R05631027_楊皓文_機器學習_作業#4

Problem 1 (Principal component analysis)

(a) Why did they obtain different results?

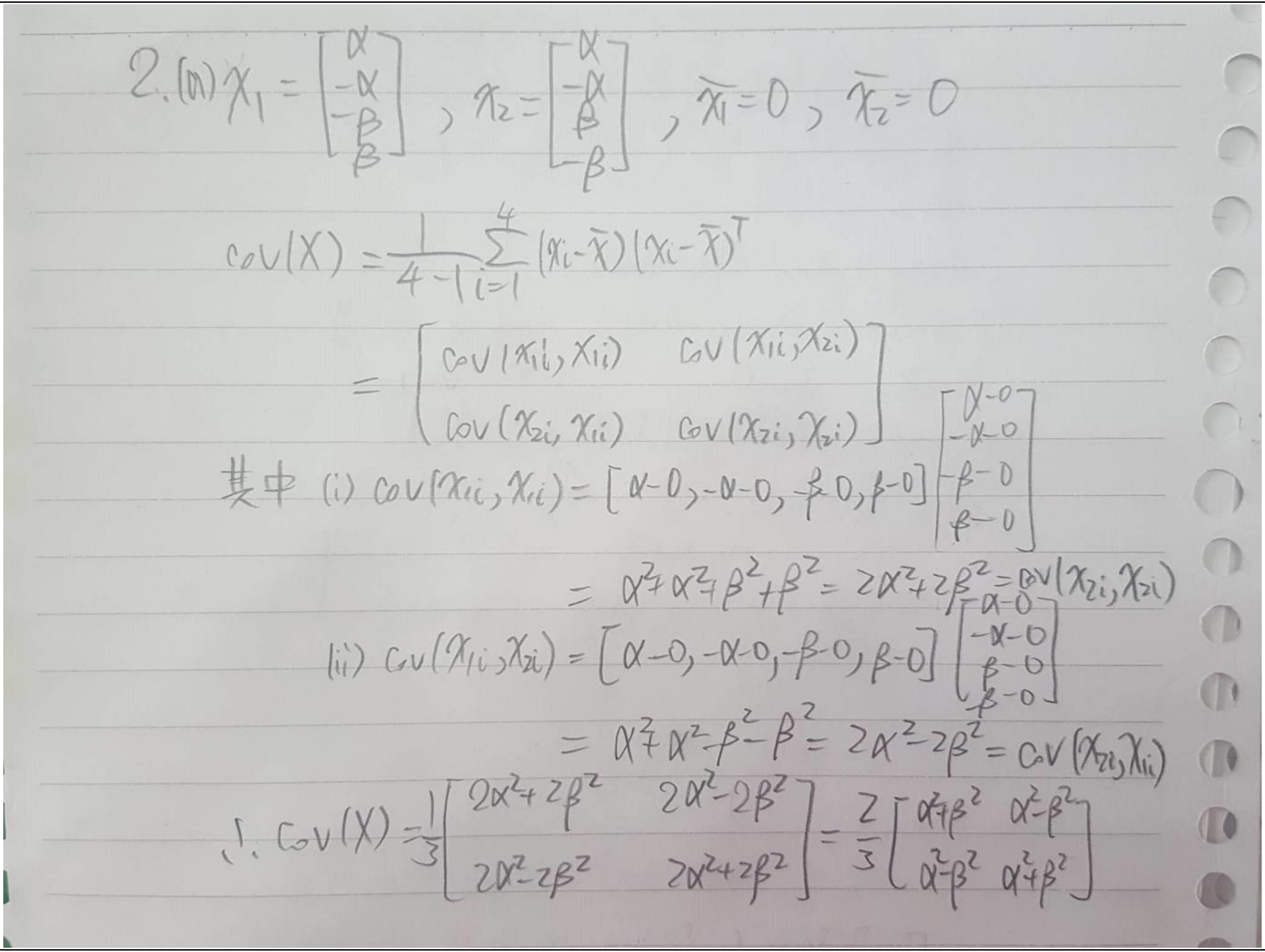
由於Joe和他的boss使用的GDP單位不同， Joe使用millions作為單位，boss使用billions作為單位，且都沒有將數據做歸一化(Normalization)的尺度調整，因此在使用PCA時，GDP的數值就相差了1000倍，造成第一大的eigenvector可以表示的變異數百分比不同。

(b) How should Joe and his Boss have treated the data before the PCA?

可以將數據進行PCA前，先將所有features數值都normalize至[-1 1]或[0 1]，如此一來，各個features的權重就會相等，進行PCA出來的主成分也會較為準確。

Problem 2 (Principal component analysis)

(a) Compute the covariance matrix cov(***X***)



2.(a) $x_1 = \begin{bmatrix} \alpha \\ -\alpha \\ -\beta \\ \beta \end{bmatrix}$, $x_2 = \begin{bmatrix} \alpha \\ -\alpha \\ \beta \\ -\beta \end{bmatrix}$, $\bar{x}_1 = 0$, $\bar{x}_2 = 0$

$$\mathrm{cov}(X) = \frac{1}{4-1}\sum_{i=1}^{4}(x_i-\bar{x})(x_i-\bar{x})^T$$

$$= \begin{bmatrix} \mathrm{cov}(x_{1i}, x_{1i}) & \mathrm{cov}(x_{1i}, x_{2i}) \\ \mathrm{cov}(x_{2i}, x_{1i}) & \mathrm{cov}(x_{2i}, x_{2i}) \end{bmatrix}$$

