hw2_sol Haoyuan Du

1.

a.

1. (a).

$$\begin{bmatrix} 0 & 1 & 1 & 2 \\ 2 & 1 & 2 & 3 \end{bmatrix}$$

cov matrix: $S_{xx}: (1^2+0+0+1^2)\frac{1}{3} = \frac{2}{3} \Rightarrow \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}$

$S_{xy}: (0+0+0+1\times1)\frac{1}{3} = \frac{1}{3}$

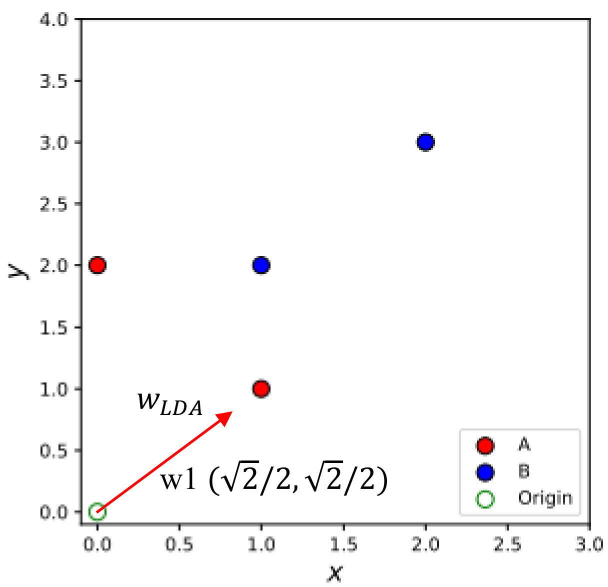
$S_{yy}: (0+(-1)^2+0+1)\frac{1}{3} = \frac{2}{3}$

$\because \det(S-\lambda I)=0$ for eigenvalue

$\therefore \begin{bmatrix} \frac{2}{3}-\lambda & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3}-\lambda \end{bmatrix} \Rightarrow \lambda^2 - \frac{4}{3}\lambda + \frac{1}{3} = 0 \Rightarrow \begin{cases} \lambda_1 = 1 \\ \lambda_2 = \frac{1}{3} \end{cases}$

When $\lambda=1$, eigenVec: $\begin{bmatrix} -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -\frac{1}{3} \end{bmatrix}\begin{bmatrix} v_1^1 \\ v_1^2 \end{bmatrix} = 0 \Rightarrow v_1^1 = v_1^2 \Rightarrow v: \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

So $w_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$



b. No. If we kept all eigen values and vectors in W which is $W^{D\times D}$, then we could reconstruct the original data by Wz^t. However, in this case, W is $W^{D\times d}$ which means we only keep some of the eigenvalues. If we use Wz^t to reconstruct the original data, we can only reconstruct a part of it. Because each of the eigenvectors represents one dimension, we can only reconstruct the dimensions that we have eigenvectors of those dimensions.

c.

S_w = S1 + S2 = (x1-m1)$(\mathrm{x1}-\mathrm{m1})^T$+(x2-m2)$(\mathrm{x2}-\mathrm{m2})^T$=$\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$ +$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$ = $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

S_B = (m1-m2)$(\mathrm{m1}-\mathrm{m2})^T$ = $\begin{bmatrix} -1 \\ -1 \end{bmatrix}$ * [-1 -1] = $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

d. w = c * S_w^{-1}(m1-m2) = c * $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ * $\begin{bmatrix} -1 \\ -1 \end{bmatrix}$ = $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ = $\begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}$

w should be the same as w1 in (1).

2.

a.

$m_1 = 1 \quad m_2 = 10$

(1) Assignment $\because (5-1)^2 < (5-10)^2$

$(6-1)^2 > (6-10)^2$

$\therefore C_1 = \{2, 4, 5\}$

$C_2 = \{6, 8, 10, 12, 14\}$

update: $m_1 = \frac{2+4+5}{3} = \frac{11}{3} \quad m_2 = \frac{6+\ldots+14}{5} = 10$

(2) Assignment: $\because (6-\frac{11}{3})^2 < (6-10)^2$

$(8-\frac{11}{3})^2 > (8-10)^2$

$\therefore C_1 = \{2, 4, 5, 6\}$

$C_2 = \{8, 10, 12, 14\}$

update: $m_1 = \frac{2+4+5+6}{4} = 4.25 \quad m_2 = \frac{8+10+12+14}{4} = 11$

(3) Assignment:

$(6-4.25)^2 < (6-11)^2$

$(8-4.25)^2 > (8-11)^2$

$\therefore C_1 = \{2, 4, 5, 6\}$

$C_2 = \{8, 10, 12, 14\}$

update: same

$\downarrow$

stop

b.

3 iterations. J = $(2-4.25)^2 + (4-4.25)^2 + (5-4.25)^2 + (6-4.25)^2 + (8-11)^2 + (10-11)^2 + (12-11)^2 + (14-11)^2$ = 28.75

c.

(1) Assignment:

$$(8-6)^2 < (8-12)^2$$

$$(10-6)^2 > (10-12)^2$$

$$\therefore C_1 = \{2, 4, 5, 6, 8\}$$

$$C_2 = \{10, 12, 14\}$$

update: $m_1 = \frac{2+\dots+8}{5} = 5 \quad m_2 = \frac{10+12+14}{3} = 12$

(2) Assignment:

$$(8-5)^2 < (8-12)^2$$

$$(10-5)^2 > (10-12)^2$$

$$\therefore C_1 = \{2, 4, 5, 6, 8\}$$

$$C_2 = \{10, 12, 14\}$$

update: same

$\downarrow$

stop

d.

2 iterations. J = $(2-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (8-5)^2 + (10-12)^2 + (12-12)^2 + (14-12)^2$ = 28

e.

The (c) is better, because (c)'s J is less than (a)'s J (28<28.75). Our aim is to minimize the J value.

3.

a. The number of iterations for convergence: 34
Entropy: 0.367
Yes. According to the plot, the convergence appears around 34.

b. 128 dimensions.
Yes.
Because PCA can reduce the dimensions from 784 to 128 which dramatically decrease the cost of calculations. The number of iterations for convergence is 22 instead of 34 which is faster, and the entropy is almost same (0.367 vs. 0.373). Thus, PCA helps clustering.

c. No. The number of iterations for convergence is 93, and the entropy is 0.674. Both are worse than (a) or (b). Because the first principal component only captures a little proportion of the variance so that if we only use the first PC, the data will have only one dimension and the new data would be bad.