EM FOR BINARY DATA

A.1 Advanced Signal Processing

Author:
MANUEL MONTOYA CATALÁ

Index

1. INTRODUCTION

The objective of this assignment is to develop an Expectation-Maximization (EM) algorithm for a mixture of multinomials (or multivariate Bernoullis) model.

$$p(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \pi_k p_k(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}_k}) = \sum_{k=1}^{K} \pi_k \underbrace{\left(\prod_{j=1}^{D} \theta_{jk}^{X_j} (1 - \theta_{jk})^{1-X_j} \right)}_{p_k(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}_k})}$$

We will derive the different equations needed for the algorithm, code the algorithm in MATLAB and use it on 2 data sets dna_amp_chr_17.mat and marker.mat.

This report consists of 2 parts:

1) Work Description: Brief answers to the assignment questions based on the theoretical background.
2) Theoretical Background: All the theory needed to do the assignment.

2. WORK DESCRIPTION

In this Section of the report, we will answer the questions of the assignment taking into account the Theoretical Background explained in Section **Theory background**.

2.1 Complete data log likelihood of the model

1. Write down the expression for the complete data log likelihood for the mixture of multinomial model

$$l_c(\boldsymbol{\theta}) = \ln p(\mathcal{D}, \mathcal{Z}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln(p(\mathbf{x}_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}))$$

where z_i are the hidden variables and $p(z_i = k) = \pi_k$.

As seen in Section **3.5.1**, we can obtain the complete data log likelihood as:

$$\begin{aligned}
I_C(\bar{\boldsymbol{\theta}}) = \ln\left(p(\mathcal{D}, \bar{\boldsymbol{Z}}|\bar{\boldsymbol{\theta}})\right) &= \sum_{i=1}^{N} ln\left(p(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_{\boldsymbol{i}}, \bar{\boldsymbol{Z}} = \bar{\boldsymbol{Z}}_{\boldsymbol{i}}|\bar{\boldsymbol{\theta}})\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k) \ln\left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k) \ln\left(\pi_{\boldsymbol{k}} \prod_{j=1}^{D} \theta_{jk}^{X_{ij}} (1 - \theta_{jk})^{1-X_{ij}}\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k) \left(ln(\pi_k) + \sum_{j=1}^{D} \left(X_{ij} \ln(\theta_{jk}) + (1 - X_{ij}) \ln\left((1 - \theta_{jk})\right)\right)\right) \\
&= \ln\left(\sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k) \pi_{\boldsymbol{k}} \prod_{j=1}^{D} \theta_{jk}^{X_{ij}} (1 - \theta_{jk})^{1-X_{ij}}\right)
\end{aligned}$$

2.2 Expected complete data log likelihood of the model

2. Write down the expression for the expected complete data log likelihood

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

As seen in Section **3.5.2**, we can obtain the expected complete data log likelihood as:

$$Q(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}}) = E\{I_C(\bar{\boldsymbol{\theta}})|D, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}}\} = E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k) \ln\left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)\right\}$$

$$= \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik} \ln(\pi_k)}_{Depends\ on\ \bar{\boldsymbol{\pi}}_{\boldsymbol{k}}} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N}\left(\sum_{j=1}^{D}\left(X_{ij} \ln(\theta_{jk}) + (1 - X_{ij}) \ln\left((1 - \theta_{jk})\right)\right)\right)}_{Depends\ on\ \bar{\bar{\boldsymbol{\theta}}}_{\boldsymbol{KxD}}}$$

$$With\ \ r_{ik} = \frac{\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})}{\sum_{\boldsymbol{j}=1}^{K} \pi_{\boldsymbol{j}} p_{\boldsymbol{j}}(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}_{\boldsymbol{j}})} = \frac{\pi_k \left(\prod_{j=1}^{D} \theta_{jm}^{X_{ij}} (1 - \theta_{jm})^{1-X_{ij}}\right)}{\sum_{m=1}^{K} \pi_m \left(\prod_{j=1}^{D} \theta_{jm}^{X_{ij}} (1 - \theta_{jm})^{1-X_{ij}}\right)}$$

Notice r_{ik} is calculated with the $\bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}} \rightarrow \{\bar{\boldsymbol{\pi}}_{\boldsymbol{k}}, \bar{\bar{\boldsymbol{\theta}}}_{\boldsymbol{KxS}}\}^{\boldsymbol{t-1}}$ parameters

2.3 Expression for the ML estimates

3. Derive the expression of the ML estimates of the new set of parameters $\boldsymbol{\theta}^t$

$$\boldsymbol{\theta}^t = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$$

As seen in Section **3.5.4** we can obtain the ML estimates of the new set of parameters $\bar{\boldsymbol{\theta}}^{\boldsymbol{t}} \rightarrow \{\bar{\boldsymbol{\pi}}_{\boldsymbol{k}}, \bar{\bar{\boldsymbol{\theta}}}_{\boldsymbol{KxD}}\}^{\boldsymbol{t}}$ as:

$$\bar{\boldsymbol{\pi}}_{\boldsymbol{K}} = [\pi_1, \ldots, \pi_k, .., \pi_K] \quad with \quad \pi_k = \frac{N_k}{N} = \frac{1}{N}\sum_{i=1}^{N} r_{ik}$$

$$\bar{\bar{\boldsymbol{\theta}}}_{KxD} = \begin{bmatrix} \bar{\boldsymbol{\theta}}_{\boldsymbol{1}} \\ \bar{\boldsymbol{\theta}}_{\boldsymbol{2}} \\ \vdots \\ \bar{\boldsymbol{\theta}}_{\boldsymbol{K}} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1K} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{D1} & \theta_{D2} & \cdots & \theta_{DK} \end{bmatrix} \quad with \quad \boldsymbol{\theta_{dk}} = \frac{\boldsymbol{1}}{\boldsymbol{N_k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{r_{ik} X_{id}}$$

2.4 EM algorithm implementation

4. Implement the EM algorithm for a mixture of K multinomials using Matlab code. The algorithm should take as input K, a matrix $\mathcal{D}$ containing the data set, the minimum increment in the log likelihood for convergence, and the maximum number of iterations. Hand in code and a high level explanation of what you algorithm does. (**This part can be done in groups**)

The steps of the EM algorithm for the multinomial distribution are derived in Section **3.5.5**
The source code for this is highly detailed and commented and is given into the.

1) Initilize parameters of the mixture model $\bar{\boldsymbol{\theta}} = \{\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}, \bar{\bar{\boldsymbol{\theta}}}_{KxD}\}$. at random according to the paramters constraints :

2) While (Stop_condition)

- **E-Step**: Update responsibility matrix as:

$$r = \begin{bmatrix} \bar{\boldsymbol{r}}_{\boldsymbol{1}} \\ \bar{\boldsymbol{r}}_{\boldsymbol{2}} \\ \vdots \\ \bar{\boldsymbol{r}}_{\boldsymbol{N}} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1K} \\ r_{21} & r_{22} & \cdots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NK} \end{bmatrix} \quad with \quad r_{ik} = \frac{\pi_k \left(\prod_{j=1}^{D} \theta_{jm}^{X_{ij}} (1 - \theta_{jm})^{1-X_{ij}} \right)}{\sum_{m=1}^{K} \pi_m \left(\prod_{j=1}^{D} \theta_{jm}^{X_{ij}} (1 - \theta_{jm})^{1-X_{ij}} \right)}$$

- **M-Step**: Update parameters of the mixture model $\bar{\boldsymbol{\theta}} = \{\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}, \bar{\bar{\boldsymbol{\theta}}}_{KxD}\}$

$$\bar{\boldsymbol{\pi}}_{\boldsymbol{K}} = [\pi_1, \dots, \pi_k, .., \pi_K] \quad with \quad \pi_k = \frac{N_k}{N} = \frac{1}{N} \sum_{i=1}^{N} r_{ik}$$

$$\bar{\bar{\boldsymbol{\theta}}}_{KxD} = \begin{bmatrix} \bar{\boldsymbol{\theta}}_{\boldsymbol{1}} \\ \bar{\boldsymbol{\theta}}_{\boldsymbol{2}} \\ \vdots \\ \bar{\boldsymbol{\theta}}_{\boldsymbol{K}} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1K} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{D1} & \theta_{D2} & \cdots & \theta_{DK} \end{bmatrix} \quad with \quad \boldsymbol{\theta_{dk}} = \frac{\boldsymbol{1}}{\boldsymbol{N_k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{r_{ik} X_{id}}$$

- Calculate the Complete Log likelihood for checking **convergence**:

$$I_C(\bar{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{I}(z_i = k) \left(ln(\pi_k) + \sum_{j=1}^{D} \left(X_{ij} \ln(\theta_{jk}) + (1 - X_{ij}) \ln\left((1 - \theta_{jk}) \right) \right) \right)$$

Computationally it is easier to calculate:

$$I_C(\bar{\boldsymbol{\theta}}) = \log_2 \left(\sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{I}(z_i = k) \pi_{\boldsymbol{k}} \prod_{j=1}^{D} \theta_{jk}^{X_{ij}} (1 - \theta_{jk})^{1-X_{ij}} \right)$$

2.5 EM algorithm implementation

5. Run your algorithm on the data sets for varying K = 2,3,4. Verify that the log likelihood increases at each step of EM. Report the log likelihoods obtained (measured in bits) and display the parameters found. Comment the performances of the algorithms in finding good clusters.

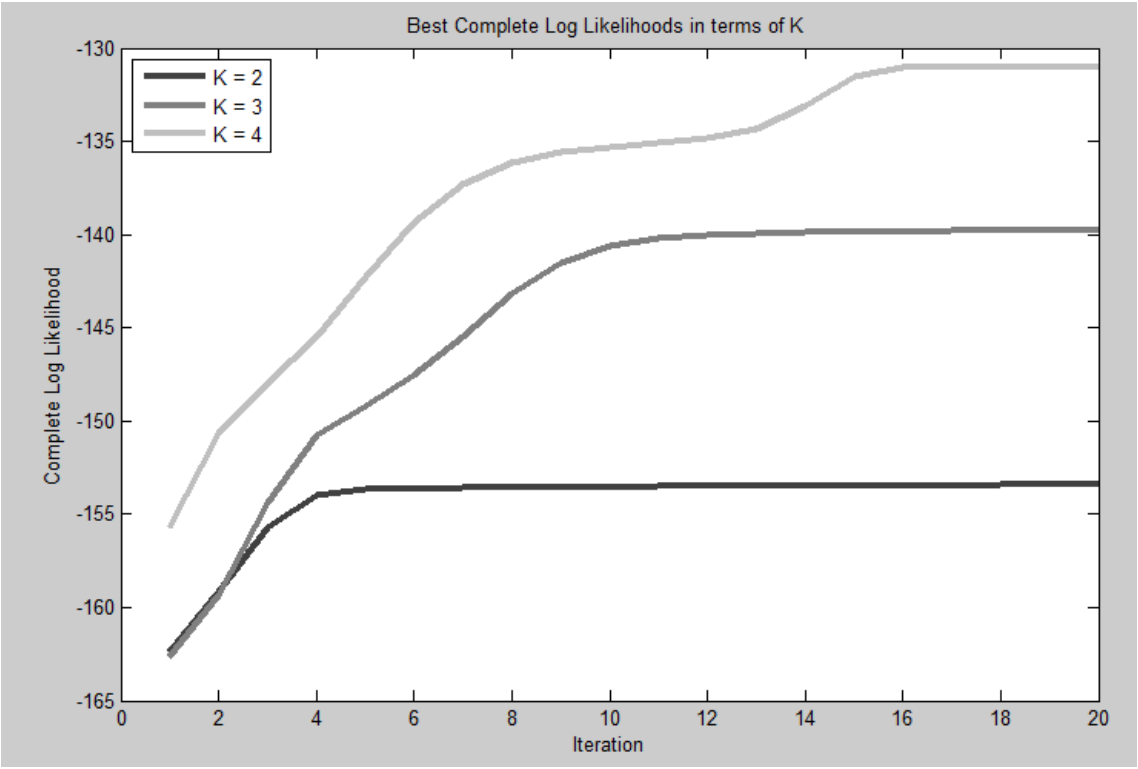
We have made an script for getting the results that:

- For different numbers of components $K = 2,3,4$
  - Runs **several times** the EM algorithm with random initialization and chooses the one with the best final complete log likelihood.
  - Plots the complete log likelihood iterations for the different values of K.
  - Displays the parameters found.

