Statistical Computing for Scientists and Engineers

Final Homework

Jiale Shi

Dec/10/2018

1 EM algorithm

Implement the EM algorithm for estimating the parameters of a mixture of Gaussians with isotropic covariances using the data provided on data resources. There are two data sets each of which is two-dimensional. You can write your own or use any available code for mixture of Gaussians (e.g. you ca use the code in the code director with some changes to account for the isotropic covariances. Also see the accompanying paper Unsupervised Learning of Finite Mixture Models, M.Figueiredo and A.K. Jain.)

Experiment with the number of mixtures and comment on the trade-off between the number of mixture and goodness of fit (i.e. log-likelihood) of the data. Plot the log-likelihood as a function of the number of components of a mixture of Gaussians to support your argument.

Find a fixed number of Gaussians that works well for each data set.

Plot the estimated Gaussians as one-sigma contours of each mixing component on top of the training data.

List the mean, covariance and mixing weights of each mixture component.

Solution: Implement the EM algorithm for estimating the parameters of a mixture of Gaussians with isotropic covariances. The isotropic covariances maxtrix means the covariance matrix is diagonal and all elements on the diagonal is equal.

$$\mathbf{C}_{\text{isotropic}} = \lambda \mathbf{I} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \tag{1}$$

The matlab code (covoption=1) can deal with diag covariances matrix. For diag cov:

$$\mathbf{C}_{\text{diag}} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \tag{2}$$

$$\begin{aligned} &\text{For diag cov} \\ &\sigma_{ik}^2 = \frac{\sum_n \frac{1}{N}(x_{in} - \mu_{ik})^2}{\sum_{n=1}^{N} \frac{1}{N}} \qquad\qquad i = 1,2 \\ &\text{For isotropic cov} \\ &\sigma_k^2 = \frac{\sum_{n=1}^{N} \frac{1}{N} ||\mathbf{x}_n - \mu_k||^2}{2\sum_n \frac{1}{N}} = \frac{\sigma_{1k}^2 + \sigma_{2k}^2}{2} \end{aligned} \tag{3}$$

Therefore, once getting the diag cov $\begin{pmatrix} \sigma_{1k}^2 & 0 \\ 0 & \sigma_{2k}^2 \end{pmatrix}$, I change it based on the diga cov to get isotropic cov.

$$\mathbf{C}_{\text{isotropic},k} = \begin{pmatrix} \sigma_k^2 & 0 \\ 0 & \sigma_k^2 \end{pmatrix} = \begin{pmatrix} \frac{\sigma_{1k}^2+\sigma_{2k}^2}{2} & 0 \\ 0 & \frac{\sigma_{1k}^2+\sigma_{2k}^2}{2} \end{pmatrix} \tag{4}$$

Other steps are the same as diag cov.

For **data1train.dat**:

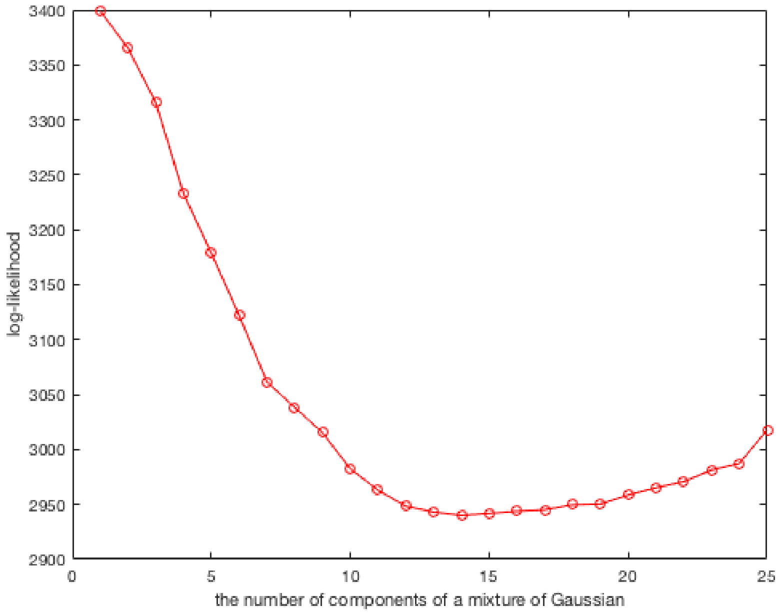

Figure 1: log-likelihood as a function of the number of components of amixture of Gaussians for **data1train.dat**

fixed number of Gaussians that works well for **data1train.dat** is **14**.

