters increases with K, and the likelihood does not increase that much for different K, both criteria were always increasing, and therefore we did not use them to choose the optimal value of K. This is illustrated in Figure 7 where we can see that for increasing number of topics K the log-likelihood always increase and so does the BIC.

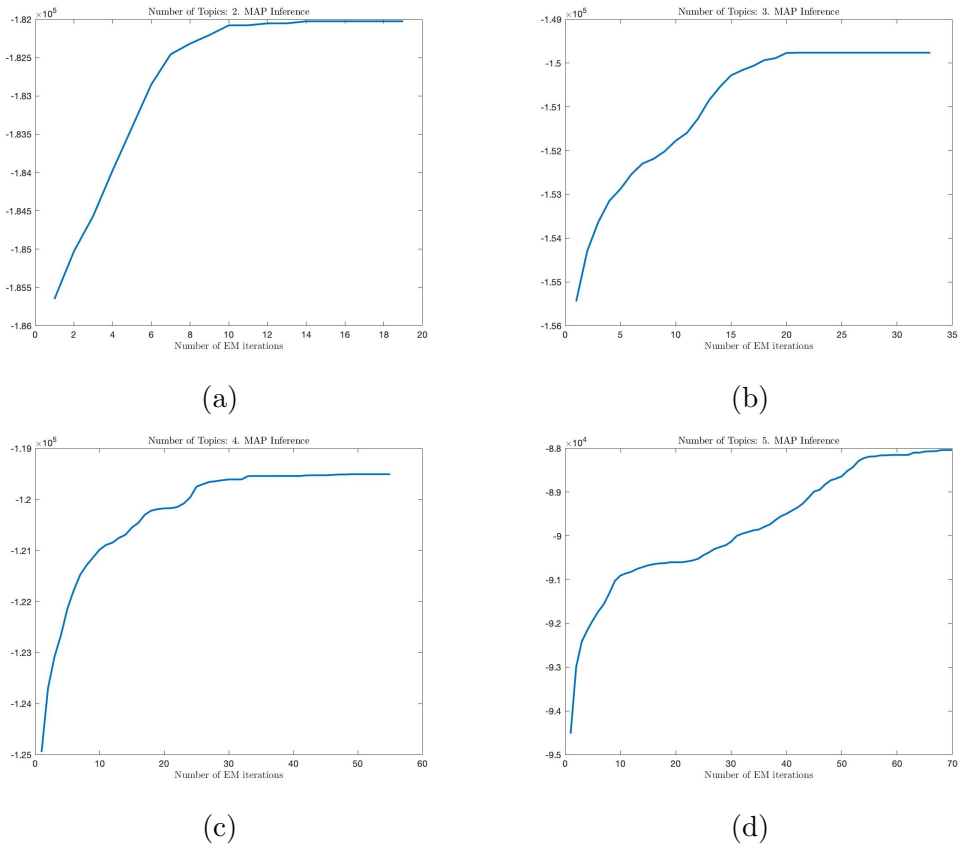


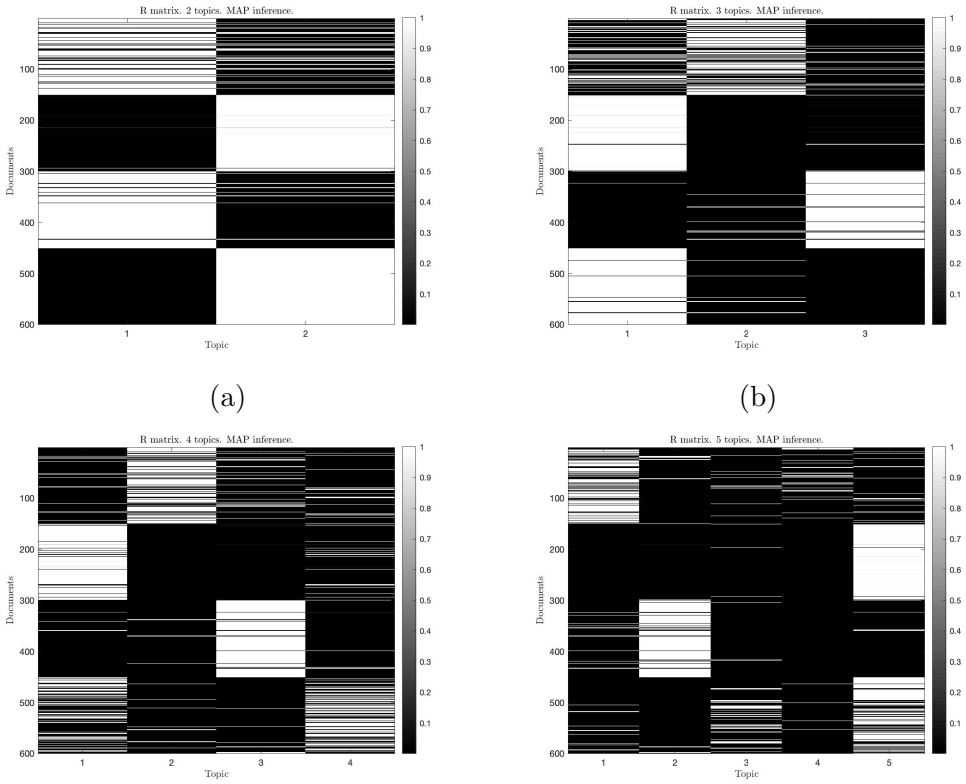
Figure 5. Increase of the posterior for different values of K.