This script is called testing.m and it is given with the report.

2.5.1 Results for Marker.mat

The Complete log-likelihoods of the EM algorithm for $K = 2,3,4$ are:



From the graph it is clear that the more components we have, the better the complete log likelihood of the Data. This is an expected result since, as we increase the number of components we can describe smaller clusters in the data, but this can produce overfitting and produce clusters that are noise.

- **Parameters for K = 2**

theta

0.1476	0.1847
0.1840	0.1847
0.0190	0.9965
0.2184	0.1900
0.1476	0.3749
0.0362	0.7614

pimix

0.2974	0.7026

- **Parameters for K = 3**

theta

0.9999	0.0498	0.0000
1.0000	0.0936	0.0000
0.4114	0.0000	0.8753
0.0000	0.2595	0.2012
0.4114	0.1723	0.2005
0.2057	0.0000	0.7780

pimix

0.6744	0.0564	0.2692

- **Parameters for K = 4**

theta

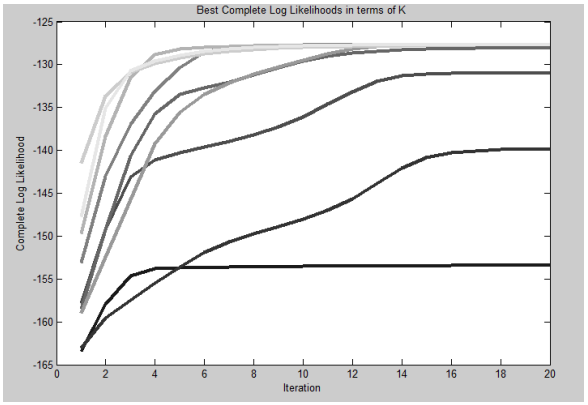
0.5891	1.0000	0.0000	0.0000
0.7364	1.0000	0.0000	0.0000
0.0000	1.0000	0.0001	0.8750
0.0000	0	0.3140	0.2000
0.0000	1.0000	0.2086	0.1995
0.0000	0.5000	0.0000	0.7778

pimix

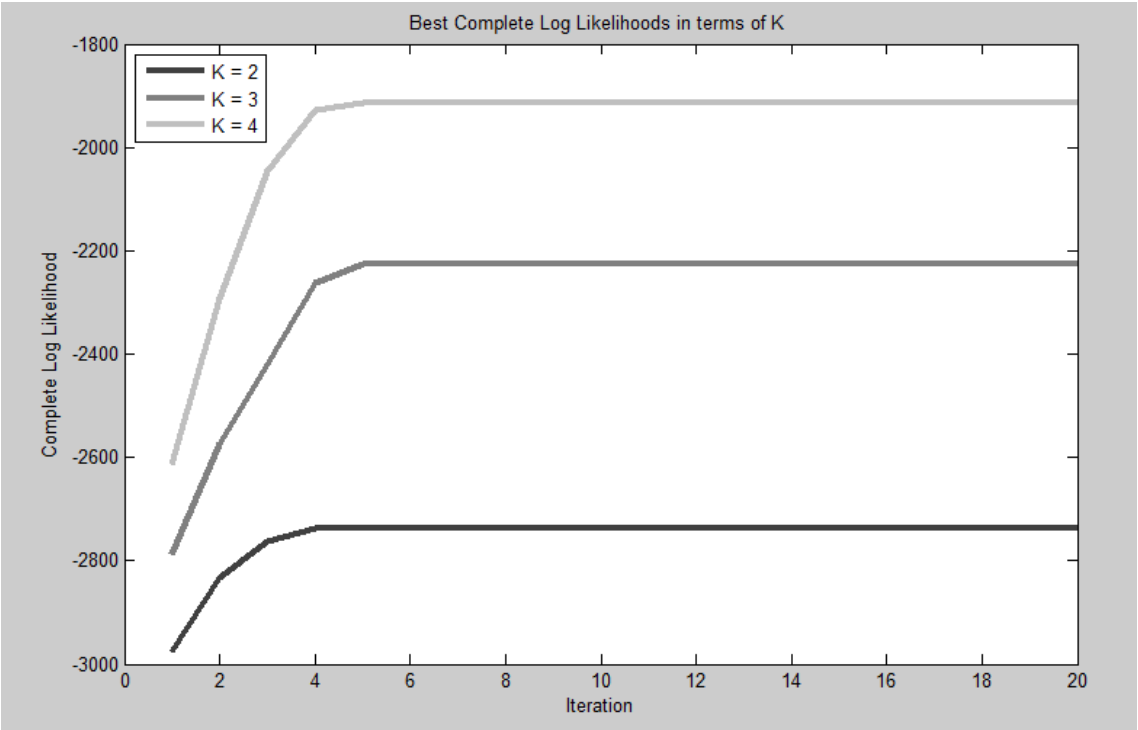
0.2712	0.2291	0.0526	0.4471

For K = 4, we see that components 1 and 2, are almost deterministic, this could indicate that we are using too many components to describe the mixture, since it is focusing in a very specific area.

The image at the right shows the complete loglikelihood for $K = 2,3,4,5,6 \ldots$ As we can see, when $K = 5 - 6$, the likelihood is saturated and no more components will increase the likelihood.



2.5.2 Results for dna_amp_chr_17.mat



From the graph it is clear that the more components we have, the better the complete log likelihood of the Data. This is an expected result since, as we increase the number of components we can describe smaller clusters in the data, but this can produce overfitting and produce clusters that are noise.

- **Parameters for K = 2**

theta

0	0.7286
0.0000	0.9406
0.0095	0.9867
0.0000	0.8214
0.1388	0.1722
0.1951	0.1855
0.3489	0.1855
0.5028	0.1722
0.4990	0.1722
0.5366	0.1722
0.6153	0.1855
0.6153	0.1855

pimix

0.4874	0.5126

- **Parameters for K = 3**

theta

0.2464	0.6719	0
0.2669	0.9062	0.0000
0.2874	0.9844	0.0000
0.2669	0.7656	0.0000

0.9033	0.0000	0.0262
1.0000	0.0156	0.0711
1.0000	0.0156	0.2499
1.0000	0.0000	0.4287
1.0000	0.0000	0.4243
1.0000	0.0000	0.4679
1.0000	0.0156	0.5595
1.0000	0.0156	0.5595

pimix

0.1871	0.6705	0.1424

- **Parameters for K = 4**

theta

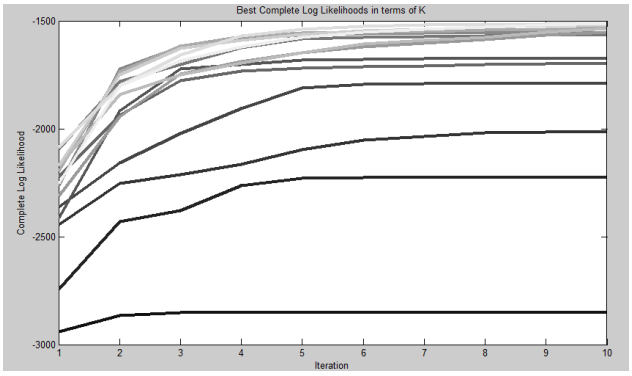
0	0.6719	0	0.2449
0.0000	0.9062	0.0000	0.2653
0.0000	0.9844	0.0000	0.2857
0.0000	0.7656	0.0000	0.2653
0.0940	0.0000	0.0000	0.8980
0.2507	0.0156	0.0000	1.0000
0.7825	0.0156	0.0427	1.0000
0.8252	0.0000	0.2745	1.0000
0.0374	0.0000	0.5728	1.0000
0.0000	0.0000	0.6478	1.0000
0.0000	0.0156	0.7749	1.0000
0.0000	0.0156	0.7749	1.0000

pimix

0.1433	0.3747	0.3171	0.1649

The cluster parameters seem ok.

The image at the right shows the complete loglikelihood for $K = 2,3,4,5,6 \ldots$ As we can see, when $K = 10 - 11,$ the likelihood is saturated and no more components will increase the likelihood.



3. THEORY BACKGROUND

3.1 Notation:

$$p(x|\mu) \rightarrow p.d.f. \quad x \sim D(\theta)$$

This is the p.d.f. of a continuous Random Variable according to a given distribution D that depends on the parameter θ. We can extend this to a case where de distribution depends on a set of parameters $\theta_1, \theta_2 \dots, \theta_N$ that can be expressed as a vector $\overline{\boldsymbol{\theta}} = [\theta_1, \theta_2 \dots, \theta_S]$.

$$p(x|\overline{\boldsymbol{\theta}}) \rightarrow p.d.f. \quad x \sim D(\overline{\boldsymbol{\theta}})$$

If we are using joint distributions of different random variables $X_1, X_2 \dots, X_L$ then we have the most general formula:

$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) \rightarrow p.d.f. \quad \overline{\boldsymbol{X}} \sim D(\overline{\boldsymbol{\theta}})$$

$$p(X_1, X_2 \dots, X_L|\theta_1, \theta_2 \dots, \theta_N) \rightarrow p.d.f. \quad \overline{\boldsymbol{X}} \sim D(\theta_1, \theta_2 \dots, \theta_N)$$

3.2 Exponential family

The exponential family is: The set of distributions over a vector random variables $\overline{\boldsymbol{X}}$, and vector of parameters $\overline{\boldsymbol{\theta}}$ that can be expressed as:

$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) = g(\overline{\boldsymbol{\theta}})\, h(\overline{\boldsymbol{X}}) e^{\left(\eta(\overline{\boldsymbol{\theta}})^{\boldsymbol{T}} \cdot \phi(\overline{\boldsymbol{X}})\right)}$$
$$= g(\overline{\boldsymbol{\theta}})\, h(\overline{\boldsymbol{X}}) e^{\left(\sum_{i=1}^{S} \eta_i(\overline{\boldsymbol{\theta}}) \cdot \phi_i(\overline{\boldsymbol{X}})\right)}$$

Where:

$\overline{\boldsymbol{X}}$: Vector of R.V. of the distribution. It can be discrete or continuous.
$\overline{\boldsymbol{\theta}}$: The *natural parameters* of the distribution.

$\phi(\overline{\boldsymbol{X}})$: Function of $\overline{\boldsymbol{X}}$ called sufficient statistic. A transformation of the space of $\overline{\boldsymbol{X}}$.
$h(\overline{\boldsymbol{X}})$: A given function of $\overline{\boldsymbol{X}}$.
$\eta(\overline{\boldsymbol{\theta}})^{\boldsymbol{T}}$: A function of the natural parameters.
$g(\overline{\boldsymbol{\theta}})$: Coefficient that ensures the distribution is normalized.

$$\int_{S_{\overline{x}}} p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) = g(\overline{\boldsymbol{\theta}}) \left[\int_{S_{\overline{x}}} h(\overline{\boldsymbol{X}}) e^{\left(\eta(\overline{\boldsymbol{\theta}})^{\boldsymbol{T}} \cdot \phi(\overline{\boldsymbol{X}})\right)} d\overline{\boldsymbol{X}} \right] = 1$$

$h(\overline{\boldsymbol{X}})$ and $\eta(\overline{\boldsymbol{\theta}})^{\boldsymbol{T}}$ are just transformations of the normal parameters of the distribution so that we can express the p.d.f. in terms of $\overline{\boldsymbol{X}}$ and $\overline{\boldsymbol{\theta}}$**.**

Another way of describing the Exponential family is:

$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) = \frac{1}{Z(\overline{\boldsymbol{\theta}})}\, h(\overline{\boldsymbol{X}}) e^{\left(\eta(\overline{\boldsymbol{\theta}})^{\boldsymbol{T}} \cdot \phi(\overline{\boldsymbol{X}})\right)}$$

Where:

$$Z(\overline{\boldsymbol{\theta}}) = \left[\int_{S_{\overline{x}}} h(\overline{\boldsymbol{X}}) e^{\left(\eta(\overline{\boldsymbol{\theta}})^{\boldsymbol{T}} \cdot \phi(\overline{\boldsymbol{X}})\right)} d\overline{\boldsymbol{X}} \right] = \frac{1}{g(\overline{\boldsymbol{\theta}})}$$

We can place $Z(\overline{\boldsymbol{\theta}})$ into the exponential resulting in:

$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) = h(\overline{\boldsymbol{X}}) e^{\left(\eta(\overline{\boldsymbol{\theta}})^{\boldsymbol{T}} \cdot \phi(\overline{\boldsymbol{X}}) - A(\overline{\boldsymbol{\theta}})\right)}$$

With:

$$A(\overline{\boldsymbol{\theta}}) = \boldsymbol{ln}\left(\boldsymbol{Z}(\overline{\boldsymbol{\theta}})\right)$$

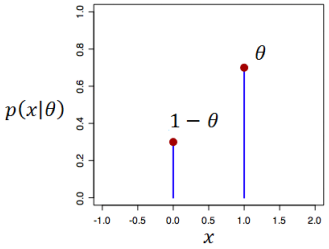
$A(\overline{\boldsymbol{\theta}})$ is called the log partition function or the cumulant function.

