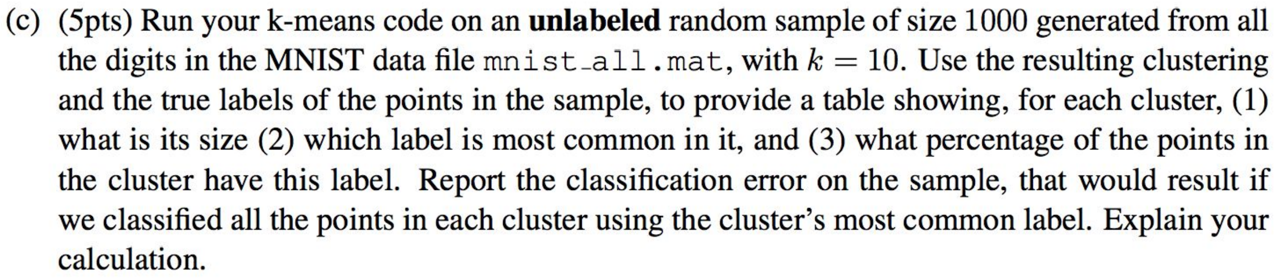
Assignment 4



(c) (5pts) Run your k-means code on an **unlabeled** random sample of size 1000 generated from all the digits in the MNIST data file `mnist_all.mat`, with $k = 10$. Use the resulting clustering and the true labels of the points in the sample, to provide a table showing, for each cluster, (1) what is its size (2) which label is most common in it, and (3) what percentage of the points in the cluster have this label. Report the classification error on the sample, that would result if we classified all the points in each cluster using the cluster's most common label. Explain your calculation.

Answer 1.C.

Cluster Id	Cluster Size	Most Common Label	Percentage with Label
1	131	7	0.5267
2	104	8	0.4038
3	76	0	0.9605
4	78	2	0.8205
5	102	9	0.3725
6	143	3	0.5734
7	88	1	0.7614
8	82	4	0.3415
9	108	1	0.6296
10	88	6	0.9432

We will now show the classification error in case we classified the samples by the most common label. We will notice that column4*column2 will give us the amount of successes per cluster. If we divide this number by 1000 we would get the success rate, and 1 minus the success rate is the classification error. We will now calculate:

$Err = 1 - sum(column2 * column4)/1000 = 1 - (69 + 42 + 73 + 64 + 38 + 82 + 67 + 28 + 68 + 83)/1000 =$ $1 - 614/1000 \;=\; 0.386 = 38.6\%$

(d) (5pts) Repeat (c) for your single linkage algorithm, again reporting the table and the classification error. Which clustering algorithm worked better for this problem?

Answer 1.D.

Cluster Id	Cluster Size	Most Common Label	Percentage with Label
1	991	1	0.1362
2	1	8	1
3	1	2	1
4	1	5	1
5	1	8	1
6	1	3	1
7	1	8	1
8	1	4	1
9	1	0	1
10	1	8	1

$Err = 1 - sum(column2 * column4)/1000 = 1 - (135\ +\ 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1)/1000 = 1 - 144/1000\ =\ 0.856 = 85.6\%$

It is easy to see that the Kmeans error rate is significantly lower than single linkage. Therefore it probably works better for this problem.

Answer 1.E.

Kmeans:

Cluster Id	Cluster Size	Most Common Label	Percentage with Label
1	140	5	0.25
2	91	6	0.8791
3	135	3	0.4444
4	225	7	0.3289
5	115	2	0.6174
6	132	1	0.7424
7	66	0	0.8939
8	96	0	0.4271

