Homework #1

Yongjin Shin
20090488
Industrial Management Engineering, POSTECH
dreimer@postech.ac.kr

October 8, 2018

Problem 1. Detailed derivation of MoG algorithm (via direct maximization of likelihood)

Solution Description

Unlike K-means algorithm which assigns each data points to only one closest cluster, in real world, it is not easy to divide data set exactly into K groups mutually exclusively. Mixture model considers things are composed of several distributions at the same time so that even though the probability density function can not be able to be achieved with certain probability model, however, it consists of a few some probability density functions. In this sense, mixture of Gaussian is the model that several different Gaussian distributions are overlapped containing different mean and covariance respectivly. Then we can define linearly combinationed Gaussian function as follows:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{1}$$

where each Gausian density $\mathcal{N}(\mu_k, \sigma_k)$ is called a component of the mixture and the parameter π_k is called mixing coefficients. If we integrate both side of equation,

$$\int p(x)dx = \int \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)dx = 1 \tag{2}$$

Since each component $\mathcal{N}(\mu_k, \sigma_k)$ and p(x) are normalized, the sum of mixing coefficients should be 1. Also $p(x) \geq 0$ and $\mathcal{N}(x|\mu_k, \sigma_k) \geq 0$ which implies that $\pi_k \geq 0$ for all k. Therefore we can think of π_k as a probability. Then we can say this by slightly different way:

$$p(x) = \sum_{k=1}^{K} p(k)p(x_n|k) \tag{3}$$

where the same with equation except $p(k)$ represents π_k, which is the prior probability of picking the k^{th} component and $p(x|k) = \mathcal{N}(\mu_k, \sigma_k)$ as the probability of x conditioned on k.

Then let's introduce some latent variable r which is composed of K elements $r \in \mathbb{R}^K$ so that a particular element is equal to 1 and others all zero. Therefore $z_k \in \{0, 1\}$ and $\sigma_k z_k = 1$. Now we can define the joint distribution $p(x|z)$ in terms of marginal distribution p(z) and a conditional distribution $p(x|z)$. The marginal distribution over z can be given as:

$$\int p(x,z)dz = \int p(z)p(x|z)dz$$
$$p(z_k = 1) = \pi_k$$

In some sense, z_k represents that the corresponding x_n to z_k is included in k^{th} group, however, it doesn't need to be only one group so that π_k can be considered as the probability of k^{th} groups prior probability like as we defined before. Because z uses one hot encoding, we can also write the distribution as follows:

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{4}$$

Also, the conditional distribution of x give for the particular z_k is a Gaussian:

$$\begin{aligned} p(x|z_k = 1) &= \mathcal{N}(x|\mu_k, \Sigma_k) \\ p(x|z) &= \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \end{aligned}$$

Now we can see this:

$$\begin{aligned} p(x|\pi, \mu, \Sigma) &= \sum_{k=1}^{K} p(z_k = 1|\pi) p(x|z_k = 1) \\ &= \sum_{k=1}^{K} \Big[\prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \Big] \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \end{aligned}$$

Before solving the maximization problem, we need to define another term called **responsibility**, such that component k takes for explaining the observation x. We see π_k as the prior probability of $z_k = 1$ and this responsibility variable is kind of posterior probability of observed x. Simply we can see how much the particular Gaussian explain the observed value x. In other words, observed value x is composed of several different Gaussian containing its own responsibilities summed up to 1. The responsibilities can be defined as follows:

$$r_{k,n} = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)} \tag{5}$$

Let X be $N \times D$ matrix and the latent variable Z is $N \times K$ matrix. If we assume data set is i.i.d, then log-likelihood is given by:

$$\begin{aligned} \mathcal{L} &= \log p(X|\pi, \mu, \Sigma) \\ &= \sum_{n=1}^{N} \log p(x_n|\pi, \mu, \Sigma) \\ &= \sum_{n=1}^{N} \log \Big(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \Big) \end{aligned}$$

Maximum likelihood estimates of π_k, μ_k, Σ_k are computed as follows:
(1) μ_k

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu_k} &= -\sum_{n=1}^{N}\Big(\frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)}\Big)\Sigma_k^{-1}(x_n-\mu_k) \\
&= -\sum_{n=1}^{N} r_{k,n}\Sigma_k^{-1}(x_n-\mu_k) \\
&= 0
\end{aligned}
$$