3.2.1 Examples

3.2.1.1 Bernoulli

The p.d.f. of a Bernoulli distribution can be expressed as:

$$p(x|\theta) = Bern(x|\theta) = \theta^x(1-\theta)^{1-x}$$



$$p(x|\mu) = e^{\ln(\theta^x(1-\theta)^{1-x})} = e^{x\,ln(\theta)+(1-x)\,ln(1-\theta)}$$
$$= e^{ln(1-\theta)+\,x(ln(\theta)-ln(1-\theta))}$$
$$= (1-\theta)e^{\left(\frac{ln(\theta)}{ln(1-\theta)}\right)x}$$

Identifying terms:

$$g(\bar{\boldsymbol{\theta}}) = (1-\theta)$$
$$h(\bar{\boldsymbol{x}}) = 1$$
$$\eta(\bar{\boldsymbol{\theta}})^{\boldsymbol{T}} = \frac{ln(\theta)}{ln(1-\theta)}$$
$$\phi(\bar{\boldsymbol{X}}) = x$$

Another way of expressing the Bernoulli p.d.f. would be:

$$p(x|\mu) = Bern(x|\theta) = \begin{bmatrix}\theta \\ (1-\theta)\end{bmatrix}[\mathbb{I}(x=1), \mathbb{I}(x=0)]$$

$$\mathbb{I}(condition) = \begin{cases}1, & if\ condition\ is\ satisfied \\ 0, & if\ condition\ is\ not\ satisfied\end{cases}$$

$$p(x|\mu\theta) = e^{\ln\left(\begin{bmatrix}\theta \\ (1-\theta)\end{bmatrix}[\mathbb{I}(x=1),\mathbb{I}(x=0)]\right)}$$

Identifying terms:

$$g(\bar{\boldsymbol{\theta}}) = 1$$
$$h(\bar{\boldsymbol{X}}) = 1$$
$$\phi(\bar{\boldsymbol{X}}) = [\mathbb{I}(x=1), \mathbb{I}(x=0)]$$
$$\eta(\bar{\boldsymbol{\theta}})^{\boldsymbol{T}} = \begin{bmatrix}\theta \\ (1-\theta)\end{bmatrix}$$

3.2.1.2 *Multidimensional Bernoulli:*

Having of a set of D independent Bernoulli distributions:

$$\bar{\boldsymbol{X}} = \{X_1, \dots X_N\} \quad X_i \sim Bern(x|\theta_i)$$

The joint distribution of the Vector is

$$p(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}) = p(X_1, \dots X_N|\bar{\boldsymbol{\theta}}) = \prod_{i=1}^{D} p(X_i|\theta_i) = \prod_{i=1}^{D} \theta_i^{X_i}(1 - \theta_i)^{1-X_i}$$

Expressed as an exponential we have:

$$p(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}) = \prod_{i=1}^{D}(1 - \theta_i)e^{\left(\frac{ln(\theta_i)}{ln(1-\theta_i)}\right)X_i} = \prod_{i=1}^{D}(1 - \theta_i)\left(e^{\sum_{i=1}^{D}\left(\frac{ln(\theta_i)}{ln(1-\theta_i)}\right)X_i}\right)$$

Identifying terms:

$$g(\bar{\boldsymbol{\theta}}) = \prod_{i=1}^{D}(1 - \theta_i)$$
$$h(\bar{\boldsymbol{X}}) = 1$$
$$\phi(\bar{\boldsymbol{X}}) = [\bar{\boldsymbol{X}}] = [X_1, \dots X_N]$$

$$\eta(\bar{\boldsymbol{\theta}})^{\boldsymbol{T}} = \begin{bmatrix} \frac{ln(1)}{ln(1-\theta_1)} \\ \vdots \\ \frac{ln(\theta_D)}{ln(1-\theta_D)} \end{bmatrix}$$

3.2.2 Properties of the Exponential family

Members of the exponential family have many important properties in common, in the next few pages we will discuss some of them that will be helpful to know.
For now on we are going to assume$\overline{\boldsymbol{\theta}} = \eta(\overline{\boldsymbol{\theta}})$**.** Since the objective of using $\eta(\overline{\boldsymbol{\theta}})$ is just so that $\overline{\boldsymbol{\theta}}$ looks nice in the definition of the distribution.

$$p(X|\overline{\boldsymbol{\theta}}) = \frac{1}{Z(\overline{\boldsymbol{\theta}})}\, h(X) e^{\left(\overline{\boldsymbol{\theta}}^{\boldsymbol{T}} \cdot \phi(X)\right)}$$

3.2.2.1 The Exponential Family: IID observations

Let $p(X|\overline{\boldsymbol{\theta}})$ be the p.d.f. of an exponential family over the random variable X and natural parameters $\overline{\boldsymbol{\theta}} = [\theta_1, \theta_2 \dots, \theta_S]$.

$$p(X|\overline{\boldsymbol{\theta}}) = \frac{1}{Z(\overline{\boldsymbol{\theta}})}\, h(X) e^{\left(\eta(\overline{\boldsymbol{\theta}})^{\boldsymbol{T}} \cdot \phi(X)\right)}$$

Let $\mathcal{D} = \{X_1, \dots X_N\}$ be a set of N IID samples drawn from the distribution.
The **likelihood o**f $\mathcal{D}$ is the probability of getting this vector $\mathcal{D}$ of values, given the parameters $\overline{\boldsymbol{\theta}}$

$$Likelihood(\mathcal{D}, \overline{\boldsymbol{\theta}}) = \mathcal{L}(\mathcal{D}, \overline{\boldsymbol{\theta}}) = p(\mathcal{D}|\overline{\boldsymbol{\theta}})$$

These N realizations of the random variable X, have the same statistical properties of $p(X|\overline{\boldsymbol{\theta}})$ The probability of getting the set $\mathcal{D}$ is:

$$p(\mathcal{D}|\overline{\boldsymbol{\theta}}) = p(X_1, \dots X_N|\overline{\boldsymbol{\theta}}) = \prod_{i=1}^{N} p(X = X_i|\overline{\boldsymbol{\theta}}) \quad \left(Law\ of\ total\ probability\ (X\ are\ IID)\right)$$

Moreover we have:

$$p(\mathcal{D}|\overline{\boldsymbol{\theta}}) = \frac{1}{Z(\overline{\boldsymbol{\theta}})^N}\left(\prod_{i=1}^{N} h(X_i)\right) e^{\left(\eta(\overline{\boldsymbol{\theta}})^{\boldsymbol{T}} \cdot \left(\sum_{i=1}^{N} \phi(X_i)\right)\right)}$$

Using the following equations:

$$H(\mathcal{D}) \triangleq \prod_{i=1}^{N} h(X_i) \qquad \phi(\mathcal{D}) \triangleq \sum_{i=1}^{N} \phi(X_i)$$

We can express the **likelihood** as:

$$p(\mathcal{D}|\overline{\boldsymbol{\theta}}) = \frac{1}{Z(\overline{\boldsymbol{\theta}})^N} H(\mathcal{D})\, e^{\left(\overline{\boldsymbol{\theta}}^{\mathrm{T}} \cdot \phi(\mathcal{D})\right)}$$

And the **log-likelihood** as:

$$\ln\left(p(\mathcal{D}|\overline{\boldsymbol{\theta}})\right) = -NA(\overline{\boldsymbol{\theta}}) + \sum_{i=1}^{N} \ln\left(h(X_i)\right) + \overline{\boldsymbol{\theta}}^{\mathrm{T}} \phi(\mathcal{D})$$

3.2.2.2 Maximum likelihood

In this topic we are going to calculate the parameter vector $\overline{\boldsymbol{\theta}}$ that maximizes the likelihood of the exponential family.

$$\overline{\boldsymbol{\theta}}_{\boldsymbol{ML}} = \boldsymbol{arg}\ \max_{\overline{\boldsymbol{\theta}}} \mathcal{L}(\mathcal{D}, \overline{\boldsymbol{\theta}}) = p(\mathcal{D}|\overline{\boldsymbol{\theta}})$$

Since $p(\mathcal{D}|\overline{\boldsymbol{\theta}})$ is a concave function (hyper-paraboloid) over $\overline{\boldsymbol{\theta}}$, to get we $\overline{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ compute the first derivative and equal to 0.

$$\frac{\partial p(\mathcal{D}|\overline{\boldsymbol{\theta}})}{\partial \overline{\boldsymbol{\theta}}} = -N \frac{\partial A(\overline{\boldsymbol{\theta}})}{\partial \overline{\boldsymbol{\theta}}} + \phi(\mathcal{D})$$

The first element is equal to:

$$\frac{\partial A(\overline{\boldsymbol{\theta}})}{\partial \overline{\boldsymbol{\theta}}} = \frac{\partial}{\partial \overline{\boldsymbol{\theta}}}\left[\ln\left(\int_{S_{\overline{x}}} h(\overline{\boldsymbol{X}}) e^{\left(\overline{\boldsymbol{\theta}}^T \cdot \phi(\overline{\boldsymbol{X}})\right)} d\overline{\boldsymbol{X}}\right)\right]$$

$$= \left(\frac{\partial}{\partial \overline{\boldsymbol{\theta}}} \int_{S_{\overline{x}}} h(\overline{\boldsymbol{X}}) e^{\left(\overline{\boldsymbol{\theta}}^T \cdot \phi(\overline{\boldsymbol{X}})\right)} d\overline{\boldsymbol{X}}\right) \frac{1}{\int_{S_{\overline{x}}} h(\overline{\boldsymbol{X}}) e^{\left(\overline{\boldsymbol{\theta}}^T \cdot \phi(\overline{\boldsymbol{X}})\right)} d\overline{\boldsymbol{X}}}$$

Due to the Property: $\frac{\partial \ln\left(Z(\overline{\boldsymbol{\theta}})\right)}{\partial \overline{\boldsymbol{\theta}}} = \left(\frac{\partial Z(\overline{\boldsymbol{\theta}})}{\partial \overline{\boldsymbol{\theta}}}\right) \frac{1}{Z(\overline{\boldsymbol{\theta}})}$

Moreover we have:

$$\frac{\partial A(\overline{\boldsymbol{\theta}})}{\partial \overline{\boldsymbol{\theta}}} = \frac{\int_{S_{\overline{x}}} \phi(\overline{\boldsymbol{X}}) h(\overline{\boldsymbol{X}}) e^{\left(\overline{\boldsymbol{\theta}}^T \cdot \phi(\overline{\boldsymbol{X}})\right)} d\overline{\boldsymbol{X}}}{e^{A(\overline{\boldsymbol{\theta}})}} = \int_{S_{\overline{x}}} \phi(\overline{\boldsymbol{X}}) h(\overline{\boldsymbol{X}}) e^{\left(\overline{\boldsymbol{\theta}}^T \cdot \phi(\overline{\boldsymbol{X}}) - A(\overline{\boldsymbol{\theta}})\right)} d\overline{\boldsymbol{X}} = \int_{S_{\overline{x}}} \phi(\overline{\boldsymbol{X}}) p(\mathcal{D}|\overline{\boldsymbol{\theta}}) d\overline{\boldsymbol{X}}$$

$$= E\{\phi(\overline{\boldsymbol{X}})\}$$

So we have:

$$\frac{\partial p(\mathcal{D}|\overline{\boldsymbol{\theta}})}{\partial \overline{\boldsymbol{\theta}}}|_{\overline{\boldsymbol{\theta}}=\overline{\boldsymbol{\theta}}_{ML}} = 0 = -NE\{\phi(\overline{\boldsymbol{X}})\} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \phi(X_i)$$