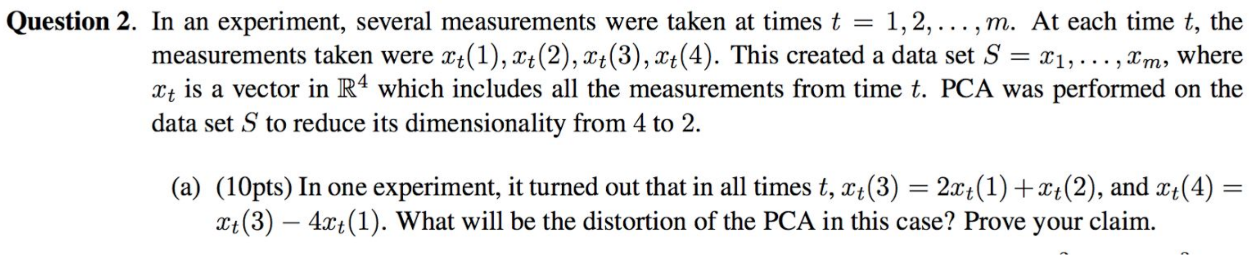
$Err = 1 - sum(column2 * column4)/1000 = 1 - 518/1000 \; = \; 0.482 = 48.2\%$

Single-Linkaged:

Cluster Id	Cluster Size	Most Common Label	Percentage with Label
1	991	7	0.1231
2	1	8	1
3	2	3	1
4	2	8	1
5	1	0	1
6	1	0	1
7	1	8	1
8	1	2	1

$Err = 1 - sum(column2 * column4)/1000 = 1 - 131/1000 \; = \; 0.869 = 86.9\%$

Both algorithms did worse with 8 clusters than with 10 clusters. It is more noticeable on Kmeans since its error rate increased by ten percents. If we look at the success percentage we can see that in some clusters the the success rate is very low. We believe that in those clusters, the algorithm was “forced” to put relatively “far” examples, since it run out of clusters. It is obvious that the number of clusters does not suitable for our problem, and it is well reflected in the results.



Question 2. In an experiment, several measurements were taken at times $t = 1, 2, \ldots, m$. At each time t, the measurements taken were $x_t(1), x_t(2), x_t(3), x_t(4)$. This created a data set $S = x_1, \ldots, x_m$, where x_t is a vector in $\mathbb{R}^4$ which includes all the measurements from time t. PCA was performed on the data set S to reduce its dimensionality from 4 to 2.

(a) (10pts) In one experiment, it turned out that in all times t, $x_t(3) = 2x_t(1) + x_t(2)$, and $x_t(4) = x_t(3) - 4x_t(1)$. What will be the distortion of the PCA in this case? Prove your claim.

Answer 2.a. We will prove that the distortion for this experiment is 0. We will provide $U \in R^{4x2}, V \in R^{2x4}$ $s.t$ $\sum_{i=1}^{m} \|x_i - UVx_i\|^2 = 0$.

We will first notice that $x_t(3) = 2 * x_t(1) + x_t(2)$, $x_t(4) = x_t(3) - 4x_t(1) =- 2x_t(1) + x_t(2)$, therefor all vectors in the expriments are from $span([1; 0; 2; -2], [0; 1; 1; 1])$

Let V= $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$ U= $\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -2 & 1 \\ -2 & 1 \end{pmatrix}$

We will now show that for each vector x_t, $UV * x_t = x_t$:

UV= $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -2 & 1 & 0 & 0 \end{pmatrix}$ $x_t = a * [1; 0; 2; -2] + b * [0; 1; 1; 1]$

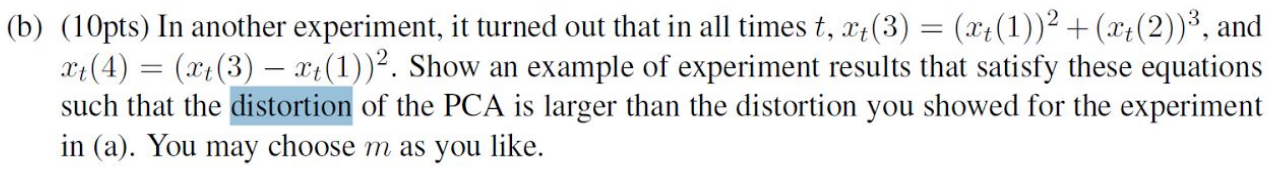
hence:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -2 & 1 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} 1 \times a \\ b \\ 2 \times a + b \\ -2 \times a + b \end{pmatrix} = \begin{pmatrix} a \\ b \\ 2 \times a + b \\ -2 \times a + b \end{pmatrix}$$

Therefore $UV * x_t = x_t$ foreach t. Now lets look again in the distortian:

$$\sum_{i=1}^{m} \|x_i - UVx_i\|^2 = \sum_{i=1}^{m} \|x_i - x_i\|^2 = \sum_{i=1}^{m} 0 = 0$$

We showed that under $U \in R^{4x2}, V \in R^{2x4}$ that we supplied the distortian is zero.



