Solutions for Week 5 Problems

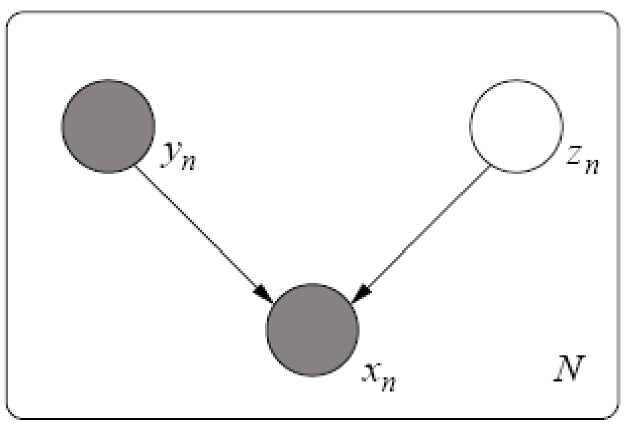
Compiled by John Nangle and Siou Lin

1. *Solution by schlecht*

   Predicting whether or not a particular airline flight will be on time, and by how much, is a very practical and important problem. For a particular flight, there are a number of obvious factors that determine its punctuality; these include airline, aircraft type, weather, departure location, and arrival location. These factors are all directly observable and should be available prior to the flight's departure in most cases. However, there is another factor that I believe influences how on-time a flight is: the experience and *goodness* of its crew.

   In my experience, some pilots are better than others at managing the affairs of an on-time arrival, particularly in events of bad weather, delays, or mechanical malfunction. I don't believe it is necessarily true that a pilot with more experience will be better than a lesser experienced one at punctuality. Because of this, I think that the goodness of a pilot is not directly observable and acts as a latent variable for explaining the differences between on-time arrivals.

   My model would be as follows. Let X be the difference of the actual arrival times for a set of *N* flights from their advertised times. The goal for a pilot is to have these values zero. Let Y be the observable random events listed above, such as weather, aircraft, *etc*. Finally, let Z be the *skill* of the pilot. The graphical model representing the joint distribution over all these is given by



2. Solution by taralove

Deriving eq. 9.17: from eq. 9.16:

$$0 = -\sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_{\mathbf{k}}, \Sigma)}{\sum_j N(x_n|\mu_{\mathbf{j}}, \Sigma_{\mathbf{j}})} \Sigma_k (x_n - \mu_{\mathbf{k}})$$

Multiplying from the right by Σ_k^{-1}, assuming it is not singular, and letting

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n|\mu_{\mathbf{k}}, \Sigma)}{\sum_j N(x_n|\mu_{\mathbf{j}}, \Sigma_{\mathbf{j}})}$$

we get:

$$\begin{aligned}
0 &= -\sum_{n=1}^{N} \gamma(z_{nk}) \Sigma_k (x_n - \mu_{\mathbf{k}}) \Sigma_{\mathbf{k}}^{\mathbf{-1}} && \Leftrightarrow \\
0 &= -\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_{\mathbf{k}}) && \Leftrightarrow \\
0 &= -\sum_{n=1}^{N} \gamma(z_{nk}) x_n + \sum_{n=1}^{N} \gamma(z_{nk}) \mu_{\mathbf{k}}) && \Leftrightarrow \\
0 &= -\sum_{n=1}^{N} \gamma(z_{nk}) x_n + \mu_{\mathbf{k}} \sum_{\mathbf{n=1}}^{\mathbf{N}} \gamma(\mathbf{z_{nk}})) && \Rightarrow \\
\mu_k &= \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})} && \Rightarrow \\
\mu_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n
\end{aligned}$$

where we have defined $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.