We get the following property called **moment matching**:

$$\boldsymbol{E}\{\boldsymbol{\phi}(\overline{\boldsymbol{X}})\} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\phi}(\boldsymbol{X_i})$$

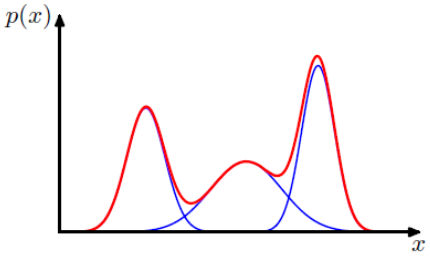
$\frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \phi(X_i)$ is the **sufficient statistic** of the MLE estimator $\overline{\boldsymbol{\theta}}_{\boldsymbol{ML}}$, dataset $\mathcal{D}$ and probability distribution $p(X|\overline{\boldsymbol{\theta}})$ since everything we need to know about $\mathcal{D}$ to estimate $\overline{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ is $\frac{1}{N} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \phi(X_i)$
We do not need to store the entire dataset itself but only the value of the sufficient statistic

To get the estimators $\overline{\boldsymbol{\theta}}_{\boldsymbol{ML}} = \left[\theta_{ML_1}, \theta_{ML_2}, \ldots, \theta_{ML_S}\right]$ of the parameters $\overline{\boldsymbol{\theta}}$ we just have to find momentums $\boldsymbol{E}\{\boldsymbol{\phi}(\overline{\boldsymbol{X}})\}$ of $p(X|\overline{\boldsymbol{\theta}})$ that depend on these parameters and use the moment maching equation to compute their value from the data set $\mathcal{D}$.

3.3 Mixture models

A mixture model is a way of creating complex probability density functions by combination of other simpler p.d.f. Let's say we have K different probability functions $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$ of the same R.V. vector $\overline{\boldsymbol{X}}$, each $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$ can be totally different and have different parameter vector $\overline{\boldsymbol{\theta}}_k$.

We can create a probability distribution $p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}})$ which is a linear combination of these N basic p.d.f. $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$.



$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \pi_k p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}_k})$$

Where:

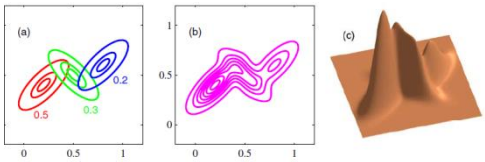
π_k : Is the probability of drawing a sample from the p.d.f. $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}_k})$.

$p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}_k})$: Is the p.d.f. of the i-th basic random distribution.

$\overline{\boldsymbol{\theta}_k}$: Is the parameter vector of the K-th component. $\overline{\boldsymbol{\theta}_k} = [\boldsymbol{\theta}_{k1}, \boldsymbol{\theta}_{k2}, \dots, \boldsymbol{\theta}_{kS}]$

For $p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}})$ to be normalized, it must be satisfied that:

$$\sum_{k=1}^{K} \pi_k = 1 \qquad \textit{Also we have } \; 0 \leq \pi_k \leq 1$$



The $\overline{\boldsymbol{\pi}}_K = [\pi_1, \pi_2, \dots, \pi_K]$ values are called the *mixing coefficients* and the $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}_k})$ distributions are called the *components* of the mixture model.

Since $p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}})$ represents the mixture model, $\overline{\boldsymbol{\theta}}$ represents all the parameters of it, this includes the mixing coefficients $\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}$ and all the vector parameters $\overline{\boldsymbol{\theta}}_k$ from the K distributions.

$$\overline{\boldsymbol{\theta}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}, \overline{\overline{\boldsymbol{\theta}}}_{KxS}\}$$

As we can see, the p.d.f. of the mixture model, is a linear combination of the p.d.f. of the basic distributions $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$. We can generate almost any p.d.f. we want by a mixture model of gaussians, $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k) = \mathcal{N}(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\pi}}_k, \overline{\overline{\boldsymbol{\Sigma}}}_k)$

To **generate data** from the mixture model distribution, what we would do is:

1) Select at random one of the *components* of the mixture according to the discrete probability distribution of the mixing coefficients. We have:

$$P\left(p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) = p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)\right) = \pi_k$$

2) Draw a sample from the distribution $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$

To define a mixture model, we must know all of its parameters $\overline{\boldsymbol{\theta}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}, \overline{\overline{\boldsymbol{\theta}}}_{KxS}\}$. If we were to estimate a mixture model from a dataset, we have to estimate both $\{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}, \overline{\overline{\boldsymbol{\theta}}}_{KxS}\}$ parameters.

3.3.1 Likelihood of a mixture model

Let $\mathcal{D} = \{\bar{\boldsymbol{X}}_1, \dots \bar{\boldsymbol{X}}_N\}$ be a set of N IID samples drawn from a mixture model $p(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}})$, where $\bar{\boldsymbol{X}} = [X_1, X_2, \dots, X_D]$ is the D-dimensional Random Vector of the mixture.

We want to be able to estimate the parameters of the mixture model $\{\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}, \bar{\bar{\boldsymbol{\theta}}}_{KxS}\}$ from dataset $\mathcal{D}$.

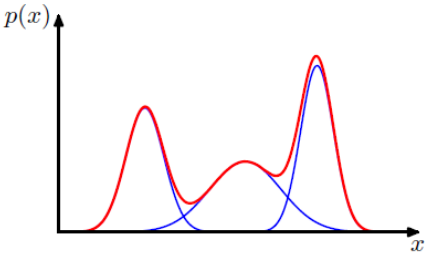
$$\bar{\boldsymbol{\pi}}_{\boldsymbol{K}} = [\pi_1, \pi_2, \dots, \pi_K] \qquad \bar{\bar{\boldsymbol{\theta}}}_{KxS} = \begin{bmatrix} \bar{\boldsymbol{\theta}}_1 \\ \vdots \\ \bar{\boldsymbol{\theta}}_K \end{bmatrix} = \begin{bmatrix} \theta_{11} & \cdots & \theta_{1S} \\ \vdots & \ddots & \vdots \\ \theta_{K1} & \cdots & \theta_{KS} \end{bmatrix}$$

The **likelihood o**f $\mathcal{D}$ is the probability of getting this vector $\mathcal{D}$ of values, given the parameters $\bar{\boldsymbol{\theta}} = \{\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}, \bar{\bar{\boldsymbol{\theta}}}_{KxS}\}$

$$Likelihood(\mathcal{D}, \bar{\boldsymbol{\theta}}) = \mathcal{L}(\mathcal{D}, \bar{\boldsymbol{\theta}}) = p(\mathcal{D}|\bar{\boldsymbol{\theta}})$$

The **Log Likelihood** of the mixture model is:

$$\ln\left(p(D|\bar{\boldsymbol{\theta}})\right) = \sum_{i=1}^{N} ln\left(p(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i|\bar{\boldsymbol{\theta}})\right) = \sum_{i=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k p_k(\bar{\boldsymbol{X}}_i|\bar{\boldsymbol{\theta}}_k)\right)$$



The MAP and ML estimates are non-convex so we cannot use the derivation property

$\frac{\partial \ln\left(p(D|\bar{\boldsymbol{\theta}})\right)}{\partial \bar{\boldsymbol{\theta}}} = 0$ to obtain $\bar{\boldsymbol{\theta}}_{\boldsymbol{ML}}$

- **We have the following problem:**
  We have to estimate $\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}$ and $\bar{\bar{\boldsymbol{\theta}}}_{KxS}$ based on the data samples $= \{\bar{\boldsymbol{X}}_1, \dots \bar{\boldsymbol{X}}_N\}$. Every sample $\bar{\boldsymbol{X}}_1$ has been actually generated from one of the K components of the mixture but we don't know which one directly from $\bar{\boldsymbol{X}}_1$.

- **If we knew the component *k* that generated any of the samples:**
  Every sample $\bar{\boldsymbol{X}}_1$ would have an associated label $\mathbf{Z}$ that indicates which component generated it. We would have couples $\{\bar{\boldsymbol{X}}_i, \boldsymbol{Z}_i\}$ where $\mathbf{Z}$ can have K different values, one per component. And the probability that any sample $\bar{\boldsymbol{X}}_i$ has the label $\boldsymbol{Z}_i = \boldsymbol{k}$ is the probability of drawing a sample from component k:

  $$\boldsymbol{P}(\boldsymbol{Z}_i = \boldsymbol{k}) = \pi_k$$

  In this case we could estimate the parameters of the components, $\bar{\boldsymbol{\theta}}_{\boldsymbol{k}}$, individually, just computing the likelihood to those samples

The $\overline{\boldsymbol{Z}}$ vector variable will stand for the ramdomness of $\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}$ in the mixture model parameters $\overline{\boldsymbol{\theta}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{k}}, \overline{\overline{\boldsymbol{\theta}}}_{KxS}\}$. So we have that $p(\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}_{\boldsymbol{k}} | \overline{\boldsymbol{\theta}}) = \pi_k$

We call $\ln\left(p(D|\overline{\boldsymbol{\theta}})\right) = \ln\left(p\left(D|\{\overline{\boldsymbol{\pi}}_{\boldsymbol{k}}, \overline{\overline{\boldsymbol{\theta}}}_{KxS}\}\right)\right)$ the **incomplete data log likelihood** because $\overline{\boldsymbol{Z}}$ are latent (hidden) variables.

We will be interested in obtaining the dataset D and $\overline{\boldsymbol{Z}}$ from $p(D, \overline{\boldsymbol{Z}} \,|\, \overline{\boldsymbol{\theta}})$ we will be able to calculate the **complete data log likelihood.**

3.3.2 Modeling the mixture model using a latent R.V. $\overline{\boldsymbol{Z}}$

We can model the mixture model in terms of the K different components $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$ and a Discrete Random Vector $\overline{\boldsymbol{Z}}$ whose job is to randomly select the component $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$ from which we draw a given sample according to the *mixing coefficients* $\boldsymbol{\pi}_k$

$\overline{\boldsymbol{Z}}$ is a random binary vector of dimension K

$$\overline{\boldsymbol{Z}} = \{0,1\}^K = [z_1, z_2, \dots, z_K]$$

It's binary components z_i are used to select the component $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$ that is going to be used at each sampling of the mixture model.

- Only one of the binary components of $\overline{\boldsymbol{Z}}$ can be set to 1, the rest is set to 0, this component corresponds to the selected $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$ distribution.