(b) (10pts) In another experiment, it turned out that in all times t, $x_t(3) = (x_t(1))^2 + (x_t(2))^3$, and $x_t(4) = (x_t(3) - x_t(1))^2$. Show an example of experiment results that satisfy these equations such that the distortion of the PCA is larger than the distortion you showed for the experiment in (a). You may choose m as you like.

Answer 2.b. We will prove that the distortion for this experiment is more than 0.
Let's observe in the experiment results $x_1 = [1, 0, 1, 0]$; $x_2 = [0, 1, 1, 1]$; $x_3 = [1, -1, 0, 1]$, we can see that experiment satisfy these equation.
We will now do PCA for those experiment results:

$$\mathrm{A} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 & 1 \end{pmatrix} =$$

$$\mathrm{A} = \begin{pmatrix} 2 & -1 & 1 & 1 \\ -1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{pmatrix}$$

$$det\left(\begin{pmatrix} 2 & -1 & 1 & 1 \\ -1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{pmatrix} - \lambda \cdot I \right) = (\lambda - 4)(\lambda - 3)(\lambda - 1) \cdot \lambda$$

$$\rightarrow the\ eigenvalues\ for \begin{pmatrix} 2 & -1 & 1 & 1 \\ -1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{pmatrix} are\ \lambda_1 = 4, \lambda_2 = 3, \lambda_3 = 1, \lambda_4 = 0.$$

Therefore, the PCA algorithm will choose the 2 highest value of λ for reducing to two dimensions, hence $\lambda_1 = 4, \lambda_2 = 3$.

And the distortion is equal to the sum of the lowest d - k eigenvalues, hence

$$\text{distortion} = \lambda_3 + \lambda_4 = 1 + 0$$

$$\rightarrow \text{ distortion} = 1 > 0. \blacksquare$$

Question 3. Consider the following distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{-1, 1\}^2$ and $\mathcal{Y} = \{-1, 1\}$.

(a) (10pts) Does the Naive-Bayes assumption hold for this distribution? Prove your claim.

Answer 3.a. We will prove Naive Bayes assumption does not hold for this distribution.
We will first notice that:

$P(Y = 1) = \sum_{i,j \in \{-1,1\}} P(X = (i,j)|Y = 1) = (5 + 0 + 11 + 14)/60 = 20/60$
$P(X = (-1,-1)|Y = 1) = \frac{5}{60}/\frac{20}{60} = 0.25$
$P(X = (-1, 1)|Y = 1) = 0/\frac{20}{60} = 0$
$P(X = (1,-1)|Y = 1) = \frac{11}{60}/\frac{20}{60} = \frac{11}{20}$
$P(X_1 =-1|Y = 1) = P(X = (-1, 1)|Y = 1) + P(X = (-1,-1)|Y = 1) = 0.25$
$P(X_2 =-1|Y = 1) = P(X = (1,-1)|Y = 1) + P(X = (-1,-1)|Y = 1) = 0.25 + \frac{11}{20} = 0.8$

The Naive Bayes assumption claims that X1 and X2 must be conditionally independent:
$P[X = x|Y = y] = \pi_{i=1}{}^{n} P[X(i) = x_i|Y = y]$
We will show that it doesn't apply on $P(X = (-1,-1)|Y = 1)$, meaning:
$P(X = (-1,-1)|Y = 1) \neq P[X_1 =-1|Y = 1] * P[X_2 =-1|Y = 1]$
Whe shoed that $P(X = (-1,-1)|Y = 1) = 0.25$
Lets calculate $P[X_1 =-1|Y = 1] * P[X_2 =-1|Y = 1]$:
$P[X_1 =-1|Y = 1] * P[X_2 =-1|Y = 1] = 0.25 * 0.8 = 0.2$
Therefore $P(X = (-1,-1)|Y = 1) \neq P[X_1 =-1|Y = 1] * P[X_2 =-1|Y = 1]$ and Naive Bayes assumption does not hold for this distribution.