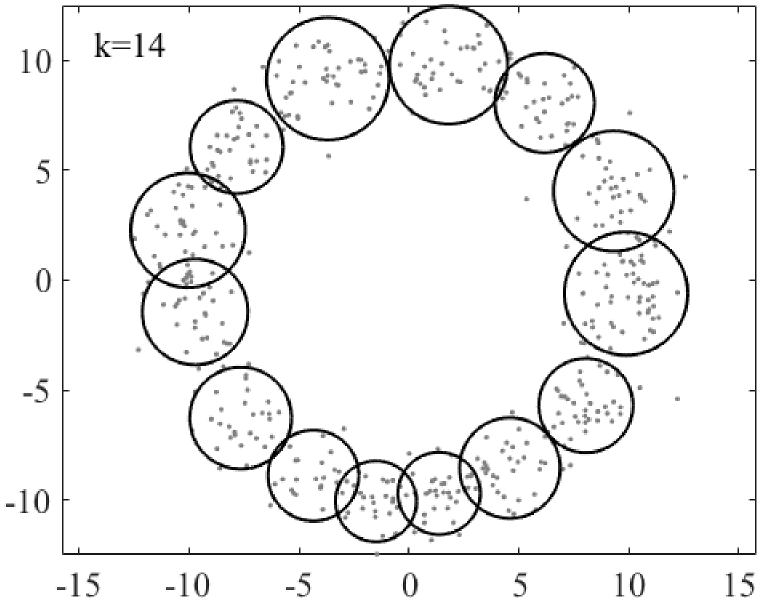

Figure 2: Optimal Gaussians mixtures that work for **data1train.dat**

List the mean, covariance and mixing weights for each mixture component.

best mean	best covariance	best mixing weight
$\begin{pmatrix} 1.3716 \\ -9.6936 \end{pmatrix}$	$\begin{pmatrix} 0.8763 & 0 \\ 0 & 0.8763 \end{pmatrix}$	0.0646
$\begin{pmatrix} 6.1686 \\ 8.0496 \end{pmatrix}$	$\begin{pmatrix} 1.2756 & 0 \\ 0 & 1.2756 \end{pmatrix}$	0.0667
$\begin{pmatrix} -4.3445 \\ -8.8821 \end{pmatrix}$	$\begin{pmatrix} 1.0669 & 0 \\ 0 & 1.0669 \end{pmatrix}$	0.0471
$\begin{pmatrix} -9.7203 \\ -1.4463 \end{pmatrix}$	$\begin{pmatrix} 1.4420 & 0 \\ 0 & 1.4420 \end{pmatrix}$	0.0651
$\begin{pmatrix} 4.5894 \\ -8.5408 \end{pmatrix}$	$\begin{pmatrix} 1.3080 & 0 \\ 0 & 1.3080 \end{pmatrix}$	0.0724
$\begin{pmatrix} 9.3126 \\ 4.0447 \end{pmatrix}$	$\begin{pmatrix} 1.8687 & 0 \\ 0 & 1.8687 \end{pmatrix}$	0.0747
$\begin{pmatrix} -10.0447 \\ 2.2659 \end{pmatrix}$	$\begin{pmatrix} 1.6937 & 0 \\ 0 & 1.6937 \end{pmatrix}$	0.0809
$\begin{pmatrix} -1.4953 \\ -10.0566 \end{pmatrix}$	$\begin{pmatrix} 0.8459 & 0 \\ 0 & 0.8459 \end{pmatrix}$	0.0510
$\begin{pmatrix} 9.8751 \\ -0.6106 \end{pmatrix}$	$\begin{pmatrix} 1.9610 & 0 \\ 0 & 1.9610 \end{pmatrix}$	0.1189
$\begin{pmatrix} -7.6513 \\ -6.2702 \end{pmatrix}$	$\begin{pmatrix} 1.3372 & 0 \\ 0 & 1.3372 \end{pmatrix}$	0.0582
$\begin{pmatrix} 1.8042 \\ 9.7706 \end{pmatrix}$	$\begin{pmatrix} 1.7976 & 0 \\ 0 & 1.7976 \end{pmatrix}$	0.0787
$\begin{pmatrix} 8.0447 \\ -5.6998 \end{pmatrix}$	$\begin{pmatrix} 1.1455 & 0 \\ 0 & 1.1455 \end{pmatrix}$	0.0753
$\begin{pmatrix} -3.6905 \\ 9.1483 \end{pmatrix}$	$\begin{pmatrix} 1.9368 & 0 \\ 0 & 1.9368 \end{pmatrix}$	0.0844
$\begin{pmatrix} -7.8340 \\ 6.0516 \end{pmatrix}$	$\begin{pmatrix} 1.1180 & 0 \\ 0 & 1.1180 \end{pmatrix}$	0.0621