The first equality can be achieved with using *equation (81)* in [3]. Suppose f is a multivariate normal distribution. Then,

$$\frac{\partial f}{\partial \mu} = f \times -\frac{1}{2}\big(\Sigma^{-1} + (\Sigma^{-1})^T\big) \times (x-\mu) \times (-1) \tag{6}$$

Since covariance matrix is symmetric, $\Sigma^{-1} = (\Sigma^{-1})^T$. Thus we can see that $\frac{\partial f}{\partial \mu} = f \times \Sigma^{-1}(x-\mu)$, so that after replacing with **responsibility** the first equality can be as above. Therefore,

$$\mu_k = \frac{\sum_{n=1}^{N} r_{k,n}x_n}{\sum_{n=1}^{N} r_{k,n}} \tag{7}$$

(2) Σ_k

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \Sigma_k^{-1}} &= -\sum_{n=1}^{N}\Big(\frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)}\Big)\frac{1}{2}\big(\Sigma_k - (x_n-\mu_k)(x_n-\mu_k)^T\big) \\
&= -\sum_{n=1}^{N} r_{k,n}\big(\Sigma k - (x_n-\mu_k)(x_n-\mu_k)^T\big) \\
&= 0
\end{aligned}
$$

Like how we did in the case of μ_k, the first equality can be derived as follows by using *equation (23)*, *equation (58)* and *equation (70)* in [3]:

$$
\begin{aligned}
\frac{\partial f}{\partial \Sigma^{-1}} &= \big[\frac{\partial det(\Sigma)^{\frac{-1}{2}}}{\partial \Sigma^{-1}} \times (rest)\big] + \big[f \times (\frac{-1}{2}(x-\mu)(x-\mu)^T)\big] \\
&= \big[\frac{1}{2}det(\Sigma)^{\frac{-1}{2}}\Sigma^T \times (rest)\big] + \big[f \times (\frac{-1}{2}(x-\mu)(x-\mu)^T)\big] \\
&= f \times \frac{1}{2}(\Sigma - (x-\mu)(x-\mu)^T)
\end{aligned} \tag{8}
$$

Therefore, we can conclude:

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{k,n}(x_n-\mu_k)(x_n-\mu_k)^T \tag{9}$$

(3) π_k

This can be regarded as constrain optimization problem since $\Sigma_{k=1}^{K}\pi_k = 1$. So apply Lagrange multiplier(λ) with the objective function then:

$$\mathcal{L}'(\pi,\lambda) = \mathcal{L} + \lambda[1 - \Sigma_{k=1}^{K}\pi_k] \tag{10}$$

Therefore, $\nabla_{\pi_k,\lambda}\mathcal{L}'(\pi,\lambda) = 0$ needs to be calculated:

$$\frac{\partial \mathcal{L}'(\pi,\lambda)}{\partial \lambda} = 1 - \Sigma_{k=1}^{K}\pi_k = 0 \tag{11}$$

$$\frac{\partial \mathcal{L}'(\pi,\lambda)}{\partial \pi_k} = \sum_{n=1}^{N} \Big(\frac{\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \Big) - \lambda = 0 \tag{12}$$

In equation (12) if we multiply π_k on both sides, then $\Sigma_{n=1}^{N} r_{k,n} = \pi_k \lambda$. The left side of this equation means that the number of elements in the kth group(N_k). Therefore,

$$\pi_k = \frac{N_k}{\lambda} \tag{13}$$

If we sum up (13) with using the constrain (or the result of equation (11)), $1 = \frac{N_1+N_2,+...+N_K}{\lambda}$. Therefore $\lambda = N$. Thus, we have:

$$\pi_k = \frac{N_k}{N} \tag{14}$$

Problem 2. Detailed derivation of EM-MoG.

Solution In the EM, we need to adapt the model with some latent variables z_n. For the simpler derivation, those latent variables will be regarded as discrete values. Then the log-likelihood is as follows:

$$\begin{aligned}\log p(x|\theta) &= \Sigma_{n=1}^{N} \log p(x|\theta) \\ &= \Sigma_{n=1}^{N} log(\Sigma_{z_n} p(x_n, z_n|\theta))\end{aligned} \tag{15}$$

This log-likelihood has the summation in order to marginalize over z, and which makes hard to calculate maximum likelihood estimation. Therefore wee need to deploy EM algorithm to approximate MLE.