(b) (10pts) Suppose we had a sample $S \sim \mathcal{D}^m$, such that the frequencies of the possible (x, y) in the data set was *exactly* the same as in the distribution, and suppose that we then ran the Naive-Bayes algorithm on this data set. What predictor would we get from this algorithm? Prove your claim.

Answer 3.b. We will first show that this distribution fits the symmetric case:
$p_1 = P[x_1 = 1 \mid Y = 1] = \frac{15}{20} = \frac{3}{4}, p_1{}' = P[x_1 =-1 \mid Y =-1] = \frac{30}{40} = \frac{3}{4} \rightarrow p_1 = p_1{}'$
$p_2 = P[x_2 = 1 \mid Y = 1] = \frac{4}{20} = \frac{1}{5}, p_2{}' = P[x_2 =-1 \mid Y =-1] = \frac{8}{40} = \frac{1}{5} \rightarrow p_2 = p_2{}'$
Therefore we are in the symmetric case.
We showed in class that the Naive-Bayes predictor in the symmetric case is:

$h^*(x) = sign(\sum_{i=0}^{n} log(\frac{p_i}{1-p_i}) * x(i))$, $p_0 = P[Y = 1]$, $x_0 = 1$

$p_0 = P[Y = 1] = \frac{20}{60} = 1/3 \rightarrow$
$h^*(x) = sign(log(\frac{1/3}{2/3}) * 1 + log(\frac{3/4}{1/4}) * x_1 + log(\frac{1/5}{4/5}) * x_2) = sign(log(\frac{1}{2}) + log(3) * x_1 + log(\frac{1}{4}) * x_2)$

(c) (5pts) Compare the Bayes-optimal predictor $h : \mathcal{X} \to \mathcal{Y}$ for the distribution $\mathcal{D}$ to the one you got in question (b). Explain why your answer does not stand in contradiction to (a).

The Bayes-Optimal predictor is $h^*(x) = argmax_{y \in Y} P_{(X,Y)}[Y=y|X=x]$.
For our case we can write $h^*(x) = \ if\ P[y = 1\ \wedge X = x] > P[y =-1\ \wedge X = x]\ then\ 1\ else\ 0$
We will now show for each X in the distribution the Bayes Optimal and the Naive Bayes outputs:

X_1	X_2	Naive-Bayes	Bayes-Optimal
1	1	$sign(log(\frac{1}{2}) + log(3) + log(\frac{1}{4})) =$ $sign(-0.98083) =-1$	$P[y = 1\ \wedge X = [1, 1]] = 4/60$ $P[y =-1\ \wedge X = [1, 1]] = 8/60\ \rightarrow$ $h^*(x) =-1$
1	-1	$sign(log(\frac{1}{2}) + log(3) - log(\frac{1}{4})) =$ $sign(1.7918) = 1$	$P[y = 1\ \wedge X = [1,-1]] = 11/60$ $P[y =-1\ \wedge X = [1,-1]] = 2/60\ \rightarrow$ $h^*(x) = 1$
-1	1	$sign(log(\frac{1}{2}) - log(3) + log(\frac{1}{4})) =$ $sign(-3.1781) =-1$	$P[y = 1\ \wedge X = [-1, 1]] = 0$ $P[y =-1\ \wedge X = [-1, 1]] = 24/60\ \rightarrow$ $h^*(x) =-1$
-1	-1	$sign(log(\frac{1}{2}) - log(3) - log(\frac{1}{4})) =$ $sign(-0.40547) =-1$	$P[y = 1\ \wedge X = [-1,-1]] = 5/60$ $P[y =-1\ \wedge X = [-1,-1]] = 6/60\ \rightarrow$ $h^*(x) =-1$

We can see that the Naive-Bayes and the Bayes-Optimal agrees, and for each sample return same label. It doesn't stand in contradiction to the claim in part a, because even though $P[X = x|Y = y] \neq \pi_{i=1}^{n} P[X(i) = x_i|Y = y]$, those values are close enough for this praticular problem.