For **data2train.dat**:

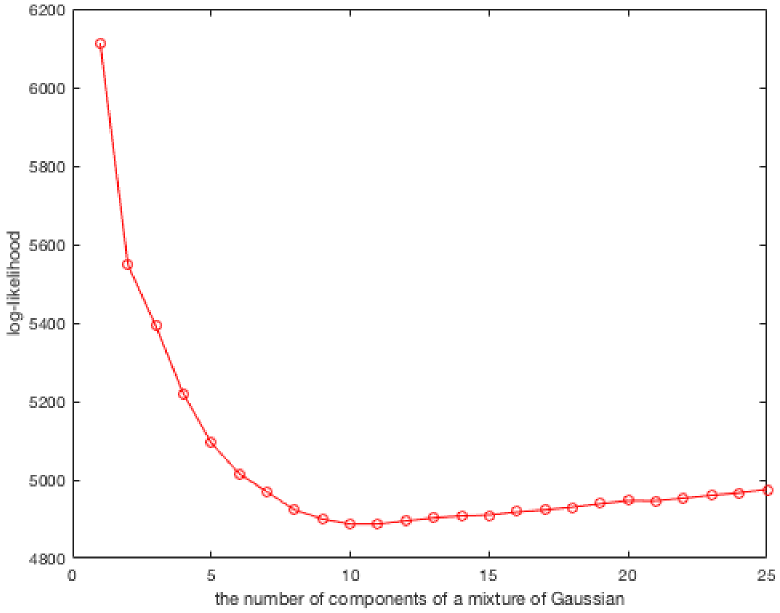

Figure 3: log-likelihood as a function of the number of components of amixture of Gaussians for **data2train.dat**

fixed number of Gaussians that works well for **data1train.dat** is **10**.

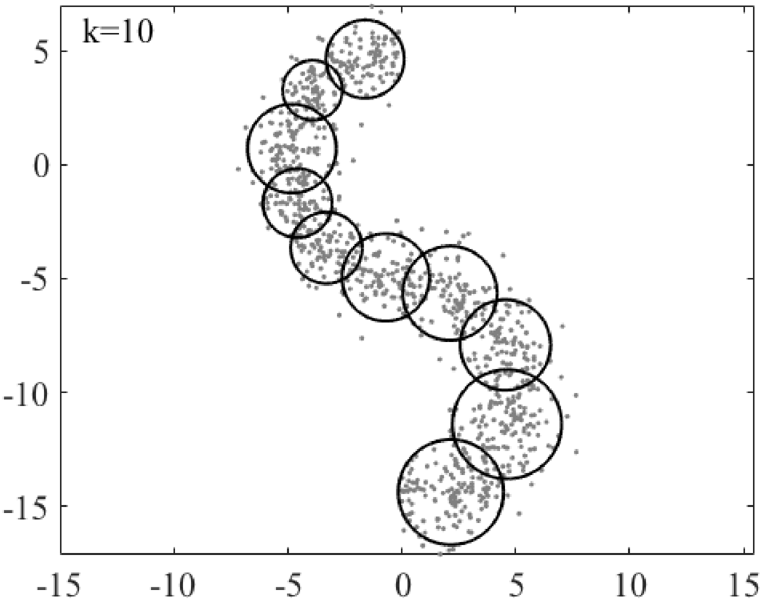

Figure 4: Optimal Gaussians mixtures that work for **data2train.dat**

List the mean, covariance and mixing weights for each mixture component.