solution by Abhishek

To get (9.19), differentiate (9.14) wrt Σ_k^{-1} to get

$$0 = \sum_{n=1}^{N} \frac{\pi_k}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)} \frac{\partial}{\partial \Sigma_k^{-1}} N(x_n|\mu_k, \Sigma_k) \tag{2.1}$$

$$N(x|\mu, \Sigma) = |\Sigma|^{-1/2}(2\pi)^{-k/2} exp \frac{-1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)$$

$$\begin{aligned}
\frac{\partial}{\partial \Sigma^{-1}} N(x|\mu, \Sigma) &= (2\pi)^{-k/2} \frac{\partial}{\partial \Sigma^{-1}} |\Sigma^{-1}|^{1/2} exp \frac{-1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) \\
&+ (2\pi)^{-k/2}|\Sigma|^{-1/2} exp \frac{-1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) \times \frac{-1}{2}(x-\mu)(x-\mu)' \\
&= N(x|\mu, \Sigma) \left(\frac{-1}{2}(x-\mu)(x-\mu)' + |\Sigma|^{-1/2} \frac{\partial}{\partial \Sigma^{-1}} |\Sigma^{-1}|^{1/2} \right)
\end{aligned} \tag{2.2}$$

$$\frac{\partial}{\partial A}|A| = |A|(A')^{-1}$$

So $\frac{\partial}{\partial \Sigma^{-1}} |\Sigma^{-1}|^{1/2} = \frac{1}{2} |\Sigma|^{1/2} |\Sigma^{-1}| \Sigma$

So (2.2) becomes

$$\frac{-1}{2} N(x|\mu, \Sigma) \left((x-\mu)(x-\mu)' - \Sigma \right)$$

Hence (2.1) is

$$0 = \sum_{n=1}^{N} \frac{\pi_k}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)} \frac{-1}{2} N(x_n|\mu_k, \Sigma_k) \left((x_n-\mu_k)(x_n-\mu_k)' - \Sigma_k \right) \tag{2.3}$$

$$\text{or } 0 = \sum_{n=1}^{N} \gamma(z_{nk}) \left((x_n-\mu_k)(x_n-\mu_k)' - \Sigma_k \right) \tag{2.4}$$

(2.4) will give expression for Σ_k as in 9.19.

Setting this to zero gives:

$$
\begin{aligned}
\frac{d}{d\Sigma_k} ln(X|\pi,\mu,\Sigma) = 0 \qquad\qquad & \Leftrightarrow \\
-\frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})\big[|\Sigma_k|^{-1} + (x_n-\mu_k)^T(x_n-\mu_k)\big] = 0 \quad & \Leftrightarrow \\
\sum_{n=1}^{N}\gamma(z_{nk})|\Sigma_k|^{-1} + \sum_{n=1}^{N}(x_n-\mu_k)^T(x_n-\mu_k) = 0 \quad & \Leftrightarrow \\
\Sigma_k^{-1} = \frac{1}{N_k}\sum_{n=1}^{N}(x_n-\mu_k)^T(x_n-\mu_k) &
\end{aligned}
$$

which is eq. 9.19.

Deriving eq. 9.22: taking the derivative of eq. 9.20 with respect to π_k:

$$
\begin{aligned}
\frac{d}{d\pi_k}\Big[ln(p(X|\pi,\mu,\Sigma) + \lambda\Big(\sum_{k=1}^{K}\pi_k - 1\Big)\Big] \qquad\qquad & = \\
\frac{d}{d\pi_k}\Big[\sum_{n=1}^{N} ln\Big\{\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)\Big\} + \lambda\Big(\sum_{k=1}^{K}\pi_k - 1\Big)\Big] \quad & = \\
\sum_{n=1}^{N}\frac{1}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)} N(x_n|\mu_k,\Sigma_k) + \lambda &
\end{aligned}
$$

Setting the above to zero and multiplying by π_k and summing over k, we get:

$$
\begin{aligned}
\sum_{n=1}^{N} \frac{1}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)} N(x_n|\mu_k,\Sigma_k) + \lambda = 0 \quad &\Leftrightarrow \\
\sum_{n=1}^{N} \frac{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)} + \lambda \sum_{k=1}^{K} \pi_k = 0 \quad &\Leftrightarrow \\
\sum_{n=1}^{N} 1 + \lambda = 0 \quad &\Rightarrow \\
\lambda = -N &
\end{aligned}
$$

since $\sum_{k=1}^{K} \pi_k = 1$. Now only multiplying by π_k (without summing over k), we get:

$$
\sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)} - N\pi_k = 0 \Leftrightarrow
$$

$$
\sum_{n=1}^{N} \gamma(z_{nk}) - N\pi_k = 0 \Rightarrow
$$

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

which is eq. 9.22.