Question 4. Let $\mathcal{X} = \{0, 1, 2\}$. Let $\Theta \subseteq [0, 1]^3$ such that for $\theta \in \Theta$, $\theta(1) + \theta(2) + \theta(3) = 1$. Define a *Trinomial* distribution $\mathcal{D}_\theta$ for $\theta \in \Theta$ as follows: $\mathbb{P}_{X \sim \mathcal{D}_\theta}[X = i] = \theta(i)$. Assume that we have a sample $S = x_1, \ldots, x_m \sim \mathcal{D}_\theta^m$.

(a) (10pts) Let $\Theta' = \{\theta \in \Theta \mid \theta(1) = 3\theta(2)\}$. Give an explicit formula for the value of the maximum likelihood estimator $\hat{\theta}$ using $x_1, \ldots, x_m$, assuming that $\theta \in \Theta'$. Prove your claim.

Answer 4.a. We will prove that $\widehat{\theta} = \frac{|\{i:x_i=0\}|+3|\{i:x_i=1\}|}{3m+6|\{i:x_i=1\}|}$.

The log likelihood function:

$$L(S;\theta) = log(\theta(1)) \cdot |\{i : x_i = 0\}| + log(\theta(2)) \cdot |\{i : x_i = 1\}| + log(\theta(3)) \cdot |\{i : x_i = 2\}| =$$

$$=> L(S;\theta) = log(3\theta(2)) \cdot |\{i : x_i = 0\}| + log(\theta(2)) \cdot |\{i : x_i = 1\}| + log(1 - \theta(1) - \theta(2)) \cdot |\{i : x_i = 2\}| =$$

$$=> L(S;\theta) = log(3\theta(2)) \cdot |\{i : x_i = 0\}| + log(\theta(2)) \cdot |\{i : x_i = 1\}| + log(1 - 4\theta(2)) \cdot |\{i : x_i = 2\}|$$

$$f(x) = log(3x) \cdot |\{i : x_i = 0\}| + log(x) \cdot |\{i : x_i = 1\}| + log(1 - 3x) \cdot |\{i : x_i = 2\}|$$

Maxmizing w.r.t $\theta(2)$ gives the ML estmiator .

$$f'(x) = \frac{3|\{i:x_i=0\}|}{3x} + \frac{|\{i:x_i=1\}|}{x} - \frac{4|\{i:x_i=2\}|}{1-4x}$$

Taking derivative w.r.t x and comaring to zero give:

$$f'(x) = 0 \Rightarrow \frac{|\{i:x_i=0\}|}{x} + \frac{|\{i:x_i=1\}|}{x} - \frac{4|\{i:x_i=2\}|}{1-4x} = 0 \Rightarrow$$

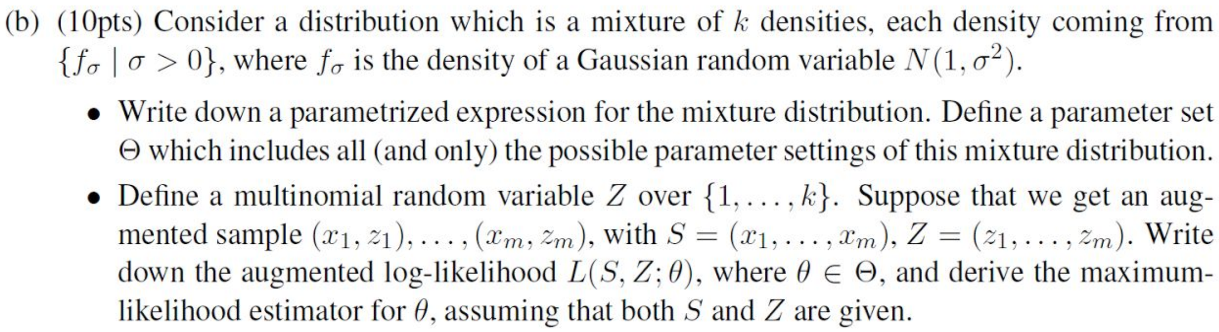
$$\Rightarrow (1-4x) \cdot |\{i : x_i = 0\}| + (1-4x) \cdot |\{i : x_i = 1\}| - 4x \cdot |\{i : x_i = 2\}| = 0$$