best mean	best covariance	best mixing weight
$\begin{pmatrix} -4.5860 \\ -1.6871 \end{pmatrix}$	$\begin{pmatrix} 0.5758 & 0 \\ 0 & 0.5758 \end{pmatrix}$	0.0793
$\begin{pmatrix} 2.1473 \\ -14.3782 \end{pmatrix}$	$\begin{pmatrix} 1.3352 & 0 \\ 0 & 1.3352 \end{pmatrix}$	0.1471
$\begin{pmatrix} -4.8326 \\ 0.7035 \end{pmatrix}$	$\begin{pmatrix} 0.9502 & 0 \\ 0 & 0.9502 \end{pmatrix}$	0.1073
$\begin{pmatrix} -3.9271 \\ 3.2646 \end{pmatrix}$	$\begin{pmatrix} 0.4316 & 0 \\ 0 & 0.4316 \end{pmatrix}$	0.0671
$\begin{pmatrix} -1.6167 \\ 4.6312 \end{pmatrix}$	$\begin{pmatrix} 0.7350 & 0 \\ 0 & 0.7350 \end{pmatrix}$	0.1017
$\begin{pmatrix} -3.3163 \\ -3.6515 \end{pmatrix}$	$\begin{pmatrix} 0.6169 & 0 \\ 0 & 0.6169 \end{pmatrix}$	0.0778
$\begin{pmatrix} 2.1058 \\ -5.6436 \end{pmatrix}$	$\begin{pmatrix} 1.0695 & 0 \\ 0 & 1.0695 \end{pmatrix}$	0.0964
$\begin{pmatrix} 4.6136 \\ -11.3956 \end{pmatrix}$	$\begin{pmatrix} 1.4347 & 0 \\ 0 & 1.4347 \end{pmatrix}$	0.1386
$\begin{pmatrix} 4.5414 \\ -7.9118 \end{pmatrix}$	$\begin{pmatrix} 0.9857 & 0 \\ 0 & 0.9857 \end{pmatrix}$	0.0924
$\begin{pmatrix} -0.7092 \\ -4.9507 \end{pmatrix}$	$\begin{pmatrix} 0.9208 & 0 \\ 0 & 0.9208 \end{pmatrix}$	0.0924

2 Resampling

Randomly generate 100 particles x^i form some distribution π of your choice, and 100 (positive) weights w^i. Normalize the weights such that $\sum_i w^i = 1$, and the weighted samples x^i, w^i to estimate the mean m of π, and denote this estimate by $\hat{m}$.

Resample the particles x^i (from the weights w^i) using multinominal resampling, and estimate the mean from the resampled (now equally weighted) samples. Denote this estimate $\hat{m}_m$.

Repeat this for systematic resampling, and denote this estimate $\hat{m}_s$.

Repeat this for stratified resampling, and denote this estimate $\hat{m}_t$.

Repeat the items above multiple times, and report an estimate of the variance for $\hat{m} - \hat{m}_m$, $\hat{m} - \hat{m}_s$, and $\hat{m} - \hat{m}_t$ respectively, conditionally on $\hat{m}$ (that is, do not sample new particles from π, but only repeat the resampling step). Which resampling scheme appears to be the preferred one, in terms of variance?

Solution:

I randomly generate 100 particles x^i using Gaussian distribution $G(0, 1)$ and calculate weights w^i, normalize the weights by

$$w^i = \frac{w^i}{sum(w^i)} \tag{5}$$

Then I use the same samples and normalized weights to resample multiple times. The algorithms of multinomial resampling, systematic resampling and stratified resampling are in the code (see the attached code)

I repeat resample 1000 times to calculate the variance of $\hat{m} - \hat{m}_m$, $\hat{m} - \hat{m}_s$, and $\hat{m} - \hat{m}_t$.

In this Figure 5, we find that the variance of $\hat{m} - \hat{m}_m$ is the largest. The variance of $\hat{m} - \hat{m}_s$ is smallest.

Analytically, in our simulation, 1000 iterations.

Variance of $\hat{m} - \hat{m}_m$ is 0.005132756202771032

Variance of $\hat{m} - \hat{m}_s$ is 0.0014684267477009338

Variance of $\hat{m} - \hat{m}_t$ is 0.002763182508388841

In terms of variance, systematic resampling is the preferred one.