3. *Solution by ruizhang*

In this case, given $p(\theta|\mathbf{X})$

$$
\begin{aligned}
p(\theta|\mathbf{X}) &= \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})} \\
\ln p(\theta|\mathbf{X}) &= \ln\{\frac{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)p(\theta)}{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z})}\}
\end{aligned}
$$

So in the E step, similar to maximum likelihood situation, the latent variables $\mathbf{Z}$ is still given by the posterior distribution $p(\mathbf{Z}|\mathbf{X},\theta)$, therefore this step remains the same.
In M step of evaluation of θ^{new}, since $\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z})$ has nothing to do with θ, it can be ignored. So $\ln\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\theta)p(\theta)\}$ is left for evaluating. Similarly, according to equation (9.30) in the text, the component to be evaluated is $\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln(p(\mathbf{X},\mathbf{Z}|\theta)) + \ln p(\theta)$

4. Solution by icrk

By visually examining the input data, it is somewhat clear that there are three clusters. Two strong clusterings at the bottom and one diffuse cluster at the top.

Setting the number of clusters to $K = 3$ in EM produces the correct clustering:

By sufficiently increasing the number of clusters it is possible that some clusters will not have representative points. Trivially, if we have more clusters than points, we can expect that while each cluster may have some responsibility to one or more points, the points will be assigned to at most a number of clusters that is equal to the number of points. Trivial case aside, it is possible to have a number of clusters smaller than the number of data points and still have clusters that are not represented. There are no guarantees that every cluster will be the one with the highest probability for any point. So, while a cluster may retain some responsibility for one or more points, it may lose out to other clusters. Running EM with $k = 20$ shows that of the 20 clusters, only 5 or 6 are observed.

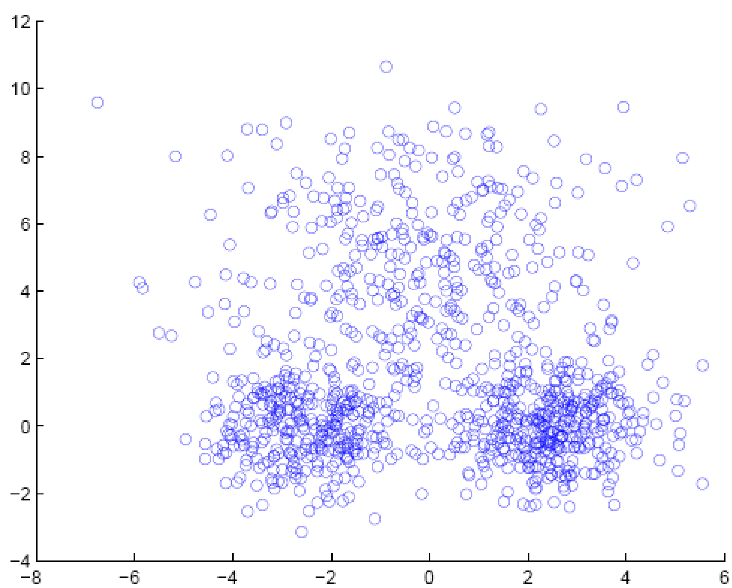


Figure 2: Clustering input data.

