Statistics 5525: Homework 3

For each homework assignment, turn in at the beginning of class on the indicated due date. Late assignments will only be accepted with special permission. Write each problem up *very* neatly (LaTeX is preferred). Show all of your work.

Problem 1

Consider the p-dimensional Gaussian Mixture Model:

$$x_i \sim \sum_k^K \pi_k p(x|\mu_k, \Sigma_k, C_k), \qquad \text{for } i = 1, \ldots, N, \tag{1}$$

where $p(x|\mu_k, \Sigma_k) = \frac{1}{\pi^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)}$.

In the special case where $\Sigma_k = \sigma I \ \forall k = 1 \ldots, K$, discuss the connections between the K-means algorithm and EM for fitting model (1). Additionally, show that as $\sigma \to 0$ the two methods coincide.

The Gaussian Mixture Model with EM algorithm is a kind of soft K-Means algorithm. In the E step, EM algorithm we calculate the expectation of the log likelihood as follows:

$$Q(\theta|\theta(t)) = E_{z|\theta(t)}[-\frac{1}{2}\sum_{i=1}^N\sum_{k=1}^K[(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) + Log\pi_k - Log\Sigma_k] \times \delta(z_i = k|x_i))]$$

To simply the above expression, we can obtain:

$$\begin{aligned} Q(\theta|\theta(t)) &= -\frac{1}{2}\sum_{i=1}^N\sum_{k=1}^K[(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) + Log\pi_k - Log\Sigma_k] \times E_{z|\theta(t)}[\delta(z_i = k|x_i))] \\ &= -\frac{1}{2}\sum_{i=1}^N\sum_{k=1}^K[(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) + Log\pi_k - Log\Sigma_k] \times P(z_i = k|x_i) \end{aligned} \tag{2}$$

Here the $P(z_i = k|x_i) = \pi_{i,k}^{(t)}$ is the posterior

$$P(z_i = k|x_i) = \pi_{i,k}^{(t)} = \frac{P(x_i|z_i = k, \theta^{(t)})Pr(z_i = k|\theta^{(t)})}{\sum_{k=1}^K P(x_i|z_i = k, \theta^{(t)})Pr(z_i = k|\theta^{(t)})}$$

with

$$P(x_i|z_i = k, \theta^{(t)}) = \frac{1}{\pi^{p/2}|\Sigma_k^{(t)}|^{1/2}} e^{-\frac{1}{2}(x_i-\mu_k^{(t)})'\Sigma_k^{-1}(x_i-\mu_k^{(t)})}$$

and

$$Pr(z_i = k|\theta^{(t)}) = \pi_k^{(t)}.$$

Here the posterior $P(z_i = k|x_i)$ gives us the probability for a data point x_i coming from the component C_k. Thus, the EM algorithm is implementing soft labeling. If Σ_k goes to zero in the EM algorithm, the $P(x_i|z_i = k, \theta^{(t)})$ goes to delta distribution. This means, we have:

$$P(x_i|z_i = k, \theta^{(t)}) = \delta(x_i - \mu_k^{(t)})$$

In addition, the posterior becomes

$$P(z_i = k|x_i) = \pi_{i,k}^{(t)} = \delta(x_i - \mu_k^{(t)})\pi_k^{(t)} / \sum_{k=1}^{K} \delta(x_i - \mu_k^{(t)})\pi_k^{(t)} \tag{3}$$

which effectively gives hard labeling (since they are step functions over the data space). A visualization of the above function (simplified version) can be found in page 512 of the text book. Thus, given $\Sigma_k \to 0$, the E step of the EM algorithm converges to the (hard/mutual exclusive) labeling step of the K-Means algorithm.

The M step maximize the $Q(\theta|\theta(t))$ in the E step with respect to the model parameters, which yields

$$\mu_k^{(t)} = \sum_{i=1}^{N} \pi_{i,k}^{(t)} x_i / \sum_{i=1}^{N} \pi_{i,k}^{(t)} \tag{4}$$

If Σ_k goes to zero, the denominator of the above expression will converge to 1, sine the posteriors $\pi_{i,k}^{(t)}$ will converge into the forms given by Eq.(3). Accordingly, we can write Eq.(4) as $\mu_k^{(t)} = \sum_{i=1}^{N} \pi_{i,k}^{(t)} x_i$, which is effectively the mean of all the data points in the component C_k. This is actually the second step of the K-Means algorithm when the new center of each cluster is calculated.