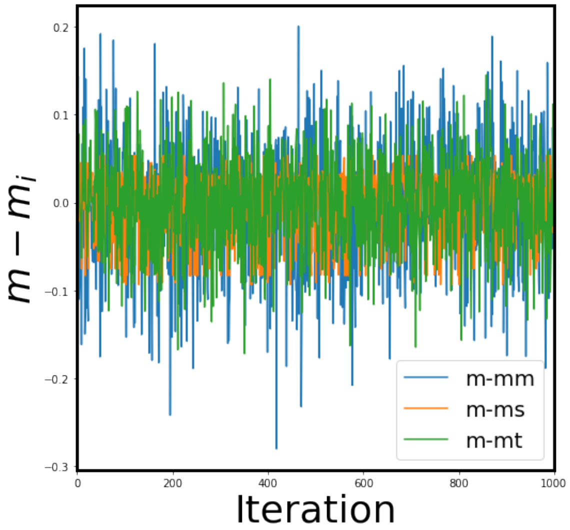

Figure 5: $\hat{m} - \hat{m}_m$, $\hat{m} - \hat{m}_s$, and $\hat{m} - \hat{m}_t$ in 1000 iterations

3 EM algorithm

Consider data,

$$D = \left\{ \begin{pmatrix} 1 \\ 3 \end{pmatrix} \begin{pmatrix} 4 \\ 5 \end{pmatrix} \begin{pmatrix} 2 \\ \star \end{pmatrix} \right\} \tag{6}$$

sampled from a two-dimensional (separable) distribution $p(x_1, x_2) = p_{x_1}(x_1)p_{x_2}(x_2)$ where:

$$p_{x_1}(x_1) = \begin{cases} \frac{1}{\theta_1} \exp\left(-\frac{x_1}{\theta_1}\right) & \text{if} \quad x_1 \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

and

$$p_{x_2}(x_2) = \begin{cases} \frac{1}{\theta_2} & \text{if} \quad 0 \leq x_2 \leq \theta_2, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

and $*$ in the dataset indicates a missing value.

(a) What can you infer from θ_2 by looking at D?

Solution:

The existing two pairs $(x_1$,$x_2)$ $\begin{pmatrix} 1 \\ 3 \end{pmatrix} \begin{pmatrix} 4 \\ 5 \end{pmatrix}$ mean that for $p_{x_2}(x_2)$,$p_3(3) \neq 0$ and $p_5(5) \neq 0$.

$0 \leq 3 \leq 5 \leq \theta_2$

Therefore, θ_2 should be greater or equal to 5.

(b) Start with an initial estimate $\theta^0 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$ and analytically calculate $Q(\theta|\theta^0)$.

This is the expected joint data log-likelihood considered in class. For this problem to compute it, you effectively have to marginalize out the missing values. This is the estimate expectation step in the EM algorithm.

Solution:

$$D = \left\{ \begin{pmatrix} x_{11} = 1 \\ x_{12} = 13 \end{pmatrix} \begin{pmatrix} x_{21} = 4 \\ x_{22} = 5 \end{pmatrix} \begin{pmatrix} x_{31} = 2 \\ x_{32} = \star \end{pmatrix} \right\}$$

$$\theta^0 = \begin{pmatrix} \theta_1^0 = 3 \\ \theta_2^0 = 6 \end{pmatrix} = (\theta_1^0, \theta_2^0)$$

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = (\theta_1, \theta_2)$$

$$\begin{aligned} p(x_1, x_2, \theta) &= p_{x_1}(x_1|\theta) p_{x_2}(x_2|\theta) \\ &= \begin{cases} \frac{1}{\theta_1} \exp\left(-\frac{x_1}{\theta_1}\right) \frac{1}{\theta_2} & \text{if} \quad x_1 \geq 0 \& 0 \leq x_2 \leq \theta_2, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{9}$$

Then we need to use the expected joint data log-likelihood $Q(\theta|\theta^0)$.