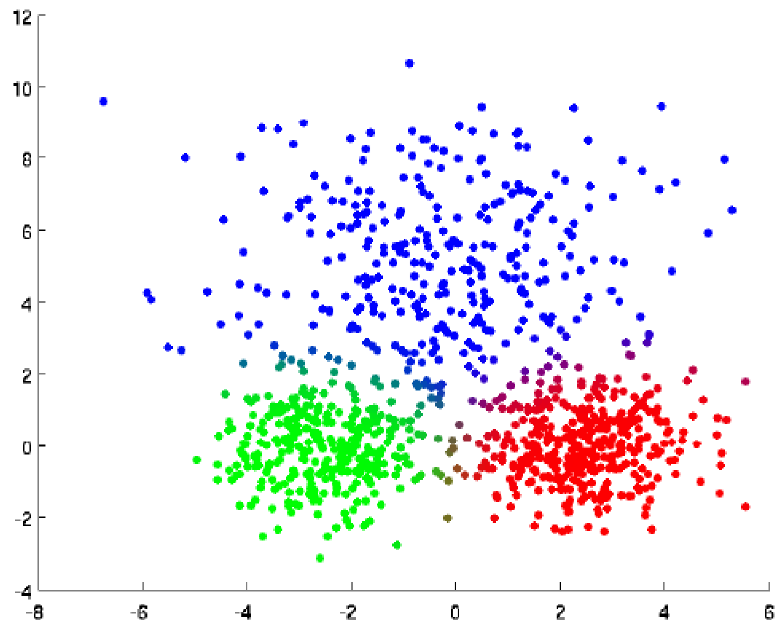


Figure 3: Clustered input data, blended colors.

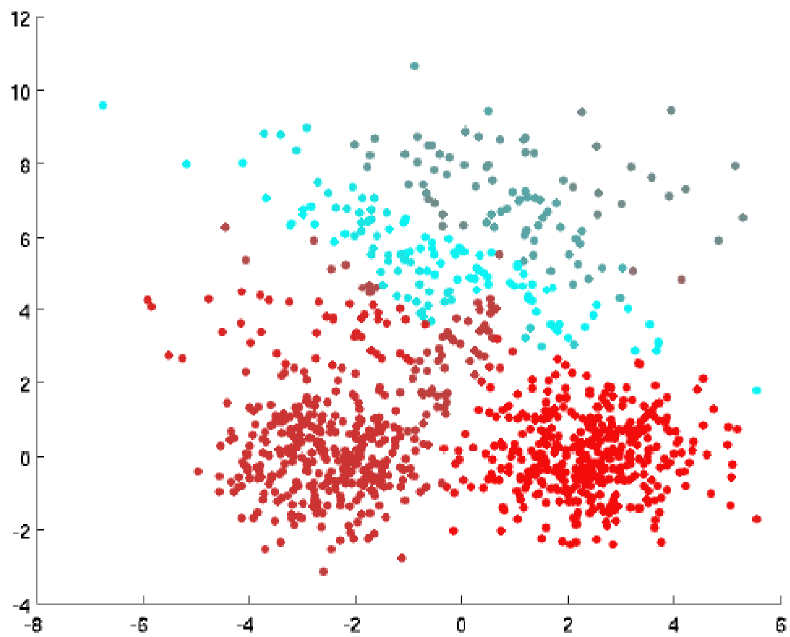


Figure 4: k=20 clustering.