First of all, the given log likelihood, let a single factor of the log-likelihood be $\mathcal{L}(\theta)$, has a lower bound as follows:

$$\begin{aligned}\mathcal{L}(\theta) &= log(\Sigma_{z_n} p(x, z_n|\theta)) \\ &= log\Sigma_{z_n} q(z_n)\big[\frac{p(x, z_n|\theta)}{q(z_n)}\big] \\ &\geq \Sigma_{z_n} q(z_n) log\big[\frac{p(x, z_n|\theta)}{q(z_n)}\big] \\ &= \mathcal{F}(q, \theta)\end{aligned} \tag{16}$$

From the above equation, the inequality can be derived by using Jensen's inequality. Then if we look at the lower bound closely,

$$\begin{aligned}\mathcal{L}(\theta) &\geq \mathcal{F}(q, \theta) \\ &= \Sigma_{z_n} q(z_n) log\big[\frac{p(x, z_n|\theta)}{q(z_n)}\big] \\ &= \Sigma_{z_n} q(z_n) log\big[\frac{p(z_n|x, \theta)}{q(z_n)} \times p(x|\theta)\big] \\ &= \Sigma_{z_n} q(z_n) \log p(x|\theta) + \Sigma_{z_n} q(z_n) log\frac{p(z_n|x, \theta)}{q(z_n)} \\ &= \log p(x|\theta) - KL[q(Z)||p(Z|x, \theta)]\end{aligned} \tag{17}$$

Therefore, a single factor of the log-likelihood has a lower bound as much as the result of KL divergence. Thus we need to choose $q(Z)$ is same as $p(Z|x, \theta)$ to make KL divergence as small as possible. This will be called **E-step**.

In the E-step, we need to compute the posterior over hidden variables, using the current estimate of parameter, θ^{old} Then we should calculate the expected complete data log-likelihood $\mathbb{E}_{p(Z|X,\theta^{old})}[\log p(X, Z|\theta)]$. From this process we can achieve the marginalized probability distribution over z.

After E-step, the parameter θ should be update in order to maximize the expected complete data log-likelihood, which is called **M-step**.

$$\begin{aligned}\theta^{new} &= argmax_{\theta} \mathbb{E}_{p(Z|X,\theta^{old})}[\log p(X, Z|\theta)] \\ &= argmax_{\theta} \int p(Z|X, \theta^{old})[\log p(X, Z|\theta)]dz\end{aligned}$$

Following will cover to find the MLE in Mixture of Gaussian with utilizing EM optimization.

First of all, we define parameters as $\theta = \{\mu, \Sigma\}$, π, and observed variables X following a certain gaussain distribution $\mathcal{N}(x_k|\mu_k, \Sigma_k)$. And Latent variables $z_n \in \{0, 1\}^K$. In the sense of one-hot encoding for z_n, we can define π_k as $p(z_{k,n} = 1)$. Then, we can notice that,

$$p(x_n|z_n,\theta)=\prod_{k=1}^{K}\mathcal{N}(x_n|\mu_k,\sigma_k)^{z_{k,n}} \tag{18}$$

$$p(z_n|\pi)=\prod_{k=1}^{K}\pi_k^{z_{k,n}} \tag{19}$$

Therefore, if we consider X i.i.d, then the complete data log likelihood is given by

$$\begin{aligned}\mathcal{L}_c&=\log p(X,Z|\theta,\pi)\\&=\sum_{n=1}^{N}\log p(x_n,z_n|\theta,\pi)\\&=\sum_{n=1}^{N}\log p(x_n|z_n|\theta)p(z_n|\pi)\\&=\sum_{n=1}^{N}\log\,(\prod_{k=1}^{K}\pi_k^{z_{k,n}}\mathcal{N}(x_n|\mu_k,\Sigma_k)^{z_{k,n}})\\&=\sum_{n=1}^{N}\sum_{k=1}^{K}z_{k,n}\log\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)\end{aligned} \tag{20}$$