The support of $\overline{\boldsymbol{Z}}$ is:

$$\overline{\boldsymbol{Z}} \,\boldsymbol{\epsilon}\, [\overline{\boldsymbol{Z}}_{\boldsymbol{1}}, \overline{\boldsymbol{Z}}_{\boldsymbol{2}}, \dots, \overline{\boldsymbol{Z}}_{\boldsymbol{K}}]$$

Where $\overline{\boldsymbol{Z}}_k$ is a binary vector with all K components set to 0, and the k-th component set to 1.

$$\overline{\boldsymbol{Z}}_k = \left[0,0,\dots,\underbrace{1}_{k-th},\dots,0\right]$$

So, the random vector $\overline{\boldsymbol{Z}}$ can only take K values and the probability that $\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}_k$ is the probability of drawing a sample from the K component, it is given by the mixing coefficients:

$$P(\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}_k) = P(z_k = 1) = \pi_k$$

Once, we know the value of $\overline{\boldsymbol{Z}}$, let's say $\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}_k$, the sample will be drawn from the distribution $p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$. Expressing this in a conditional probability form:

$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}_k, \overline{\boldsymbol{\theta}}) = p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)$$

The joint probability $\bar{X}$ and $\bar{Z}$ can be seen as:

> $p(\bar{X}, \bar{Z}|\bar{\theta})$: Probability of drawing the sample $\bar{X}$ when it is drawn from the distribution selected by $\bar{Z}$. That is, the probability of getting $\bar{X}$ for the component selected by $\bar{Z}$

Using the Law of Total probability:

$$p(\bar{X}, \bar{Z}|\bar{\theta}) = p(\bar{X}|\bar{Z}, \bar{\theta})P(\bar{Z}|\bar{\theta})$$

We can express the mixture model $p(\bar{X}|\bar{\theta})$ in terms of the $\bar{X}$ and $\bar{Z}$ as the marginal probability of $\bar{X}$ over the joint distribution $p(\bar{X}, \bar{Z}|\bar{\theta})$.

$$p(\bar{X}|\bar{\theta}) = \sum_{Z} p(\bar{X}, \bar{Z}|\bar{\theta}) = \sum_{Z} p(\bar{X}|\bar{Z}, \bar{\theta})P(\bar{Z}|\bar{\theta})$$

Since $\bar{P}(\bar{Z}|\bar{\theta}) = P(\bar{Z}) \quad (With\ \bar{\pi}_k\ set\ by\ \bar{\theta} = \{\bar{\pi}_k, \bar{\bar{\theta}}_{KxS}\})$

$$p(\bar{X}|\bar{\theta}) = \sum_{Z} p(\bar{X}|\bar{Z}, \bar{\theta})P(\bar{Z}) = \sum_{k=1}^{K} p(\bar{X}|\bar{Z} = \bar{Z}_k, \bar{\theta})P(\bar{Z} = \bar{Z}_k)$$

Since:

$$p(\bar{X}|\bar{Z} = \bar{Z}_k, \bar{\theta}) = p_k(\bar{X}|\bar{\theta}_k) \quad and \quad P(\bar{Z} = \bar{Z}_k) = \bar{\pi}_k$$

We arrive to the initial expression:

$$p(\bar{X}|\bar{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\bar{X}|\bar{\theta_k}) = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_K \end{bmatrix} \begin{bmatrix} p_1(\bar{X}|\bar{\theta}_1) & \cdots & p_K(\bar{X}|\bar{\theta}_K) \end{bmatrix} = \bar{\pi}_K \cdot \bar{p}_K(\bar{X}|\bar{\theta}_K)$$

Since we have represented $p(\bar{X}|\bar{\theta})$ as the marginal distribution in the form:

$$p(\bar{X}|\bar{\theta}) = \sum_{Z} p(\bar{X}, \bar{Z}|\bar{\theta})$$

It follows that for every observed data point $\bar{X}_i$ there is a corresponding latent variable $\bar{Z}_k$ that tells us from which component the data point was generated.

We have therefore found an equivalent formulation of the mixture model involving an explicit latent variable $\bar{Z}$. We are now able to work with the joint distribution $p(\bar{X}, \bar{Z}|\bar{\theta})$ instead of the marginal distribution $p(\bar{X}|\bar{\theta})$and this will lead to significant simplifications in the data log likelihood used by the EM algorithm.

- **General expression of $\boldsymbol{p(\overline{X}, \overline{Z}|\overline{\theta})}$**

We already saw that:

$$p(\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}}) = p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{Z}}, \overline{\boldsymbol{\theta}})P(\overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}}) = p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{Z}}, \overline{\boldsymbol{\theta}})P(\overline{\boldsymbol{Z}})$$

We can express the probability function of $\overline{\boldsymbol{Z}}$ as:

$$P(\overline{\boldsymbol{Z}}) = \prod_{k=1}^{K} \pi_i^{\mathbb{I}(\overline{Z} = \overline{Z}_k)} = \prod_{k=1}^{K} \pi_k^{z_k}$$

In the same way, we have:

$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}_k, \overline{\boldsymbol{\theta}}) = p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k) \quad \rightarrow \quad p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{Z}}, \overline{\boldsymbol{\theta}}) = \prod_{k=1}^{K} p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)^{z_k}$$

Why can we express it like this ?

Well, remember $\overline{\boldsymbol{Z}} = [z_1, z_2, \dots, z_K]$ and $\overline{\boldsymbol{Z}}_{\boldsymbol{i}} = \left[0,0, \dots, \underbrace{1}_{k-th}, \dots, 0\right]$

So we can express the joint distribution as:

$$p(\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}}) = p(\overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}})p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{Z}}, \overline{\boldsymbol{\theta}}) = \prod_{k=1}^{K} \left(\pi_k p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)\right)^{\mathbb{I}(\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}_k)} = \prod_{k=1}^{K} \left(\pi_k p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_k)\right)^{z_k}$$

Basically $\overline{\boldsymbol{Z}}$ (its $z_k = 1$ position) selects which of the K components is going to be responsible for generating $\overline{\boldsymbol{X}}_i$.

Let be a set of N IID samples drawn from a mixture model $p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}})$, where $\overline{\boldsymbol{X}} = [X_1, X_2, \dots, X_D]$ is the D-dimensional Random Vector of the mixture.

- **Porpous and advantajes of $\boldsymbol{p(\overline{X}, \overline{Z}|\overline{\theta})}$**

If we have several observations $\mathcal{D} = \{\overline{\boldsymbol{X}}_1, \dots \overline{\boldsymbol{X}}_N\}$ it follows that for every observed data point $\overline{\boldsymbol{X}}_{\boldsymbol{i}}$ there is a corresponding latent variable $\boldsymbol{Z_i}$. $Pairs \rightarrow \{\overline{\boldsymbol{X}}_{\boldsymbol{i}}, \boldsymbol{Z_i}\}$
We have therefore found an equivalent formulation mixture model involving the explicit latent variable $\overline{\boldsymbol{Z}}$.
We are now able to work with the joint distribution $\boldsymbol{p(\overline{X}, \overline{Z}|\overline{\theta})}$ instead of the marginal distribution $p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}})$ and this will lead to significant simplifications, most notably through the introduction of the expectation-maximization (EM) algorithm.

- **Complete log-likelihood**

We define the **Complete data log likelihood** as:

$$
\begin{aligned}
I_C(\bar{\boldsymbol{\theta}}) = \ln\left(p(D,\bar{\boldsymbol{Z}}|\bar{\boldsymbol{\theta}})\right) &= \sum_{i=1}^{N} ln\left(p(\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{\boldsymbol{i}},\bar{\boldsymbol{Z}}=\bar{\boldsymbol{Z}}_{\boldsymbol{i}}|\bar{\boldsymbol{\theta}})\right) \\
&= \sum_{i=1}^{N} ln\left(p(\bar{\boldsymbol{Z}}=\bar{\boldsymbol{Z}}_{\boldsymbol{i}}|\bar{\boldsymbol{\theta}})p(\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{i}|\bar{\boldsymbol{Z}}=\bar{\boldsymbol{Z}}_{\boldsymbol{i}},\bar{\boldsymbol{\theta}})\right) \\
&= \sum_{i=1}^{N} \ln\left(\prod_{\boldsymbol{k=1}}^{\boldsymbol{K}}\left(\pi_{\boldsymbol{k}}p_{\boldsymbol{k}}(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)^{\mathbb{I}(z_i=k)}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i=k)\ln\left(\pi_{\boldsymbol{k}}p_{\boldsymbol{k}}(\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{i}|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i=k)\ln\left(\pi_{\boldsymbol{k}}p_{\boldsymbol{k}}(\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{i}|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)
\end{aligned}
$$

This is the likelihood of the data $\{D,\bar{\boldsymbol{Z}}\}$

- **Responsibility:**

Another quantity that will play an important role is $p(\bar{\boldsymbol{Z}}|\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{i},\bar{\boldsymbol{\theta}})$, the conditional probability of $\bar{\boldsymbol{Z}}$ given $\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{\boldsymbol{i}}$. Given a sample $\bar{\boldsymbol{X}}_i$ of the mixture model, it tells us what is the probability that it comes from any of the K different distributions.

So $p(\bar{\boldsymbol{Z}}=\bar{\boldsymbol{Z}}_{\boldsymbol{k}}|\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{i},\bar{\boldsymbol{\theta}})$ indicates what is the probability that the sample $\bar{\boldsymbol{X}}_i$ belongs to the k-th component, $p_k(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}_k)$, of the mixture model.

By Bayes Theorem we have:

$$
p(\bar{\boldsymbol{Z}}=\bar{\boldsymbol{Z}}_{\boldsymbol{k}}|\bar{\boldsymbol{X}},\bar{\boldsymbol{\theta}}) = \frac{\overbrace{p(\bar{\boldsymbol{X}}|\bar{\boldsymbol{Z}}=\bar{\boldsymbol{Z}}_{\boldsymbol{k}},\bar{\boldsymbol{\theta}})}^{p_k(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})}\overbrace{p(\bar{\boldsymbol{Z}}=\bar{\boldsymbol{Z}}_{\boldsymbol{k}}|\bar{\boldsymbol{\theta}})}^{\pi_{\boldsymbol{k}}}}{p(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}})} = \frac{\pi_{\boldsymbol{k}}p_{\boldsymbol{k}}(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})}{\sum_{\boldsymbol{j=1}}^{K}\pi_{\boldsymbol{j}}p_{\boldsymbol{j}}(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}_{\boldsymbol{j}})}
$$

We define the ***responsibility*** of the K-component to $\bar{\boldsymbol{X}}_i$

$$
r_{ik} = p(\bar{\boldsymbol{Z}}=\bar{\boldsymbol{Z}}_{\boldsymbol{k}}|\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{i},\bar{\boldsymbol{\theta}}) = E\{\mathbb{I}(z_i=k)|\bar{\boldsymbol{X}}_i\} = \frac{\pi_{\boldsymbol{k}}p_{\boldsymbol{k}}(\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{i}|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})}{\sum_{\boldsymbol{j=1}}^{K}\pi_{\boldsymbol{j}}p_{\boldsymbol{j}}(\bar{\boldsymbol{X}}=\bar{\boldsymbol{X}}_{i}|\bar{\boldsymbol{\theta}}_{\boldsymbol{j}})}
$$

The responsibility r_{ik} can be viewed as the responsibility that component K takes for 'explaining' the observation $\bar{\boldsymbol{X}}_i$. In other words, it gives us a measure of how likely is $\bar{\boldsymbol{X}}_i$ to belong to $p_k(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}_k)$.

r_{ik} is also called the posterior probability, of component k given data point X_i.

For every sample $\bar{X}_i$ we will have a responsibility vector $\bar{r}_i = [r_1, \dots, r_k, \dots, r_K]$ that tells us, how likely is that sample to belong to any of the K components. It actually gives us the probability that $\bar{X}_i$ belongs to any of the K components.

$\bar{r}_i$ is an estimator of the $\bar{Z}_i$ random vector.

- While $\bar{Z}_i$ a is binary vector 1-toK that tells us which is the component generated $\bar{X}_i$.
- $\bar{r}_i$ tells us, for every component k, what is the probability that it generated from that component

$$\bar{r}_i = [r_1, \dots, r_k, \dots, r_K] \quad \bar{Z}_i = [z_1, \dots, z_k, \dots, z_K]$$

It follows that:

$$\sum_{k=1}^{K} r_k = 1 \qquad \sum_{k=1}^{K} z_k = 1$$

- **Incomplete log-likelihood, complete log-likelihood and responsibility**

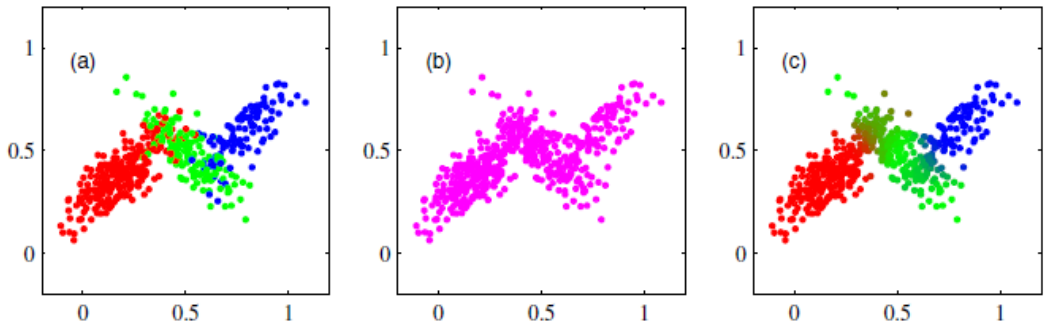


Figure 9.5 Example of 500 points drawn from the mixture of 3 Gaussians shown in Figure 2.23. (a) Samples from the joint distribution $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ in which the three states of $\mathbf{z}$, corresponding to the three components of the mixture, are depicted in red, green, and blue, and (b) the corresponding samples from the marginal distribution $p(\mathbf{x})$, which is obtained by simply ignoring the values of $\mathbf{z}$ and just plotting the $\mathbf{x}$ values. The data set in (a) is said to be *complete*, whereas that in (b) is *incomplete*. (c) The same samples in which the colours represent the value of the responsibilities $\gamma(z_{nk})$ associated with data point $\mathbf{x}_n$, obtained by plotting the corresponding point using proportions of red, blue, and green ink given by $\gamma(z_{nk})$ for $k = 1, 2, 3$, respectively