$$\begin{aligned} &Q(\theta|\theta^0) = E_{x_{32}}[\ln p(x_v, x_h, \theta)|\mathcal{D}, \theta^0] \\ &= \int_{-\infty}^{\infty} [\ln p(x_{11}, x_{12}|\theta) + \ln p(x_{21}, x_{22}|\theta) + \ln p(x_{31}, x_{32}|\theta)] p(x_{32}|x_{31}, \theta^0) dx_{32} \\ &= \ln p(x_{11}, x_{12}|\theta) + \ln p(x_{21}, x_{22}|\theta) + \int_{-\infty}^{\infty} [\ln p(x_{31}, x_{32}|\theta)] \frac{p(x_{31}, x_{32}|\theta^0)}{\int_{-\infty}^{\infty} p(x_{31}, x_{32}^{'}|\theta^0) dx_{32}^{'}} dx_{32} \end{aligned} \tag{10}$$

$$\begin{aligned} \ln p(x_{11}, x_{12}|\theta) + \ln p(x_{21}, x_{22}|\theta) &= \ln\left(\frac{1}{\theta_1 \theta_2} e^{-\frac{x_{11}}{\theta_1}}\right) + \ln\left(\frac{1}{\theta_1 \theta_2} e^{-\frac{x_{21}}{\theta_1}}\right) \\ &= \ln\left(\frac{1}{\theta_1 \theta_2} e^{-\frac{1}{\theta_1}}\right) + \ln\left(\frac{1}{\theta_1 \theta_2} e^{-\frac{4}{\theta_1}}\right) \\ &= -2 \ln \theta_1 \theta_2 - \frac{5}{\theta_1} \end{aligned} \tag{11}$$

$$p(x_{31} = 2, x_{32}|\theta^0) = p_{x_{31}}(x_{31} = 2|\theta^0) p_{x_{32}}(x_{32}|\theta^0) \tag{12}$$

$$\int_{-\infty}^{\infty} p(x_{31} = 2, x_{32}^{'}|\theta^0) dx_{32}^{'} = p_{x_{31}}(x_{31} = 2|\theta^0) \tag{13}$$

$$\int_{-\infty}^{\infty} [\ln p(x_{31}, x_{32}|\theta)] \frac{p(x_{31}, x_{32}|\theta^0)}{\int_{-\infty}^{\infty} p(x_{31}, x_{32}^{'}|\theta^0) dx_{32}^{'}} dx_{32} = \int_{-\infty}^{\infty} [\ln p(x_{31}, x_{32}|\theta)] p_{x_{32}}(x_{32}|\theta^0) dx_{32} \tag{14}$$

$$[\ln p(x_{31}, x_{32}|\theta)] p_{x_{32}}(x_{32}|\theta^0) \neq 0 \text{ if } 0 \leq x_{32} \leq \min(\theta_2, \theta_2^0 = 6) = \min(\theta_2, 6)$$

$$\int_{-\infty}^{\infty}[\ln p(x_{31},x_{32}|\theta)]p_{x_{32}}(x_{32}|\theta^0)dx_{32} = \int_0^{\min(\theta_2,6)}(-\ln\theta_1\theta_2 - \frac{2}{\theta_1})(\frac{1}{6})dx_{32} \tag{15}$$

for $5 \leq \theta_2 < 6, \theta_2 = \min(\theta_2, 6)$

$$\begin{aligned}&\int_{-\infty}^{\infty}[\ln p(x_{31},x_{32}|\theta)]p_{x_{32}}(x_{32}|\theta^0)dx_{32}\\ &= \int_0^{\theta_2}(-\ln\theta_1\theta_2 - \frac{2}{\theta_1})(\frac{1}{6})dx_{32}\\ &= \frac{\theta_2}{6}(-\ln\theta_1\theta_2 - \frac{2}{\theta_1})\end{aligned} \tag{16}$$

for $6 \leq \theta_2, 6 = \min(\theta_2, 6)$

$$\begin{aligned}&\int_{-\infty}^{\infty}[\ln p(x_{31},x_{32}|\theta)]p_{x_{32}}(x_{32}|\theta^0)dx_{32}\\ &= \int_0^{6}(-\ln\theta_1\theta_2 - \frac{2}{\theta_1})(\frac{1}{6})dx_{32}\\ &= (-\ln\theta_1\theta_2 - \frac{2}{\theta_1})\end{aligned} \tag{17}$$

Therefore, combine (11)(16)(17).

$$\begin{aligned}&\text{for } 5 \leq \theta_2 < 6\\ &Q(\theta|\theta^0) = -2\ln\theta_1\theta_2 - \frac{5}{\theta_1} + \frac{\theta_2}{6}(-\ln\theta_1\theta_2 - \frac{2}{\theta_1}) = -(2+\frac{\theta_2}{6})\ln\theta_1\theta_2 - (5+\frac{\theta_2}{3})\frac{1}{\theta_1}\\ &\text{for } 6 \leq \theta_2\\ &Q(\theta|\theta^0) = -2\ln\theta_1\theta_2 - \frac{5}{\theta_1} + (-\ln\theta_1\theta_2 - \frac{2}{\theta_1}) = -3\ln\theta_1\theta_2 - \frac{7}{\theta_1}\end{aligned} \tag{18}$$

(c) Find the θ that maximizes your $Q(\theta|\theta^0)$. This is the maximization step of the EM algorithm.