By the result of the derivation above, **E-step** of MoG is to start with computing the expected complete-data log-likelihood w.r.t the posterior distribution over hidden variables $p(Z|X)$:

$$\mathbb{E}_{p(Z|X)}[\mathcal{L}_c]=\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{p(Z|X)}\Big[z_{k,n}\log\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)\Big] \tag{21}$$

Since $p(Z|X,\theta,\pi)\propto p(Z|\pi)p(X|Z,\theta)=\prod_{n=1}^{N}\prod_{k=1}^{K}(\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k))^{z_{n,k}}$ and hence factorizes over n so that under the posterior distribution the z_n are independent. Therefore

$$\begin{aligned}\mathbb{E}_{p(Z|X)}[z_{k,n}]&=p(z_{k,n}=1|x_n)\\&=\frac{p(x_n|z_{k,n}=1)\pi_k}{\sum_{k'=1}^{K}p(x_n|z_{k',n}=1)\pi_{k'}}\\&=\frac{\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{k'=1}^{K}\mathcal{N}(x_n|\mu_{k'},\Sigma_{k'})\pi_{k'}}\\&=r_{k,n}\end{aligned} \tag{22}$$

Thus,

$$\begin{aligned}\mathbb{E}_{p(Z|X)}[\mathcal{L}_c]&=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{k,n}\log\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)\\&\propto\sum_{n=1}^{N}\sum_{k=1}^{K}r_{k,n}\big[\log\pi_k-\frac{1}{2}\log|\Sigma_k|-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\big]\end{aligned} \tag{23}$$

In the **M-step**, parameters θ and π are updated such that the expected complete-data log-likelihood is maximized:

(1)μ_k

Similar to the way we did in the problem 1, by using *equation (81)* in [3]:

$$\frac{\partial \mathbb{E}_{p(Z|X)}[\mathcal{L}_c]}{\partial \mu_k} = \sum_{n=1}^{N} r_{n,k} \Sigma^{-1}(x_n - \mu_k) = 0 \tag{24}$$

Then

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} r_{k,n} x_n}{\sum_{n=1}^{N} r_{k,n}} \tag{25}$$

(2)Σ_k

By using *equation (58)* and *equation (70)* in [3]:

$$\frac{\partial \mathbb{E}_{p(Z|X)}[\mathcal{L}_c]}{\partial \Sigma_k^{-1}} = \frac{1}{2}\big(\sum_{n=1}^{N} r_{n,k}(\Sigma^T - (x_n - \mu_k)(x_n - \mu_k)^T)\big) = 0 \tag{26}$$

Thus,

$$\Sigma_k^{new} = \frac{\sum_{n=1}^{N} r_{n,k}(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T)}{\sum_{n=1}^{N} r_{k,n}} \tag{27}$$

(3) π_k

This can be regarded as constrain optimization problem since $\Sigma_{k=1}^{K}\pi_k = 1$. So apply Lagrange multiplier(λ) with the objective function then:

$$\mathcal{L}' = \mathbb{E}_{p(Z|X)}[\mathcal{L}_c] + \lambda[1 - \Sigma_{k=1}^{K}\pi_k] \tag{28}$$

Therefore, $\nabla_{\pi_k,\lambda}\mathcal{L}' = 0$ needs to be calculated:

$$\frac{\partial \mathcal{L}'}{\partial \lambda} = 1 - \Sigma_{k=1}^{K}\pi_k = 0 \tag{29}$$

$$\frac{\partial \mathcal{L}'}{\partial \pi_k} = \sum_{n=1}^{N} r_{k,n}\frac{1}{\pi_k} - \lambda = 0 \tag{30}$$

Then $\Sigma_{n=1}^{N} r_{k,n} = \pi_k \lambda$. Since $\sum_{k=1}^{K} \pi_k = 1$, $\lambda = N$. Thus, we have:

$$\pi_k^{new} = \frac{\sum_{k=1}^{K} r_{k,n}}{N} = \frac{N_k}{N} \tag{31}$$

Problem 3. Implement EM-MoG and run your algorithm on toy dataset (for instance, you generate 2-dimensional examples from three Gaussians (with different means and covariance matrices) and determine three clusters using EM-MoG)

Solution E-step was calculated firstly by equation (22) and M-step was deployed with equation (25), (27) and (31)