3.4 EM algorithm

- Expected complete data log likelihood (Expectation step or E step)

$$Q(\theta, \theta^{t-1}) = E\{l_c(\theta)|\mathcal{D}, \theta^{t-1}\}$$

- Maximization step (M step)
  - ML Estimation

$$\theta^t = \underset{\theta}{\operatorname{argmax}} Q(\theta, \theta^{t-1})$$

  - MAP Estimation

$$\theta^t = \underset{\theta}{\operatorname{argmax}} Q(\theta, \theta^{t-1}) + \ln p(\theta)$$

- Convergence: $l_c(\theta^t) \geq l_c(\theta^{t-1})$

3.4.1 Expectation Step

In this step we just calculate the responsibility vector of each component k, to each sample i.

$$r_{ik} = \frac{\pi_k p_k(\bar{X} = \bar{X}_i|\bar{\theta}_k)}{\sum_{j=1}^{K} \pi_j p_j(\bar{X} = \bar{X}_i|\bar{\theta}_j)}$$

Where X_{ij} is the binary value of $j-th$ component of the $i-th$ sample $\bar{X}_i = [X_1, .., X_j, .., X_D]$

This is equal to compute the matrix:

$$r = \begin{bmatrix} \bar{r}_1 \\ \bar{r}_2 \\ \vdots \\ \bar{r}_N \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1K} \\ r_{21} & r_{22} & \cdots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NK} \end{bmatrix} \quad Remember \sum_{k=1}^{K} r_{ik} = 1 \ (Rows\ sum\ 1)$$

In the algorithm there is no need for calculating $\sum_{j=1}^{K} \pi_j p_j(\bar{X} = \bar{X}_i|\bar{\theta}_j)$ since this is a normalization factor for a given observation $\bar{X}_i$.
Due to the fact that:

$$\sum_{k=1}^{K} r_{ik} = 1$$

We just need to calculate all K values $\pi_k p_k(\bar{X} = \bar{X}_i|\bar{\theta}_k)$ for a given $\bar{X}_i$ and normalize.

3.4.2 Maximization Step

In the maximization step we will estimate the new model $\bar{\boldsymbol{\theta}} = \{\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}, \bar{\bar{\boldsymbol{\theta}}}_{KxD}\}^{t}$ using the past parameters $\bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}} = \{\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}, \bar{\bar{\boldsymbol{\theta}}}_{KxD}\}^{t-1}$ and the updated r_{ik} values from the E-step.

The new parameters $\bar{\boldsymbol{\theta}}^{\boldsymbol{t}}$ are those that maximize $Q(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$ over all possible values of $\bar{\boldsymbol{\theta}}$

$$\bar{\boldsymbol{\theta}}^{\boldsymbol{t}} = \boldsymbol{arg}\ \max_{\bar{\boldsymbol{\theta}}} Q(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$$

As calculated before, the complete log-likelihood is expressed as:

$$I_C(\bar{\boldsymbol{\theta}}) = \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k) \ln\left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i | \bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)$$

The Expected **complete data log-likelihood** is:

$$\begin{aligned} Q(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}}) &= E\{I_C(\bar{\boldsymbol{\theta}}) | D, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}}\} = E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k) \ln\left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i | \bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)\right\} \\ &= E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k | \bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_{\boldsymbol{i}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}}) \ln\left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i | \bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)\right\} \\ &= \sum_{k=1}^{K}\sum_{i=1}^{N} E\{\mathbb{I}(z_i = k | \bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_{\boldsymbol{i}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}})\} \ln\left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i | \bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right) \end{aligned}$$

Since:

$$E\{\mathbb{I}(z_i = k | \bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_{\boldsymbol{i}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}})\} = p(\bar{\boldsymbol{Z}} = \bar{\boldsymbol{Z}}_{\boldsymbol{k}} | \bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}}) = r_{ik}$$

Notice r_{ik} is calculated with the $\bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}} \rightarrow \{\bar{\boldsymbol{\pi}}_{\boldsymbol{k}}, \bar{\bar{\boldsymbol{\theta}}}_{KxS}\}^{t-1}$ parameters

We have that:

$$\begin{aligned} Q(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}}) &= \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik} \ln\left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i | \bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right) \\ &= \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik} \ln(\pi_{\boldsymbol{k}})}_{Depends\ on\ \bar{\pi}_k} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik} \ln\left(p_{\boldsymbol{k}}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i | \bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)}_{Depends\ on\ \bar{\bar{\boldsymbol{\theta}}}_{\boldsymbol{KxD}}} \end{aligned}$$

So we need to maximize this over $\bar{\boldsymbol{\theta}} = \{\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}, \bar{\bar{\boldsymbol{\theta}}}_{KxD}\}$.
Since $Q(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$ is a concave function over $\bar{\boldsymbol{\theta}}$, we could perform the derivative with respect to $\bar{\boldsymbol{\theta}}$ and set it to 0, to get the $\bar{\boldsymbol{\theta}}$ that maximizes the function $Q(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$

$$\frac{\partial Q(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \bar{\boldsymbol{\theta}}}\Big|_{\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}^{\boldsymbol{t}}} = 0$$

3.4.2.1 *Value of the mixing coefficients*

We are going to calculate the value of the mixing coefficients $\bar{\pi}_k$ that maximizes $Q(\bar{\theta}, \bar{\theta}^{t-1})$ to obtain the next value of these coefficients $\bar{\pi}_k^t$:

$$\frac{\partial Q(\bar{\theta}, \bar{\theta}^{t-1})}{\partial \bar{\pi}_k} |_{\bar{\pi}_k = \bar{\pi}_k^t} = 0$$

Since we are subject to the **constraint** that:

$$\sum_{k=1}^{K} \pi_k = 1$$

We add a **Lagrange multiplier** with that constraint so we must solve:

$$\frac{\partial}{\partial \bar{\pi}_k} \left(\underbrace{\sum_{k=1}^{K} \sum_{i=1}^{N} r_{ik} \ln(\pi_k)}_{Depends\ on\ \bar{\pi}_k} + \underbrace{\lambda \left(\sum_{k=1}^{K} \pi_k = 1 \right)}_{Lagrange\ multiplier} \right) = 0$$

Taking partial derivatives with respect to the elements of $\bar{\pi}_K$ we get, for the element k, of $\bar{\pi}_K$**:**

$$\frac{1}{\pi_k} \sum_{i=1}^{N} r_{ik} + \lambda = 0$$

We have:

$$\pi_k = -\frac{\sum_{i=1}^{N} r_{ik}}{\lambda} = -\frac{N_k}{\lambda} \qquad with \quad N_k = \sum_{i=1}^{N} r_{ik}$$

N_k : Can be interpreted as the effective number of points assigned to component k.

Since we have the constraint:

$$\sum_{k=1}^{K} \pi_k = 1 \quad \rightarrow \quad \sum_{k=1}^{K} -\frac{N_k}{\lambda} = 1 \quad \rightarrow \quad \boldsymbol{\lambda = -\sum_{k=1}^{K} N_k = -N}$$

So we finally have:

$$\boldsymbol{\pi_k = \frac{N_k}{N} = \frac{1}{N} \sum_{i=1}^{N} r_{ik}}$$

3.4.2.2 Value of the parameters of the distributions

We are going to calculate the value of the parameter vector $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$ of the k-component that maximizes $Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$ to obtain the next value of these coefficients $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$:

$$\frac{\partial Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}} |_{\overline{\boldsymbol{\theta}}_{\boldsymbol{k}} = \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\boldsymbol{t}}} = 0$$

Remember: $\overline{\boldsymbol{\theta}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}, \overline{\overline{\boldsymbol{\theta}}}_{KxD}\}$ and $\overline{\overline{\boldsymbol{\theta}}}_{KxD} = [\overline{\boldsymbol{\theta}}_{\boldsymbol{1}}, \dots, \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}, \dots, \overline{\boldsymbol{\theta}}_{\boldsymbol{K}}]$ so we are trying to get the value of one of there $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$ vectors.

$$\frac{\partial Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}} |_{\overline{\boldsymbol{\theta}}_{\boldsymbol{k}} = \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\boldsymbol{t}}} = \frac{\partial}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}} \left(\underbrace{\sum_{i=1}^{N} r_{ik} \ln \left(p_{\boldsymbol{k}}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{X}}_i | \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}) \right)}_{Depends\ on\ \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}} \right) = 0$$

So, to get $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\boldsymbol{t}}$ we must solve the equation:

$$\sum_{i=1}^{N} r_{ik} \left[\frac{\partial}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}} \left(\ln \left(p_{\boldsymbol{k}}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{X}}_i | \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}) \right) \right) \right] = 0$$

3.4.3 EM Algorithm for the Exponential Family

The exponential family has a p.d.f. with the form:

$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) = h(\overline{\boldsymbol{X}})e^{\left(\overline{\boldsymbol{\theta}}^{\boldsymbol{T}}\cdot\phi(\overline{\boldsymbol{X}})-A(\overline{\boldsymbol{\theta}})\right)}$$

In a **mixture model** of this family, the k-th component will therefore have the p.d.f.:

$$p_k(\overline{\boldsymbol{X}}\,|\overline{\boldsymbol{\theta}}_k) = h_k(\overline{\boldsymbol{X}})e^{\left(\overline{\boldsymbol{\theta}_k}^{\boldsymbol{T}}\cdot\phi_k(\overline{\boldsymbol{X}})-A_k(\overline{\boldsymbol{\theta}}_k)\right)}$$

The joint probability distribution $p(\overline{\boldsymbol{X}},\overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}})$ of the mixture model can be expressed as:

$$p(\overline{\boldsymbol{X}},\overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}}) = \prod_{\boldsymbol{k=1}}^{\boldsymbol{K}}\left(\pi_k h(\overline{\boldsymbol{X}})e^{\left(\overline{\boldsymbol{\theta}}^{\boldsymbol{T}}\cdot\phi(\overline{\boldsymbol{X}})-A(\overline{\boldsymbol{\theta}})\right)}\right)^{\mathbb{I}(\overline{\boldsymbol{Z}}=\overline{\boldsymbol{Z}}_{\boldsymbol{k}})} = \prod_{\boldsymbol{k=1}}^{\boldsymbol{K}}\left(h(\overline{\boldsymbol{X}})e^{\left(\overline{\boldsymbol{\theta}}^{\boldsymbol{T}}\cdot\phi(\overline{\boldsymbol{X}})-A(\overline{\boldsymbol{\theta}})+\boldsymbol{ln}(\pi_k)\right)}\right)^{\mathbb{I}(\overline{\boldsymbol{Z}}=\overline{\boldsymbol{Z}}_{\boldsymbol{k}})}$$

Taking the logarithm:

$$\ln\left(p(\overline{\boldsymbol{X}},\overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}})\right) = \sum_{\boldsymbol{k=1}}^{K}\mathbb{I}(\overline{\boldsymbol{Z}}\ =\ \overline{\boldsymbol{Z}}_{\boldsymbol{k}})\ln\left(h(\overline{\boldsymbol{X}})e^{\left(\overline{\boldsymbol{\theta}}^{\boldsymbol{T}}\cdot\phi(\overline{\boldsymbol{X}})-A(\overline{\boldsymbol{\theta}})+ln(\pi_k)\right)}\right)$$
$$= \sum_{\boldsymbol{k=1}}^{K}\mathbb{I}(\overline{\boldsymbol{Z}}\ =\ \overline{\boldsymbol{Z}}_{\boldsymbol{k}})\left(\left(ln(\pi_k)-A(\overline{\boldsymbol{\theta}})\right)+\left(\ln\left(h(\overline{\boldsymbol{X}})\right)+\ \overline{\boldsymbol{\theta}}^{\boldsymbol{T}}\cdot\phi(\overline{\boldsymbol{X}})\right)\right)$$