$$\Rightarrow 1 \cdot |\{i : x_i = 0\}| + 1 \cdot |\{i : x_i = 1\}| = 4x \cdot |\{i : x_i = 0\}| + 4x \cdot |\{i : x_i = 1\}| + 4x \cdot |\{i : x_i = 2\}|$$

$$\Rightarrow x = \frac{|\{i:x_i=0\}|+|\{i:x_i=1\}|}{4|\{i:x_i=0\}|+4|\{i:x_i=1\}|+4|\{i:x_i=2\}|} = \frac{|\{i:x_i=0\}|+|\{i:x_i=1\}|}{4m} .$$

We know that $\widehat{\theta} = argmax_{(\theta)}\ L(S;\Theta)$, than:

$$\Rightarrow \widehat{\theta} = \frac{|\{i:x_i=0\}|+|\{i:x_i=1\}|}{4m} .$$



(b) (10pts) Consider a distribution which is a mixture of k densities, each density coming from $\{f_\sigma \mid \sigma > 0\}$, where f_σ is the density of a Gaussian random variable $N(1, \sigma^2)$.

- Write down a parametrized expression for the mixture distribution. Define a parameter set Θ which includes all (and only) the possible parameter settings of this mixture distribution.
- Define a multinomial random variable Z over $\{1, \ldots, k\}$. Suppose that we get an augmented sample $(x_1, z_1), \ldots, (x_m, z_m)$, with $S = (x_1, \ldots, x_m)$, $Z = (z_1, \ldots, z_m)$. Write down the augmented log-likelihood $L(S, Z; \theta)$, where $\theta \in \Theta$, and derive the maximum-likelihood estimator for θ, assuming that both S and Z are given.

Answer 4.b.

Define $\Theta = \{p_1 .... p_k, \sigma_1 .... \sigma_k | \ 1 \le i \le k \ : \sigma_i > 0, \sum_{i=1}^{k} p_i = 1\}, \ p = (p_{1....} p_k), \ \sigma = (\sigma_{1.....} \sigma_k)$

X is mixture distribution than we can say:

$$X \sim p_1 \cdot N(1, \sigma_1{}^2) + p_2 \cdot N(1, \sigma_2{}^2) + ... + p_k \cdot N(1, \sigma_k{}^2) = \sum_{i=1}^{k} p_i \cdot N(1, \sigma_i{}^2)$$

$$f_{\sigma_j} = N(1, \sigma_j{}^2)$$

$$L(S, Z, \Theta) = \sum_{i=1}^{m} log(\pi_{j=1}{}^{k}(p_j * f_{\sigma_j}(x_i))^{I[z_i=j]}) = \sum_{i=1}^{m} \sum_{j=1}^{k} I[z_i = j](log(p_j) + log(f_{\sigma_j}(x_i))$$

$$= \sum_{j=1}^{k} log(p_j)(\sum_{i=1}^{m} I[z_i = j]) + \sum_{i=1}^{m} \sum_{j=1}^{k} I[z_i = j] * log(f_{\sigma_j}(x_i))$$

$$= \sum_{j=1}^{k-1} log(p_j)(\sum_{i=1}^{m} I[z_i = j]) + log(p_k)(\sum_{i=1}^{m} I[z_i = k]) + \sum_{i=1}^{m} \sum_{j=1}^{k} I[z_i = j] * log(f_{\sigma_j}(x_i))$$

$$= \sum_{j=1}^{k-1} log(p_j)(\sum_{i=1}^{m} I[z_i = j]) + log(1 - p_1 - p_2 .... - p_{k-1})(\sum_{i=1}^{m} I[z_i = k]) + \sum_{i=1}^{m} \sum_{j=1}^{k} I[z_i = j] * log(f_{\sigma_j}(x_i))$$