Solution:

M step: solving $\bigtriangledown Q(\theta|\theta^0) = 0$ or get the maximization of $Q(\theta|\theta^0)$

We have the boundary conditions for θ_2, so we start from θ_2.

$$\frac{\partial Q(\theta|\theta^0)}{\partial\theta_2} \tag{19}$$

for $5 \leq \theta_2 < 6$

$$\begin{aligned}&\frac{\partial Q(\theta|\theta^0)}{\partial\theta_2} = -\frac{\ln\theta_1}{6} - \frac{1}{3\theta_1} - \frac{2}{\theta_2} - \frac{\ln\theta_2}{6} - \frac{1}{6}\\ &= F(\theta_1) + G(\theta_2) - \frac{1}{6}\end{aligned} \tag{20}$$

$$F(\theta_1) = -\frac{\ln\theta_1}{6} - \frac{1}{3\theta_1} \tag{21}$$

$$\frac{\partial F(\theta_1)}{\partial\theta_1} = \frac{1}{6}(\frac{1}{\theta_1} - \frac{2}{(\theta_1)^2}) = 0 \to \theta_1 = 2 \tag{22}$$

$$F(\theta_1) \leq F(2) = -\frac{1}{6}(\ln 2 + 1) < 0 \tag{23}$$

$$G(\theta_2) = -\frac{2}{\theta_2} - \frac{\ln\theta_2}{6} < 0 \text{for} \quad 5 \leq \theta_2 < 6 \tag{24}$$

Therefore,

$$\begin{aligned} &\text{for } 5 \leq \theta_2 < 6 \\ &\quad \frac{\partial Q(\theta|\theta^0)}{\partial\theta_2} < 0 \end{aligned} \tag{25}$$

$$\begin{aligned} &\text{for } 6 \leq \theta_2 \\ &\quad \frac{\partial Q(\theta|\theta^0)}{\partial\theta_2} = -\frac{3}{\theta_2} < 0 \end{aligned} \tag{26}$$

And $Q(\theta|\theta^0)$ is continuous at $\theta = 6$. Therefore, for $5 \leq \theta_2 < 6$ and $6 \leq \theta_2$, $Q(\theta|\theta^0)$ is monotonically continuous decreasing.

When $\theta_2 = 5$, $Q(\theta|\theta^0)$ gets the largest value $Q(\theta_1, \theta_2 = 5|\theta^0)$. Then we fix $\theta_2 = 5$ and solve θ_1.

$$Q(\theta_1, \theta_2 = 5|\theta^0) = -(\frac{17}{6})\ln 5\theta_1 - (\frac{20}{3})\frac{1}{\theta_1} \tag{27}$$

$\theta_1 > 0$

$$\frac{\partial Q(\theta_1, \theta_2 = 5|\theta^0)}{\partial\theta_1} = -(\frac{17}{6})\frac{1}{\theta_1} + (\frac{20}{3})\frac{1}{(\theta_1)^2} = 0 \to \theta_1 = \frac{40}{17} \tag{28}$$

$$\begin{aligned} &\frac{\partial Q(\theta_1, \theta_2 = 5|\theta^0)}{\partial\theta_1} > 0 \quad \text{if } 0 < \theta_1 < \frac{40}{17} \\ &\frac{\partial Q(\theta_1, \theta_2 = 5|\theta^0)}{\partial\theta_1} < 0 \quad \text{if } \frac{40}{17} < \theta_1 \end{aligned} \tag{29}$$

$Q(\theta_1, \theta_2 = 5|\theta^0)$ gets the largest value when $\theta_1 = \frac{40}{17}$.

In conclusion, the $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \frac{40}{17} \\ 5 \end{pmatrix}$ maximizes $Q(\theta|\theta^0)$.