The complete data log likelihood of the mixture model is:

$$I_C(\overline{\boldsymbol{\theta}}) = \ln\left(p(D,\overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}})\right)\ = \sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{I}(z_i=k)\ln\left(\pi_k p_k(\overline{\boldsymbol{X}}=\overline{\boldsymbol{X}}_i|\overline{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)$$
$$= \sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{I}(z_i=k)\ln\left(\pi_k h_k(\overline{\boldsymbol{X}})e^{\left(\overline{\boldsymbol{\theta}_k}^{\boldsymbol{T}}\cdot\phi_k(\overline{\boldsymbol{X}})-A_k(\overline{\boldsymbol{\theta}}_k)\right)}\right)$$
$$= \sum_{\boldsymbol{k=1}}^{K}\sum_{\boldsymbol{i=1}}^{N}\mathbb{I}(\overline{\boldsymbol{Z}}\ =\ \overline{\boldsymbol{Z}}_{\boldsymbol{k}})\left(\left(ln(\boldsymbol{\pi_k})-A(\overline{\boldsymbol{\theta}})\right)+\left(\ln\left(h(\overline{\boldsymbol{X}})\right)+\ \overline{\boldsymbol{\theta}}^{\boldsymbol{T}}\cdot\phi(\overline{\boldsymbol{X}})\right)\right)$$

The new parameters $\overline{\boldsymbol{\theta}}^{\boldsymbol{t}}$ are those that maximize $Q\left(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}}\right)$ over all possible values of $\overline{\boldsymbol{\theta}}$

$$\overline{\boldsymbol{\theta}}^{\boldsymbol{t}} = \boldsymbol{arg}\ \max_{\overline{\boldsymbol{\theta}}} Q\left(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}}\right)$$

With:

$$Q\left(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}}\right) = \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N}r_{ik}\ln(\pi_k)}_{Depends\ on\ \overline{\boldsymbol{\pi}}_{\boldsymbol{k}}} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N}r_{ik}\left(H(\mathcal{D})+\ \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\ \boldsymbol{T}}\cdot\phi(\overline{\mathcal{D}})-A(\overline{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)}_{Depends\ on\ \overline{\overline{\boldsymbol{\theta}}}_{\boldsymbol{KxD}}}$$

So we need to maximize this over $\overline{\boldsymbol{\theta}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}},\overline{\overline{\boldsymbol{\theta}}}_{KxD}\}$.

Since $Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$ is a concave function over $\overline{\boldsymbol{\theta}}$, we could perform the derivative with respect to $\overline{\boldsymbol{\theta}}$ and set it to 0, to get the $\overline{\boldsymbol{\theta}}$ that maximizes the function $Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$

$$\frac{\partial Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \overline{\boldsymbol{\theta}}}|_{\overline{\boldsymbol{\theta}}=\overline{\boldsymbol{\theta}}^{\boldsymbol{t}}} = 0$$

3.4.3.1 Value of the mixing coefficients

From the general case, we already know that:

$$\boldsymbol{\pi_k} = \frac{\boldsymbol{N_k}}{\boldsymbol{N}} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{r_{ik}}$$

3.4.3.2 Value of the parameters of the distributions

We are going to calculate the value of the parameter vector $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$ of the k-component that maximizes $Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$ to obtain the next value of these coefficients $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$:

$$\frac{\partial Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}}|_{\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}=\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\boldsymbol{t}}} = 0$$

Remember: $\overline{\boldsymbol{\theta}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}, \overline{\overline{\boldsymbol{\theta}}}_{KxD}\}$ *and* $\overline{\overline{\boldsymbol{\theta}}}_{KxD} = [\overline{\boldsymbol{\theta}}_{\mathbf{1}}, \dots, \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}, \dots, \overline{\boldsymbol{\theta}}_{\boldsymbol{K}}]$ so we are trying to get the value of one of there $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$ vectors.

$$\frac{\partial Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}}|_{\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}=\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\boldsymbol{t}}} = \frac{\partial}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}} \left(\underbrace{\sum_{i=1}^{N} r_{ik} \left(\overline{\boldsymbol{\theta}}_k{}^{\boldsymbol{T}} \cdot \phi(\mathcal{D}) - A(\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}) \right)}_{Depends\ on\ \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}} \right) = 0$$

As we have seen earlier, the moment matching of the exponential family implies:

$$\frac{\partial A(\overline{\boldsymbol{\theta}}_{\boldsymbol{k}})}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}} = E\{\phi(\overline{\boldsymbol{X}})\} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\phi}(\boldsymbol{X_i})$$

We have that:

$$\frac{\partial Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}}|_{\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}=\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\boldsymbol{t}}} = \sum_{i=1}^{N} r_{ik}(\phi(\mathcal{D}) - E\{\phi(\overline{\boldsymbol{X}})\}) = 0$$

3.5 EM Algorithm for the multinomial distribution

Let $\overline{\boldsymbol{X}}$ be a D −dimensional binary random vector that follows a multinomial (or multivariate Bernoullis) distribution:

$$\overline{\boldsymbol{X}} = \{X_1, \dots, X_j, \dots, X_N\} \quad X_j \sim Bern(x|\theta_i)$$

If the Bernoullis are independent of each other, the joint distribution of the $\overline{\boldsymbol{X}}$ is:

$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) = p(X_1, \dots X_N|\overline{\boldsymbol{\theta}}) = \prod_{j=1}^{D} \theta_j^{X_j}(1 - \theta_j)^{1-X_j} = \prod_{j=1}^{D}(1 - \theta_j)\left(e^{\Sigma_{j=1}^{D}\left(\frac{ln(\theta_j)}{ln(1-\theta_j)}\right)X_j}\right)$$

The parameters of the multinomial distribution are the parameters of the binomials:

$$\overline{\boldsymbol{\theta}} = [\bar{\theta}_1, \dots, \bar{\theta}_j, .., \bar{\theta}_D]$$

- Now consider a **mixture model** of different D −dimensional multivariate Bernoullis

$$p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \pi_k p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}_k}) = \sum_{k=1}^{K} \pi_k \underbrace{\left(\prod_{j=1}^{D} \theta_{jk}^{X_j}(1 - \theta_{jk})^{1-X_j}\right)}_{p_k(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}_k})}$$

Where $\overline{\boldsymbol{\theta}_k} = [\bar{\theta}_{k1}, \dots, \bar{\theta}_{kj}, .., \bar{\theta}_{kD}]$ is the parameter vector of the K compoment of the mixture.

$$\overline{\boldsymbol{\theta}} = \{\bar{\pi}_k, \overline{\overline{\boldsymbol{\theta}}}_{KxD}\} \qquad \overline{\overline{\boldsymbol{\theta}}}_{KxD} = \begin{bmatrix} \overline{\boldsymbol{\theta}}_1 \\ \vdots \\ \overline{\boldsymbol{\theta}}_K \end{bmatrix} = \begin{bmatrix} \theta_{11} & \cdots & \theta_{1D} \\ \vdots & \ddots & \vdots \\ \theta_{K1} & \cdots & \theta_{KD} \end{bmatrix}$$

Examples of this kind of data are binary images with D bits.

- Using explicitly the latent variable Z, we get the joint distribution $p(\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}})$ as:

$$p(\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}}) = p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{Z}}, \overline{\boldsymbol{\theta}})P(\overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}}) = p(\overline{\boldsymbol{X}}|\overline{\boldsymbol{Z}}, \overline{\boldsymbol{\theta}})P(\overline{\boldsymbol{Z}}) = \prod_{\boldsymbol{k=1}}^{\boldsymbol{K}} \left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)^{\mathbb{I}(\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}_{\boldsymbol{k}})}$$

$$p(\overline{\boldsymbol{X}}, \overline{\boldsymbol{Z}}|\overline{\boldsymbol{\theta}}) = \prod_{\boldsymbol{k=1}}^{\boldsymbol{K}} \left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\overline{\boldsymbol{X}}|\overline{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)^{z_k} = \prod_{\boldsymbol{k=1}}^{\boldsymbol{K}} \left(\pi_{\boldsymbol{k}} \left(\prod_{j=1}^{D} \theta_{jk}^{X_j}(1 - \theta_{jk})^{1-X_j}\right)\right)^{\mathbb{I}(\overline{\boldsymbol{Z}} = \overline{\boldsymbol{Z}}_{\boldsymbol{k}})}$$

- Complete and incomplete likelihoods:

Let $\mathcal{D} = \{\bar{\boldsymbol{X}}_{\boldsymbol{1}}, \ldots \bar{\boldsymbol{X}}_{\boldsymbol{N}}\}$ be a set of N IID samples drawn from the distribution.
The **likelihood o**f $\mathcal{D}$ is the probability of getting this vector $\mathcal{D}$ of vectors, given the parameters $\bar{\boldsymbol{\theta}}$

$$Likelihood(\mathcal{D}, \bar{\boldsymbol{\theta}}) = \mathcal{L}(\mathcal{D}, \bar{\boldsymbol{\theta}}) = p(\mathcal{D}|\bar{\boldsymbol{\theta}})$$

$$p(\mathcal{D}|\bar{\boldsymbol{\theta}}) = p(X_1, \ldots X_N|\bar{\boldsymbol{\theta}}) = \prod_{i=1}^{N} p(X = X_i|\bar{\boldsymbol{\theta}}) \quad \left(Law\ of\ total\ probability\ (\bar{X}\ are\ IID)\right)$$

$$p(\mathcal{D}|\bar{\boldsymbol{\theta}}) = \prod_{i=1}^{N} \prod_{\boldsymbol{k=1}}^{\boldsymbol{K}} \left(\pi_{\boldsymbol{k}} \left(\prod_{j=1}^{D} \theta_{jk}^{X_{ij}} \left(1 - \theta_{jk}\right)^{1-X_{ij}} \right) \right)^{\mathbb{I}(\bar{\boldsymbol{Z}} = \bar{\boldsymbol{Z}}_{\boldsymbol{k}})}$$

Being X_{ij}: The binary value of $j-th$ component of the $i-th$ sample $\bar{X}_i = \left[X_1, .., X_j, .., X_D\right]$

The **complete likelihood o**f $\mathcal{D}$ is the probability of getting this vector $\mathcal{D}$ of values, given the parameters $\bar{\boldsymbol{\theta}}$

$$Complete\ Likelihood(\{\mathcal{D}, Z\}, \bar{\boldsymbol{\theta}}) = \mathcal{L}(\{\mathcal{D}, Z\}, \bar{\boldsymbol{\theta}}) = p(\mathcal{D}, \bar{\boldsymbol{Z}}|\bar{\boldsymbol{\theta}})$$

$$p(\mathcal{D}|\bar{\boldsymbol{\theta}}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \underbrace{\left(\prod_{j=1}^{D} \theta_{jk}^{X_{ij}} \left(1 - \theta_{jk}\right)^{1-X_{ij}} \right)}_{p_k(\bar{\boldsymbol{X}}|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})}$$

3.5.1 Complete data Log-likelihood

The complete data log likelihood for the multinomial distribution is:

$$\begin{aligned}
I_C(\bar{\boldsymbol{\theta}}) = \ln\left(p(D, \bar{\boldsymbol{Z}}|\bar{\boldsymbol{\theta}})\right) &= \sum_{i=1}^{N} ln\left(p(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_{\boldsymbol{i}}, \bar{\boldsymbol{Z}} = \bar{\boldsymbol{Z}}_{\boldsymbol{i}}|\bar{\boldsymbol{\theta}})\right) \\
&= \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{I}(z_i = k) \ln\left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_i|\bar{\boldsymbol{\theta}}_{\boldsymbol{k}})\right) \\
&= \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{I}(z_i = k) \ln\left(\pi_{\boldsymbol{k}} \prod_{j=1}^{D} \theta_{jk}^{X_{ij}} \left(1 - \theta_{jk}\right)^{1-X_{ij}} \right) \\
&= \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{I}(z_i = k) \left(ln(\pi_k) + \sum_{j=1}^{D} \left(X_{ij} \ln\left(\theta_{jk}\right) + \left(1 - X_{ij}\right) \ln\left(\left(1 - \theta_{jk}\right)\right)\right)\right)
\end{aligned}$$