My Matlab implementation of the EM algorithm is below, the log likelihood output (each five iterations) is below that.

```
load -ascii data/a5_data.txt

%initialization
%%%%%%%%%%%%%%%%%%%%%%%%%%%
k=3;          %how many clusters
maxiter=100; %how many iterations?

%M=moviein(maxiter);
%set(gca, 'NextPlot', 'replacechildren')
%randomly assign data to clusters
for i=1:1000;
a5_data(i,3)=mod(i,k)+1;
end;


%mean
mu_k=zeros(k,2);
for i=1:1000;
```

```
mu_k(cast(a5_data(i,3),'uint8'),1)=mu_k(cast(a5_data(i,3),'uint8'),1)+1;
end;

sum=zeros(k,2);
numthings=0;
temp_mu_k=zeros(k,2)
for i=1:3;
for j=1:1000;
if cast(a5_data(j,3), 'uint8')==i;
            sum(i,1)=sum(i,1)+a5_data(j,1);
            sum(i,2)=sum(i,1)+a5_data(j,2);
            numthings=numthings+1;
end;
end;
    temp_mu_k(i,1)=sum(i,1)/numthings;
    temp_mu_k(i,2)=sum(i,2)/numthings;
end;
mu_k=temp_mu_k;

%mixing coeff (a_1,...,a_k) s.t. a_n=1000/k
for i=1:k;
pi_k(i)=1000/k;
end;
gamma_znk=zeros(1000,k);
for i=1:1000;
    gamma_znk(i,cast(a5_data(i,3),'uint8'))=1;
end;

N_k=zeros(1,k);
for n=1:k;
    for i=1:1000;
        N_k(1,n)=N_k(1,n)+gamma_znk(i,n);
    end;
end;

%covariance
sum=zeros(2)
for i=1:k;
    for j=1:1000;
       sum=sum+gamma_znk(j,i)*(a5_data(j,1:2)-mu_k(i,:))'*(a5_data(j,1:2)-mu_k(i,:));
    end;
    sigma_k(:,:,i)=(1/N_k(i))*sum;
end;

for iter=1:maxiter;
%Expectation
%%%%%%%%%%%%%%%%%%%%%%%%%%
a=(1/(2*pi));
```

```
E_denom=0;
for i=1:1000;
    for j=1:k;
        E_denom=E_denom+(pi_k(j)*a*(1/(det(sigma_k(:,:,j))^(.5)))*
exp(-.5*((a5_data(i,1:2)-mu_k(j,:))*inv(sigma_k(:,:,j))*(a5_data(i,1:2)-mu_k(j,:))')));
    end;
    for j=1:k;
        E_num=pi_k(j)*a*(1/(det(sigma_k(:,:,j))^(.5)))*
exp(-.5*((a5_data(i,1:2)-mu_k(j,:))*inv(sigma_k(:,:,j))*(a5_data(i,1:2)-mu_k(j,:))'));
        gamma_znk(i,j)=E_num/E_denom;
end;
    E_denom=0;
end;
%Maximization
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%N_k
N_k(1,:)=zeros(1,k);
for n=1:k;
    for i=1:1000;
        N_k(1,n)=N_k(1,n)+gamma_znk(i,n);
    end;
end;

%mu_k(k,2)
sum=zeros(1,2);
for i=1:k;
    for j=1:1000;
        sum=sum+(gamma_znk(j,i)*a5_data(j,1:2));
    end;
    mu_k(i,:)=(1/N_k(i))*sum;
    sum=zeros(1,2);
end;

%sigma_k
sum=zeros(2,2);
for i=1:k;
    for j=1:1000;
        sum=sum+gamma_znk(j,i)*((a5_data(j,1:2)-mu_k(i,:))'*(a5_data(j,1:2)-mu_k(i,:)));
    end;
    sigma_k(:,:,i)=(1/N_k(i))*sum;
    sum=zeros(2,2);
end;

for i=1:k;
    pi_k(i)=N_k(i)/1000;
end;

%Log likelihood
```

```
a=1/(2*pi);
sum=0;
LL=0;
for i=1:1000;
    for j=1:k;
        sum=sum+(pi_k(j)*a*(1/(det(sigma_k(:,:,j))^(.5)))*
exp(-.5*(a5_data(i,1:2)-mu_k(j,:))*inv(sigma_k(:,:,j))*(a5_data(i,1:2)-mu_k(j,:))'));
    end;
    LL=LL+log(sum);
    sum=0;
end;

LL

%M(iter)=getframe;
end;

%since k=3, use gamma to blend
scatter(a5_data(:,1), a5_data(:,2), 20, gamma_znk, 'filled')
```

Iteration	Log Likelihood
1	-4.7355e+03
5	-4.5372e+03
10	-4.5140e+03
15	-4.5097e+03
20	-4.5067e+03
25	-4.5039e+03
30	-4.5026e+03
35	-4.5020e+03
40	-4.5015e+03
45	-4.5005e+03
50	-4.4976e+03
55	-4.4916e+03
60	-4.4752e+03
65	-4.4099e+03
70	-4.3318e+03
75	-4.3277e+03
80	-4.3275e+03
85	-4.3275e+03
90	-4.3275e+03
95	-4.3275e+03
100	-4.3275e+03