Algorithm 1: EM - MoG

```
input  : Dataset $\mathcal{D} = \{x_1, ..., x_N\}$
output: $\pi_k, \mu_k, \Sigma_k, r_{n,k}$ s.t $k = \{1, ..., K\}$, $n = \{1, ..., N\}$
Choose an initial setting for the parameters $\theta^{old}, \pi^{old}$
while NOT converging OR iteration < max_iteration do
    $r_{k,n} = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}$
    $\mu_k = \frac{\sum_{n=1}^{N} r_{k,n} x_n}{\sum_{n=1}^{N} r_{k,n}}$
    $\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{k,n}(x_n - \mu_k)(x_n - \mu_k)^T$
    $\pi_k = \frac{N_k}{N}$
end
Return
```

Toy data set was generated with three different Gaussian distributions as belows. The first one has $\mu_A = (2, 3)$, $\Sigma_A = (1, 1.5; 1.5, 3)$ and the number of samples from this distribution is 100. The second one has $\mu_B = (-1, -1)$, $\Sigma_B = (2, -1.5; -1.5, 2)$ and the number of samples from this distribution is 60. And the last one has $\mu_C = (4, 1)$, $\Sigma_C = (1, -1.5; -1.5, 3)$ and the number of samples from this distribution is 80. Therefore $\pi_1 = 0.416$, $\pi_2 = 0.25$ and $\pi_3 = 0.333$.

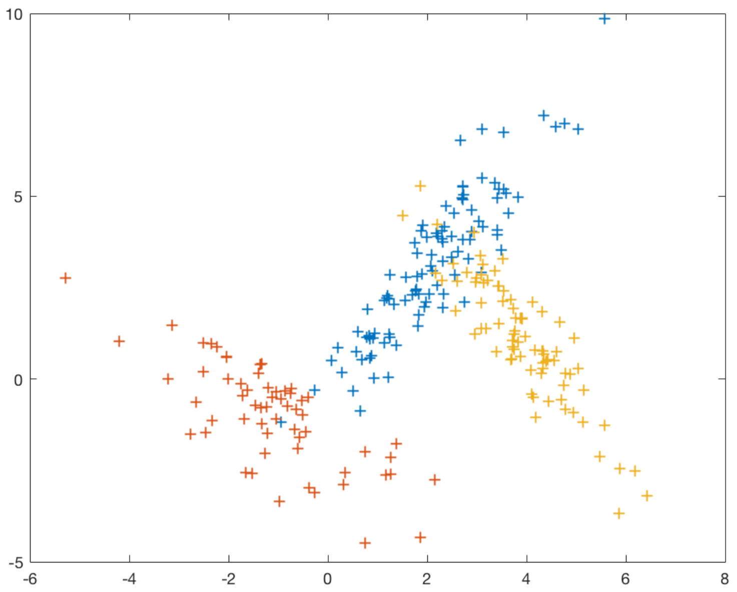
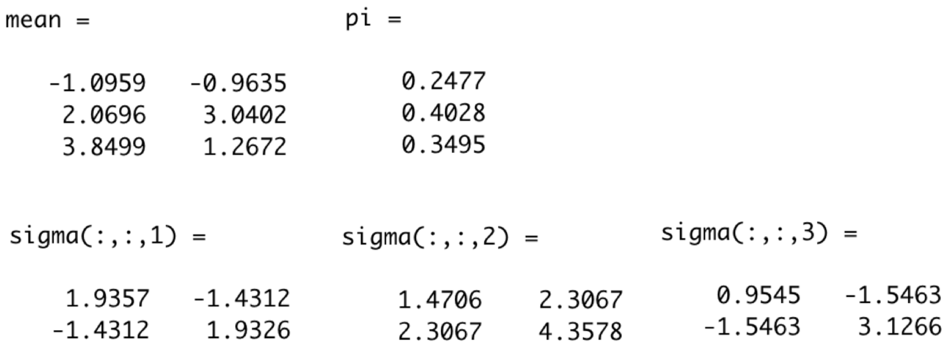