To sum up, the GMM with EM algorithm converges to the K-Means algorithm when the Σ_k (or the σ) goes to zero.

Problem 2

Download "ClusterSet1.txt" from the course webpage. Apply the k-means clustering procedure to this data set. You may code this up from scratch or use the built in functions in either R or Matlab. Discuss how you selected 'K', and why you believe it is correct.

I used R package to do the K-Means clustering. I also used the same K-Means function in R to calculate the Within Sum of Squares (WSS) and the Total Sum of Squares (TSS). Then I plotted the $w = WSS/TSS$ values with increasing value of K. The scree plot is shown in below.

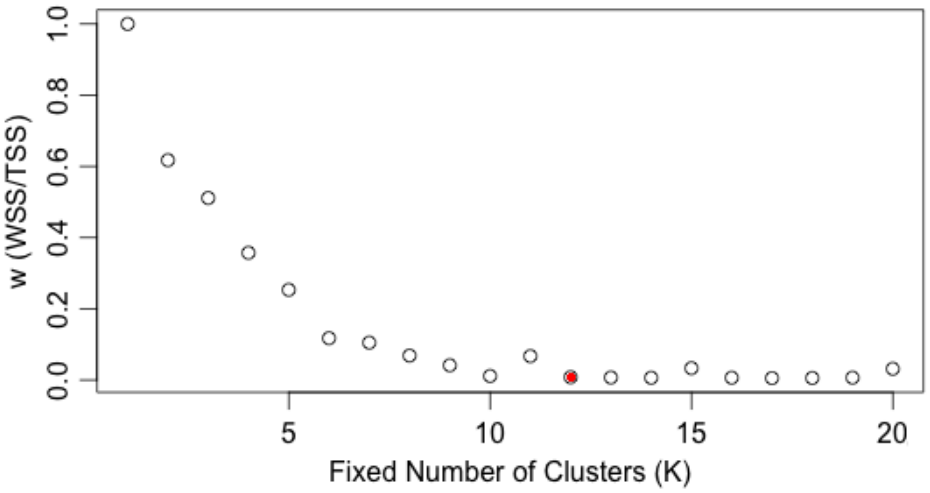

Figure 1: Scree plot of K-Means clustering, the optimal K is 12. The y axis is the ratio of Within Sum of Squares (WSS) and Total Sum of Squares (TSS).

It is very obvious from the scree plot that the optimal K is 12, where the w value reaches minimal while K is kept still small. The w values for K=1,2,3,...,20 can also be obtained from the R package: 1.000000000, 0.617371303, 0.511107268, 0.357233846, .253278036, 0.117447680, 0.105069464, 0.068609355, 0.041399420, 0.011358326, 0.067122556, 0.008019367, 0.007076398, 0.006339394, 0.033207078, 0.006506954, 0.005367669, 0.005656337, 0.006719339, 0.031540968. It is also obvious from reading the data that the w stabilizes at very small values once K reaches 12. The clustering result is shown in Fig.2. (Note different clusters may be dyed with one color.)

Problem 3

Recall that the EM algorithm for fitting model (1) iterates over the following updates: For $t = 1, \ldots, T$

1. $\pi_{i,k}^{(t)} = p(x_i \in C_k | \mu_k^{(t-1)}, \Sigma_k^{(t-1)})$
2. $\mu_k^{(t)} = \sum_{i=1}^N \pi_{i,k}^{(t)} x_i / \sum_{i=1}^N \pi_{i,k}^{(t)}$
3. $\Sigma_k^{(t)} = \sum_{i=1}^N \pi_{i,k}^{(t)} (x_i - \mu_k^{(t)})(x_i - \mu_k^{(t)})' / \sum_{i=1}^N \pi_{i,k}^{(t)}$

part a

Given $\pi_{i,k}$s, show that the M.L.E for μ_ks are given by $\sum_{i=1}^N \pi_{i,k} x_i / \sum_{i=1}^N \pi_{i,k}$.