6 WordCloud visualization

In order to gain more visual intuition of the solution, we show two Wordclouds generated for Experiment 3 5.2.3 in Figure 8. From the documents belonging to a same topic we obtain the most common word of that certain document so that for each topic we can create a Wordcloud with the most common words per topic. Whit this results it is rather visual to check if a topic is representing a field of research of the provided data.

7 Conclusions

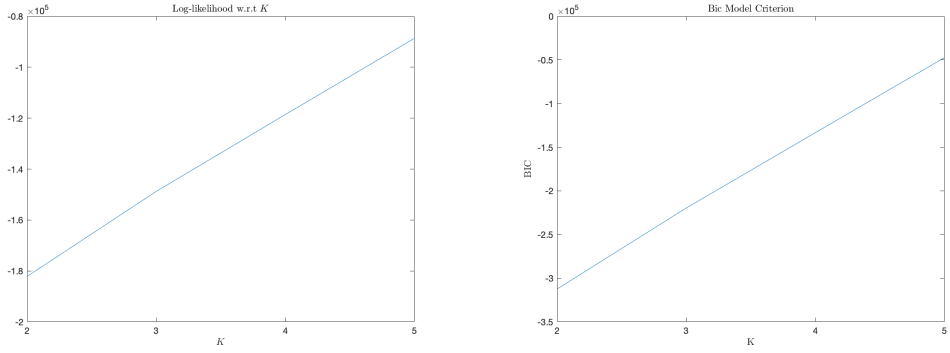
Throughout this project we have studied the maximum likelihood and the maximum a posterior approach for the Expectation Maximization algorithm applied to a collection of documents following a categorical distribution. It has been shown that using a model in which we know some prior information can achieve better performance characteristics and that the number of mixtures is an important hyperparameter that must be selected according to some criterion. In our case, given the big number of parameters of the model and the small change in the likelihoods for different values of K we could not use model information criteria and therefore we analyzed the corresponding outputs of the model.



(a) (b)

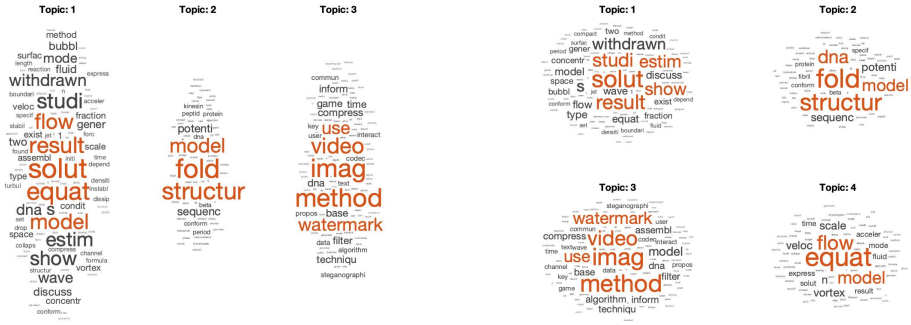
(c) (d)

Figure 6. Responsibilities for each document for different values of K.



(a) (b)

Figure 7. Figure 7a shows the log-likelihood increasing for different values of K. Figure 7b shows the BIC model criterion.



(a) (b)

Figure 8. The most common words in Topic 1 in Figure 8a are present in both Topic 1 and Topic 4 8b. Therefore, Topic 1 and 4 could be merged together. Notice that each topic is related to a certain field of research from the abstracts.

References

[1] K. P. Murphy, *Machine learning: A probabilistic perspective. adaptive computation and machine learning*, 2012.

[2] R. Eisele, *The log-sum-exp trick in machine learning*. [Online]. Available: https://www.xarg.org/2016/06/the-log-sum-exp-trick-in-machine-learning/.

[3] A. Artés-Rodríguez, *Notes for advanced signal processing*, 2018.

[4] C. Fraley and A. E. Raftery, "Bayesian regularization for normal mixture estimation and model-based clustering," *Journal of classification*, vol. 24, no. 2, pp. 155–181, 2007.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006. [Online]. Available: http://research.microsoft.com/en-us/um/people/cmbishop/prml/.