Figure 1: Toy Data Set

The result of EM optimization is follows:



```
mean =                         pi =

   -1.0959   -0.9635               0.2477
    2.0696    3.0402               0.4028
    3.8499    1.2672               0.3495


sigma(:,:,1) =             sigma(:,:,2) =             sigma(:,:,3) =

    1.9357   -1.4312            1.4706    2.3067           0.9545   -1.5463
   -1.4312    1.9326            2.3067    4.3578          -1.5463    3.1266
```

Figure 2: Result of EM

From the result, we can notice that cluster 1 of result is showing that B distribution, cluster 2 is A distribution and cluster 3 is C distribution. The results do not find exact mean and covariance values of the given dataset, however, it can work very well in the task. From figure 3 to figure 5, each clusters were plotted considering its responsibilities. For example, figure 3 shows that the brighter the marker is, the higher probability the cluster has. A certain overlapped area in the middle can have a bit lower probability, which has a bit darker color, however, the predicted clustered has a nice fitting distribution on overall data points.

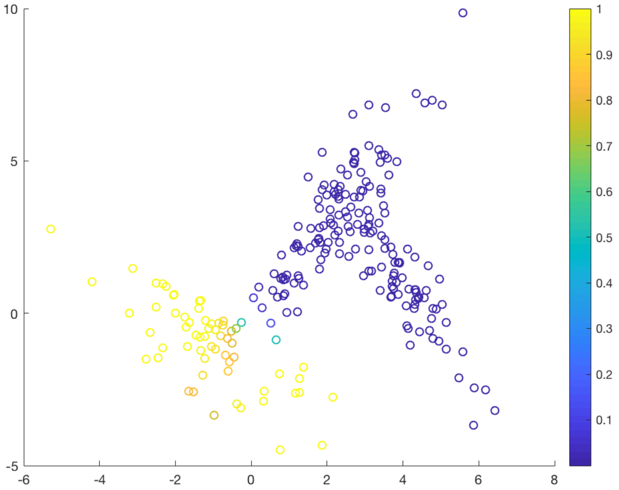


Figure 3: Cluster 1 = B dist

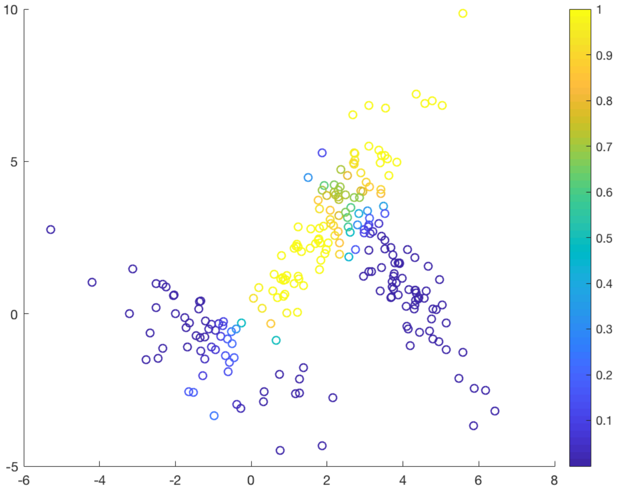


Figure 4: Cluster 2 = A dist

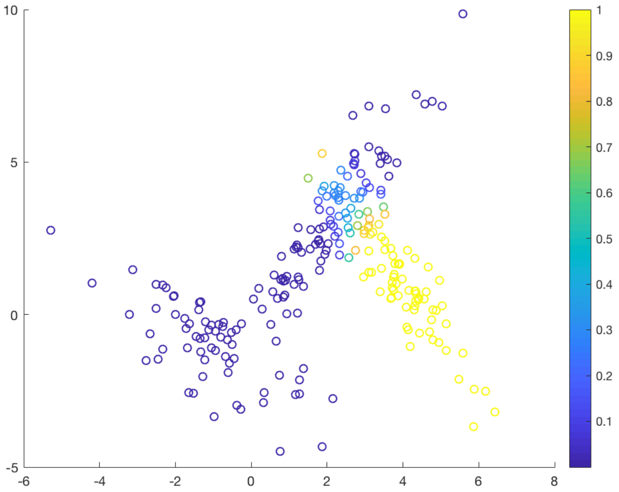


Figure 5: Cluster 3 = C dist

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning.*

[2] Seungjin Choi. *Lecture Note.*

[3] Kaare Brandt Petersen. *The Matrix Cookbook.*