With the $Q(\theta, \theta(t))$ as derived in Eq.(2), the $\mu_k^{(t)}$ can be obtained by solving

$$\frac{\partial Q(\theta, \theta(t))}{\partial \mu_k} = 0 \tag{5}$$

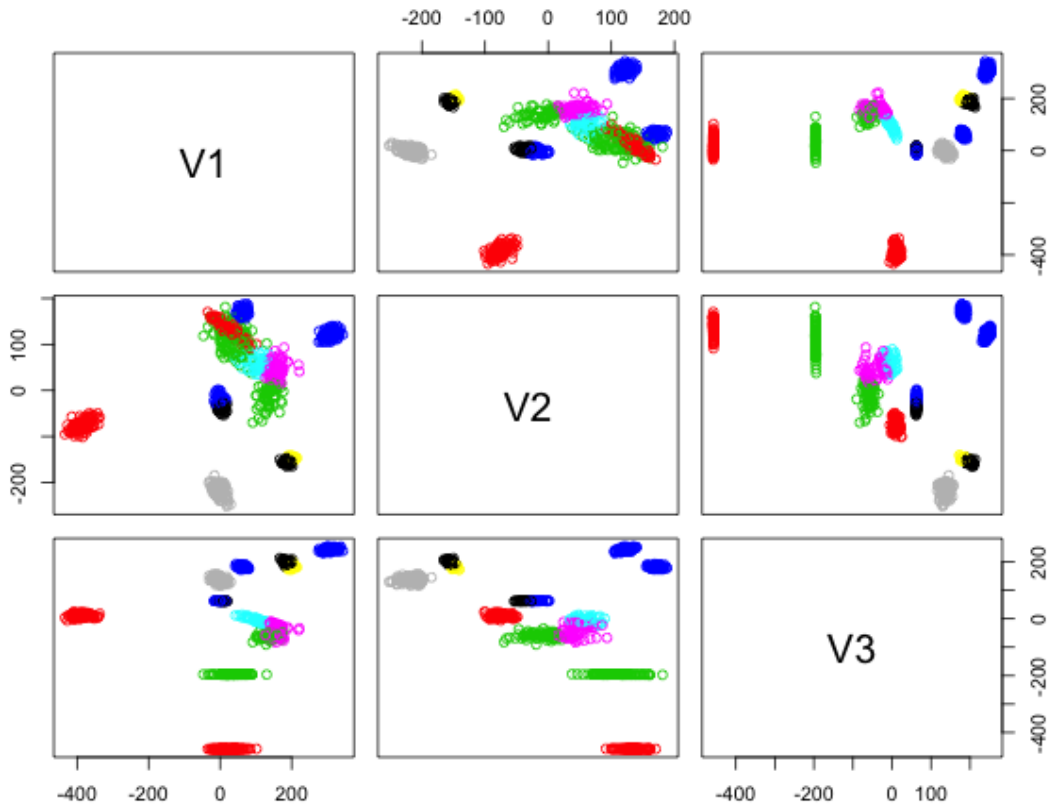

Figure 2: Result of K-Means clustering. Different pairwise combinations of (v_1, v_2, v_3) are used to display the clustering result.

This yields

$$\sum_{i=1}^{N}\sum_{k'=1}^{K}\frac{\partial[(x_i-\mu_{k'})^T\Sigma_{k'}^{-1}(x_i-\mu_{k'})\pi_{ik'}]}{\partial\mu_k}=0 \tag{6}$$

or

$$\sum_{i=1}^{N}\frac{\partial[(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\pi_{ik}]}{\partial\mu_k}=0 \tag{7}$$

According to vector calculus, the above equation can be simplified into

$$\sum_{i=1}^{N}(x_i-\mu_k)\pi_{ik}=0 \tag{8}$$

which ultimately gives

$$\mu_k^{(t)}=\sum_{i=1}^{N}\pi_{ik}^{(t)}x_i/\sum_{i=1}^{N}\pi_{ik}^{(t)} \tag{9}$$

part b

Given μ_ks and $\pi_{i,k}$s, show that the M.L.E for Σ_ks are given by $\sum_{i=1}^{N} \pi_{i,k}(x_i - \mu_k)(x_i - \mu_k)' / \sum_{i=1}^{N} \pi_{i,k}$ (Given μ_ks).