3.5.2 Expected complete data log-likelihood

The expected complete data log likelihood for the multinomial distribution is:

$$Q(\bar{\theta}, \bar{\theta}^{t-1}) = E\{I_C(\bar{\theta})|D, \bar{\theta}^{t-1}\} = E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k) \ln\left(\pi_k p_k(\bar{X} = \bar{X}_i|\bar{\theta}_k)\right)\right\}$$
$$= E\left\{\sum_{k=1}^{K}\sum_{i=1}^{N} \mathbb{I}(z_i = k|\bar{X} = \bar{X}_i, \bar{\theta}^{t-1}) \ln\left(\pi_k p_k(\bar{X} = \bar{X}_i|\bar{\theta}_k)\right)\right\}$$
$$= \sum_{k=1}^{K}\sum_{i=1}^{N} E\{\mathbb{I}(z_i = k|\bar{X} = \bar{X}_i, \bar{\theta}^{t-1})\} \ln\left(\pi_k p_k(\bar{X} = \bar{X}_i|\bar{\theta}_k)\right)$$

Since:

$$E\{\mathbb{I}(z_i = k|\bar{X} = \bar{X}_i, \bar{\theta}^{t-1})\} = p(\bar{Z} = \bar{Z}_k|\bar{X} = \bar{X}_i, \bar{\theta}^{t-1}) = r_{ik} = \frac{\pi_k p_k(\bar{X} = \bar{X}_i|\bar{\theta}_k)}{\sum_{j=1}^{K} \pi_j p_j(\bar{X} = \bar{X}_i|\bar{\theta}_j)}$$

Notice r_{ik} is calculated with the $\bar{\theta}^{t-1} \rightarrow \{\bar{\pi}_k, \bar{\bar{\theta}}_{KxS}\}^{t-1}$ parameters

We have that:

$$Q(\bar{\theta}, \bar{\theta}^{t-1}) = \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik} \ln\left(\pi_k p_k(\bar{X} = \bar{X}_i|\bar{\theta}_k)\right)$$
$$= \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik} \ln(\pi_k) + \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik} \ln\left(p_k(\bar{X} = \bar{X}_i|\bar{\theta}_k)\right)$$
$$= \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik} \ln(\pi_k)}_{Depends\ on\ \bar{\pi}_k} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N}\left(\sum_{j=1}^{D}\left(X_{ij} \ln(\theta_{jk}) + (1 - X_{ij}) \ln\left((1 - \theta_{jk})\right)\right)\right)}_{Depends\ on\ \bar{\bar{\theta}}_{KxD}}$$

3.5.3 Expectation Step

In this step we just calculate the responsibility vector of each component k, to each sample i.

$$r_{ik} = \frac{\pi_k p_k(\bar{X}|\bar{\theta}_k)}{\sum_{j=1}^{K} \pi_j p_j(\bar{X}|\bar{\theta}_j)} = \frac{\pi_k \left(\prod_{j=1}^{D} \theta_{jm}^{X_{ij}} (1 - \theta_{jm})^{1-X_{ij}}\right)}{\sum_{m=1}^{K} \pi_m \left(\prod_{j=1}^{D} \theta_{jm}^{X_{ij}} (1 - \theta_{jm})^{1-X_{ij}}\right)}$$

Where X_{ij} is the binary value of $j-th$ component of the $i-th$ sample $\bar{X}_i = [X_1, .., X_j, .., X_D]$

This is equal to compute the matrix:

$$r = \begin{bmatrix} \bar{r}_1 \\ \bar{r}_2 \\ \vdots \\ \bar{r}_N \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1K} \\ r_{21} & r_{22} & \cdots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NK} \end{bmatrix} \quad Remember \sum_{k=1}^{K} r_{ik} = 1\ (Rows\ sum\ 1)$$

3.5.4 Maximization Step

In the maximization step we will estimate the new model $\overline{\boldsymbol{\theta}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}, \overline{\overline{\boldsymbol{\theta}}}_{KxD}\}^{t}$ using the past arameters $\overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}, \overline{\overline{\boldsymbol{\theta}}}_{KxD}\}^{t-1}$.

The new parameters $\overline{\boldsymbol{\theta}}^{\boldsymbol{t}}$ are those that maximize $Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$ over all possible values of $\overline{\boldsymbol{\theta}}$

$$\overline{\boldsymbol{\theta}}^{\boldsymbol{t}} = \boldsymbol{arg}\ \max_{\boldsymbol{\theta}} Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$$

With:

$$Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}}) = \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik}\ln(\pi_k)}_{Depends\ on\ \overline{\boldsymbol{\pi}}_{\boldsymbol{k}}} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N}\left(\sum_{j=1}^{D}\left(X_{ij}\ln(\theta_{jk}) + (1 - X_{ij})\ln\left((1 - \theta_{jk})\right)\right)\right)}_{Depends\ on\ \overline{\overline{\boldsymbol{\theta}}}_{\boldsymbol{KxD}}}$$

Using the updated r_{ik} values from the E-step.

We have that:

$$\begin{aligned} Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}}) &= \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik}\ln\left(\pi_{\boldsymbol{k}} p_{\boldsymbol{k}}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{X}}_i|\overline{\boldsymbol{\theta}}_{\boldsymbol{k}})\right) \\ &= \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik}\ln(\pi_{\boldsymbol{k}})}_{Depends\ on\ \overline{\pi}_k} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik}\ln\left(p_{\boldsymbol{k}}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{X}}_i|\overline{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)}_{Depends\ on\ \overline{\overline{\boldsymbol{\theta}}}_{\boldsymbol{KxD}}} \end{aligned}$$

So we need to maximize this over $\overline{\boldsymbol{\theta}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}, \overline{\overline{\boldsymbol{\theta}}}_{KxD}\}$.
Since $Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$ is a concave function over $\overline{\boldsymbol{\theta}}$, we could perform the derivative with respect to $\overline{\boldsymbol{\theta}}$ and set it to 0, to get the $\overline{\boldsymbol{\theta}}$ that maximizes the function $Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$

$$\frac{\partial Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \overline{\boldsymbol{\theta}}}|_{\overline{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}^{\boldsymbol{t}}} = 0$$

3.5.4.1 Value of the mixing coefficients

From the general case, we already know that:

$$\boldsymbol{\pi_k} = \frac{\boldsymbol{N_k}}{\boldsymbol{N}} = \frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{r_{ik}}$$

3.5.4.2 Value of the parameters of the distributions

We are going to calculate the value of the parameter vector $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$ of the k-component that maximizes $Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})$ to obtain the next value of these coefficients $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$:

$$\frac{\partial Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}}|_{\overline{\boldsymbol{\theta}}_{\boldsymbol{k}} = \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\boldsymbol{t}}} = 0$$

Remember: $\overline{\boldsymbol{\theta}} = \{\overline{\boldsymbol{\pi}}_{\boldsymbol{K}}, \overline{\overline{\boldsymbol{\theta}}}_{KxD}\}$ and $\overline{\overline{\boldsymbol{\theta}}}_{KxD} = [\overline{\boldsymbol{\theta}}_{\boldsymbol{1}}, \dots, \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}, \dots, \overline{\boldsymbol{\theta}}_{\boldsymbol{K}}]$ so we are trying to get the value of one of there $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$ vectors.

$$\frac{\partial Q(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}^{\boldsymbol{t-1}})}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}}|_{\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}=\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\boldsymbol{t}}} = \frac{\partial}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}}\left(\underbrace{\sum_{i=1}^{N} r_{ik} \ln\left(p_{\boldsymbol{k}}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{X}}_i | \overline{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)}_{Depends\ on\ \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}}\right) = 0$$

So, to get $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}^{\boldsymbol{t}}$ we must solve the equation:

$$\sum_{i=1}^{N} r_{ik}\left[\frac{\partial}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}}\left(\ln\left(p_{\boldsymbol{k}}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{X}}_i | \overline{\boldsymbol{\theta}}_{\boldsymbol{k}})\right)\right)\right] = 0$$

$$\sum_{i=1}^{N} r_{ik}\left[\frac{\partial}{\partial \overline{\boldsymbol{\theta}}_{\boldsymbol{k}}}\left(\sum_{j=1}^{D}\left(X_{ij}\ln(\theta_{jk}) + (1 - X_{ij})\ln\left((1 - \theta_{jk})\right)\right)\right)\right] = 0$$

For the $j - th$ element of $\overline{\boldsymbol{\theta}}_{\boldsymbol{k}}$ we have:

$$\sum_{i=1}^{N} r_{ik}\left[\frac{\partial}{\partial \theta_{jk}}\left(\sum_{j=1}^{D}\left(X_{ij}\ln(\theta_{jk}) + (1 - X_{ij})\ln\left((1 - \theta_{jk})\right)\right)\right)\right] = 0$$

$$\sum_{i=1}^{N} r_{ik}\left[\frac{\partial}{\partial \theta_{jk}}\left(X_{ij}\ln(\theta_{jk}) + (1 - X_{ij})\ln\left((1 - \theta_{jk})\right)\right)\right] = 0$$

$$\sum_{i=1}^{N} r_{ik}\left[\left(\frac{X_{ij}}{\theta_{jk}} - \frac{(1 - X_{ij})}{(1 - \theta_{jk})}\right)\right] = 0$$

From this equation we finally get:

$$\boldsymbol{\theta_{jk}} = \frac{\boldsymbol{1}}{\boldsymbol{N_k}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{r_{ik}X_{ij}}$$

3.5.5 Pseudo-code for the EM for multinomial distribution

3) Initilize parameters of the mixture model $\bar{\boldsymbol{\theta}} = \{\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}, \bar{\bar{\boldsymbol{\theta}}}_{KxD}\}$. at random according to the paramters constraints :

4) While (Stop_condition)

- E-Step: Update responsibility matrix as:
-

$$r = \begin{bmatrix} \bar{\boldsymbol{r}}_{\boldsymbol{1}} \\ \bar{\boldsymbol{r}}_{\boldsymbol{2}} \\ \vdots \\ \bar{\boldsymbol{r}}_{\boldsymbol{N}} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1K} \\ r_{21} & r_{22} & \cdots & r_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NK} \end{bmatrix} \quad with \quad r_{ik} = \frac{\pi_k \left(\prod_{j=1}^{D} \theta_{jm}^{X_{ij}} (1 - \theta_{jm})^{1-X_{ij}} \right)}{\sum_{m=1}^{K} \pi_m \left(\prod_{j=1}^{D} \theta_{jm}^{X_{ij}} (1 - \theta_{jm})^{1-X_{ij}} \right)}$$

- M-Step: Update parameters of the mixture model $\bar{\boldsymbol{\theta}} = \{\bar{\boldsymbol{\pi}}_{\boldsymbol{K}}, \bar{\bar{\boldsymbol{\theta}}}_{KxD}\}$

$$\bar{\boldsymbol{\pi}}_{\boldsymbol{K}} = [\pi_1, \ldots, \pi_k, .., \pi_K] \quad with \quad \pi_k = \frac{N_k}{N} = \frac{1}{N} \sum_{i=1}^{N} r_{ik}$$

$$\bar{\bar{\boldsymbol{\theta}}}_{KxD} = \begin{bmatrix} \bar{\boldsymbol{\theta}}_{\boldsymbol{1}} \\ \bar{\boldsymbol{\theta}}_{\boldsymbol{2}} \\ \vdots \\ \bar{\boldsymbol{\theta}}_{\boldsymbol{K}} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1K} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{D1} & \theta_{D2} & \cdots & \theta_{DK} \end{bmatrix} \quad with \quad \boldsymbol{\theta_{dk}} = \frac{\mathbf{1}}{\boldsymbol{N_k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{r_{ik} X_{id}}$$

- Calculate the Complete Log likelihood for checking convergence:

$$I_C(\bar{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{I}(z_i = k) \left(ln(\pi_k) + \sum_{j=1}^{D} \left(X_{ij} \ln(\theta_{jk}) + (1 - X_{ij}) \ln\left((1 - \theta_{jk}) \right) \right) \right)$$

Computationally it is easier to calculate:

$$I_C(\bar{\boldsymbol{\theta}}) = \ln \left(\sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{I}(z_i = k) \pi_{\boldsymbol{k}} \prod_{j=1}^{D} \theta_{jk}^{X_{ij}} (1 - \theta_{jk})^{1-X_{ij}} \right)$$