Now we will derivate this function for Θ, hence foreach p_j, σ_j :

Foreach p_j :

$$\frac{d(L)}{d(p_j)} = \frac{(\sum_{i=1}^{m} I[z_i=j])}{p_j} - \frac{\sum_{i=1}^{m} I[z_i=k]}{1 - \sum_{i=1}^{k-1} p_i}, \quad \frac{d(L)}{d(p_j)} = 0 \rightarrow \frac{(\sum_{i=1}^{m} I[z_i=j])}{p_j} = \frac{\sum_{i=1}^{m} I[z_i=k]}{1 - \sum_{i=1}^{k-1} p_i} = \frac{\sum_{i=1}^{m} I[z_i=k]}{p_k}$$

We will notice that we arbitrary chose p_k, and any parameter p_j could use as p_k.

Therefore the foreach p_j the ratio $\frac{(\sum_{i=1}^{m} I[z_i=j])}{p_j}$ is some constant value, we will label this ratio as t.

$$\frac{(\sum_{i=1}^{m} I[z_i=j])}{p_j} = t \rightarrow (\sum_{i=1}^{m} I[z_i = j]) = t * p_j \rightarrow t(p_1 + ... + p_k) = (\sum_{i=1}^{m} I[z_i = 1]) + ...I[z_i = k]) \rightarrow t * 1 = m$$

so t=m. In conclution the max value $\frac{d(L)}{d(p_j)} = 0 \ when \ p_j = \frac{(\sum_{i=1}^{m} I[z_i=j])}{m}$.

We will now continue to derivate L foreach σ_j .

$$\frac{d(L)}{d(\sigma_j)} = (\sum_{i=1}^{m} \sum_{j=1}^{k} I[z_i = j] * log(f_{\sigma_j}(x_i)))' = (\sum_{i=1}^{m} \sum_{j=1}^{k} I[z_i = j] * log(\frac{1}{\sigma_j * \sqrt{2 * \pi}} * exp(- \frac{(x_i - 1)^2}{2 * \sigma_j{}^2})))'$$

$$= (\sum_{i=1}^{m}\sum_{j=1}^{k} I[z_i = j] * (-\frac{(x_i-1)^2}{2*\sigma_j^2} - log(\sigma_j) - log(\sqrt{2\pi}))' = (-m * log(\sqrt{2\pi}) * \sum_{i=1}^{m}\sum_{j=1}^{k} I[z_i = j] * (-\frac{(x_i-1)^2}{2*\sigma_j^2} - log(\sigma_j))'$$

$$= (-m * log(\sqrt{2\pi}) * \sum_{i=1}^{m} I[z_i = j] * (\frac{(x_i-1)^2}{\sigma_j^3} - \frac{1}{\sigma_j}), \quad \frac{d(L)}{d(\sigma_j)} = 0 \rightarrow \sum_{i=1}^{m} I[z_i = j] * \frac{(x_i-1)^2}{\sigma_j^3} = \sum_{i=1}^{m} I[z_i = j] * \frac{1}{\sigma_j} \rightarrow$$

$\sigma_j = \sqrt{\frac{\sum_{i=1}^{m} I[z_i=j]*(x_i-1)^2}{\sum_{i=1}^{m} I[z_i=j]}}$ In conclution the max value $\frac{d(L)}{d(\sigma_j)} = 0 \; when \; \sigma_j = \sqrt{\frac{\sum_{i=1}^{m} I[z_i=j]*(x_i-1)^2}{\sum_{i=1}^{m} I[z_i=j]}}$.

So based on Z we maximaized any value in θ, hence we maximized θ.