Similarly, with the $Q(\theta, \theta(t))$ as derived in Eq.(2), the $\Sigma_k^{(t)}$ can be obtained by solving

$$\frac{\partial Q(\theta, \theta(t))}{\partial \Sigma_k} = 0 \tag{10}$$

which is equivalent to

$$\sum_{i=1}^{N} \sum_{k'=1}^{K} \frac{\partial [Log|\Sigma_{k'}^{-1}| - (x_i - \mu_{k'})^T \Sigma_{k'}^{-1}(x_i - \mu_{k'})\pi_{ik'}]}{\partial \Sigma_k} = 0 \tag{11}$$

or

$$\sum_{i=1}^{N} \frac{\partial [Log|\Sigma_{k}^{-1}| - (x_i - \mu_{k})^T \Sigma_{k}^{-1}(x_i - \mu_{k})\pi_{ik}]}{\partial \Sigma_k} = 0 \tag{12}$$

This can be simplified into

$$\sum_{i=1}^{N} [\Sigma_k - (x_i - \mu_k)(x_i - \mu_k)^T]\pi_{ik} = 0 \tag{13}$$

which ultimately gives

$$\Sigma_k^{(t)} = \sum_{i=1}^{N} \pi_{ik}^{(t)}(x_i - \mu_k^{(t)})(x_i - \mu_k^{(t)})^T / \sum_{i=1}^{N} \pi_{ik}^{(t)} \tag{14}$$

part c

Implement the EM algorithm and using 'ClusterSet1.txt', compare results to those found in Problem 2.

For the 'ClusterSet1.txt' data, the implementation of the EM algorithm produced results as shown in Table 1. (See Appendix for code.)

1	2	3	4	5	6	7	8	9	10	11	12
97.00	94.00	13.00	103.00	17.00	105.00	109.00	96.00	82.00	110.00	93.00	81.00
0.10	0.10	0.01	0.10	0.02	0.10	0.11	0.10	0.08	0.11	0.09	0.08

Table 1: The number of data points in each cluster and the estimated values for π_k. $(k = 1, 2, ..., 12)$.

The number of data points in each cluster is evenly distributed from the EM algorithm. Also, we can calculate the Within Sum of Square (WSS) and Total Sum of Square (TSS) for the EM clustering results. Then we can obtain w =WSS/TSS=0.008116935. The K-Means with 12 clusters has w =0.008019367. Thus, the w(EM) is slightly larger than w(K-Means), meaning the performance of K-Means is slightly better.

Problem 4

Using a hierarchical clustering method with 'ClusterSet1.txt', compare results to those found in Problem 2 and 3. Show dendrograms, and discuss the distance function you settled on for your link function.

In this problem I used hierarchical clustering with Ward's method. The distance function I used is the Within Sum of Squares of Euclidean Distances (WSSED).

$$WSSED = \sum_{k=1}^{K} \sum_{i<j, x_i \epsilon C_k, x_j \epsilon C_k} \|x_i - x_j\|_2^2$$

Here $\|x_i - x_j\|_2$ is the Euclidean distance between data point x_i and data point x_j that belong to the same component C_k. The dendrograms is shown in Fig.3.

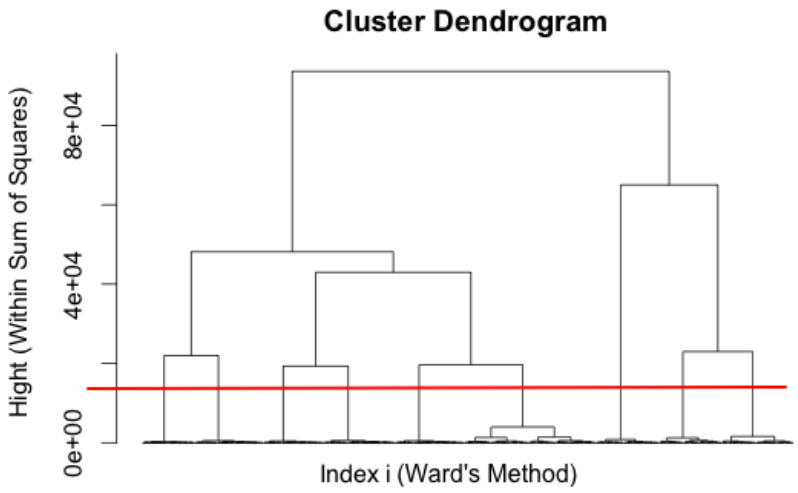

Figure 3: Dendrogram of Hierarchical Clustering with the Ward's method. The Hight is the WSSED.

From the dendrogram plotted for this Hierarchical algorithm, we can easily find the best number of clusters is 9, see the red line in Fig.3 (which crosses 9 branches in total). After labeling each data point and calculating the $w = WSS/TSS$ for the clustering result with 9 clusters. We obtain $w = 0.01152203$, which is a little bigger than the result of K-Means (0.008019367) and the result of EM (0.008116935). Note, the total number of clusters in the latter two algorithms is 11 instead of 9. Accordingly, we cannot claim this Hierarchical clustering result is worse than those of the K-Means algorithm and the EM algorithm.

Problem 5

Download "ClusterSet2.txt" from the course webpage. Using any method you find appropriate, determine the number of clusters and and assignment labels for each data point.

Here I used the Hierarchical Clustering with the Ward's method. The Dendrogram below clearly indicates we should set the number of clusters to be 10, see the red line in Fig.4 (which crosses 10 branches in total).

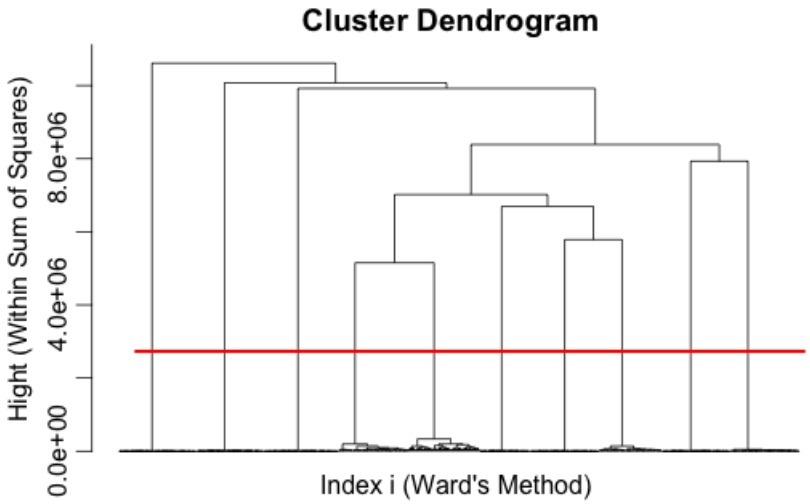

Figure 4: Dendrogram of Hierarchical Clustering with the Ward's method.

The labels for $x_1, x_2, ..., x_{1000}$ given by R code are shown in Fig.5.

```
 [1]  1  2  3  3  1  4  1  5  6  7  2  8  2  3  6  7  6  5  2  2  4  5  4  2  9  1  1  5  1 10  3  4  9  3  7  1
[37]  3  5  3  3  9  3  1 10  4  2  3  8  2  6 10  4  4  5  8  5 10  5  5 10  3  6  2  3 10  6  2  4 10  8  9  2
[73]  1  4  7  9  4  3  5  5  5  9  2  6  8  7  2  7  6  3  3  8 10  7  3  6  1  7  7  1  5  1  5  4  5  3  2  4
[109] 1  5  6  2  3  4  5  6  5  4  9  5 10  4  1  4  2  2  2  8  9  8  1  9  9  9  9  4  1  5  5 10  8  7  4  4
[145] 3  6  2  1  4  5  6  7  5  9  1  3  4  4  7  9  4  8  5  8  1  6  9  1  1  5  3  2  9  9  3 10  9  2 10  5
[181] 2  4  2 10  1  4  1  1  9  9  4  3 10  1  6 10  1  3  3  2 10 10  6  6  8  9  1  3  2 10 10  9  6  6 10  1
[217] 6  9  6  1  6 10  2  3  2  1  8  1  1  7  1 10  2  7  1  6  2  1  4  5  9  6  2  8  3  5  8  3  8  5  3  2
[253] 10  8  1  1  7 10  2  8  6  2  1  9  2  9  9  2  8  5  9  8  5  3  2  5  9  5  4  7  2  2  6  9  7  1  3  3
[289] 4  3  1  5  5  2  4  2  5  1  6  4  3  5  6  3  5 10  6  6  3  6  6  5  2  6  8  5  4 10  2  4  5  8  6  8
[325] 1 10  6  9  4  5  5  1  3  8 10  3  1  7  5  4  5 10  6  7 10  6  4  3 10  1  4 10 10  5  4 10  2  5  4  4
[361] 8  2  2 10  8  6  1  2  4  2  6  5  1  8  7  8  6  6  6  2  7  3 10  3  1  2  9  2 10  9  7  9  9  3  9  9
[397] 4  6  1  1  1  1  6  9  2 10  1  9  2  9  2  2  4  7  2  9  1  5  5  8  3  5  4  2  1  2  2  4  3  9 10  7
[433] 2  6  9  7  2  5  6  5  7  6  3  8  4  7  8  3  6  1 10  1  4  5  4  5  3 10  3  2 10  6  8 10  9 10  2  3
[469] 2  2  6  5  8 10  8  8  1  1  5  2  5  7  8 10  8  1  3  6  1  5  9  8  8  7  7  5  9  3 10 10  2  8  4  6
[505] 5  8  2  1  4  4  6  4  9  9  1  4  6  5  1  3  8  5  4  7  9  2  3  2 10  3  2 10  3  8  4  5  7 10  7  6
[541] 4  4  9  9  8  2  6  3  1  5  7  6  9  2  5  5  9  8 10  1  5  7  6  7  9  3  6  8  7  3  9  1  8  5  2  1
[577] 5  7 10  4  2  1  3  5  6  6  3  5  4  8  3  7  1 10  5  7  6  9  9  4  6  7  3  6 10  3  1  2  1  7  1  7
[613] 1  7  5  4  7  6 10  8  1  9  6  5  7  1  8  9  7  4  2 10  8  3 10  8  6  9 10  4  5  4  8  9  6  5  9  6
[649] 4  1  3  2  3  2  4  5  4 10  5  3  3 10  8  2  5  8  1  8  8  4  6 10  3  8  7  4  3  4  5  7  3  8  3  7
[685] 6  9  3  9 10  8 10  6  9  4 10  3  8  6  5  9  9  9  1  8  8  9  9  8  3 10  7  2  4  3  4  5  4  3  1  8
[721] 6  2  4  3  4  2 10  6  7  5  9  3  8  4  9  2  1  5  3  3 10  6  1  7  6  9  2  8  5  5  2  1  5  3 10  6
[757] 3  9  7  1  7  6  5  6 10 10  3  2  2  7  8  5  1  4  6  5  2  8 10  6 10  5  6  5  4  4 10 10  9 10  6  6
[793] 9  7  5 10  3  9  8 10  5  9  7  5  3  3  1  3  5  3  6  2  9  9  4  7  1 10  6  1 10  2  8  1  2  8  6  9
[829] 3  9  5  8  5  5  4  4  6  4  2  4  5  9  3  6  4  9  8  9 10  2  6  7  1  9  3  3  5  5  8 10  4  3  5  9
[865] 4  3  7  5  6  9  6  8  1 10  3  1  5  5  9 10  1  1  9  3  6  9  9  4  3  9  9  5  8  8  7 10  5  6  6  5
[901] 10 10 10  8  6  9  4  2  7  7  9  8  7 10  2  5  7  2  9  7  7  1  4  3  4  8 10  3 10 10  3  3  9  9  2  3
[937] 8 10  9  4  2 10  4  7 10  3 10  4  9  6  8  4  3  1  1  6  8  6  1  3  4  8  9  3  7  2  1  3  5  8  8  7
[973] 7  8 10  4 10 10 10  5  4  8  6  8  7  3 10  4  4  9  2  4  9 10  8  7 10  8  5  2
```

Figure 5: The labels for the data points $x_1, x_2, ..., x_{1000}$, a result of the Hierarchical Clustering with the Ward's method.