其中 (i) $\mathrm{cov}(x_{1i}, x_{1i}) = [\alpha-0, -\alpha-0, -\beta-0, \beta-0]\begin{bmatrix} \alpha-0 \\ -\alpha-0 \\ -\beta-0 \\ \beta-0 \end{bmatrix}$

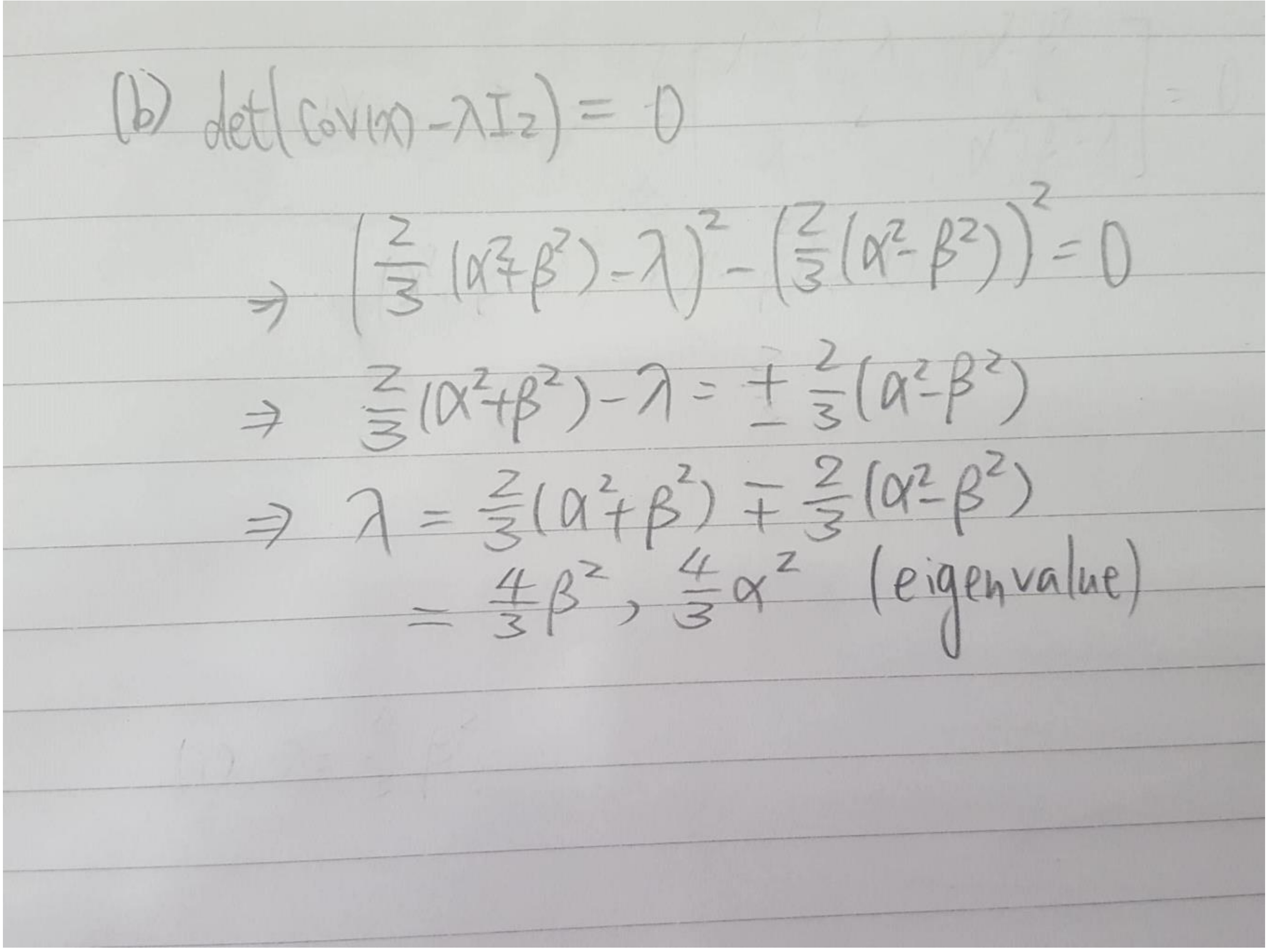
$= \alpha^2+\alpha^2+\beta^2+\beta^2 = 2\alpha^2+2\beta^2 = \mathrm{cov}(x_{2i}, x_{2i})$

(ii) $\mathrm{cov}(x_{1i}, x_{2i}) = [\alpha-0, -\alpha-0, -\beta-0, \beta-0]\begin{bmatrix} \alpha-0 \\ -\alpha-0 \\ \beta-0 \\ -\beta-0 \end{bmatrix}$

$= \alpha^2+\alpha^2-\beta^2-\beta^2 = 2\alpha^2-2\beta^2 = \mathrm{cov}(x_{2i}, x_{1i})$

$$\therefore \mathrm{cov}(X) = \frac{1}{3}\begin{bmatrix} 2\alpha^2+2\beta^2 & 2\alpha^2-2\beta^2 \\ 2\alpha^2-2\beta^2 & 2\alpha^2+2\beta^2 \end{bmatrix} = \frac{2}{3}\begin{bmatrix} \alpha^2+\beta^2 & \alpha^2-\beta^2 \\ \alpha^2-\beta^2 & \alpha^2+\beta^2 \end{bmatrix}$$

(b) Compute the eigenvectors and eigenvalues of cov(***X***)



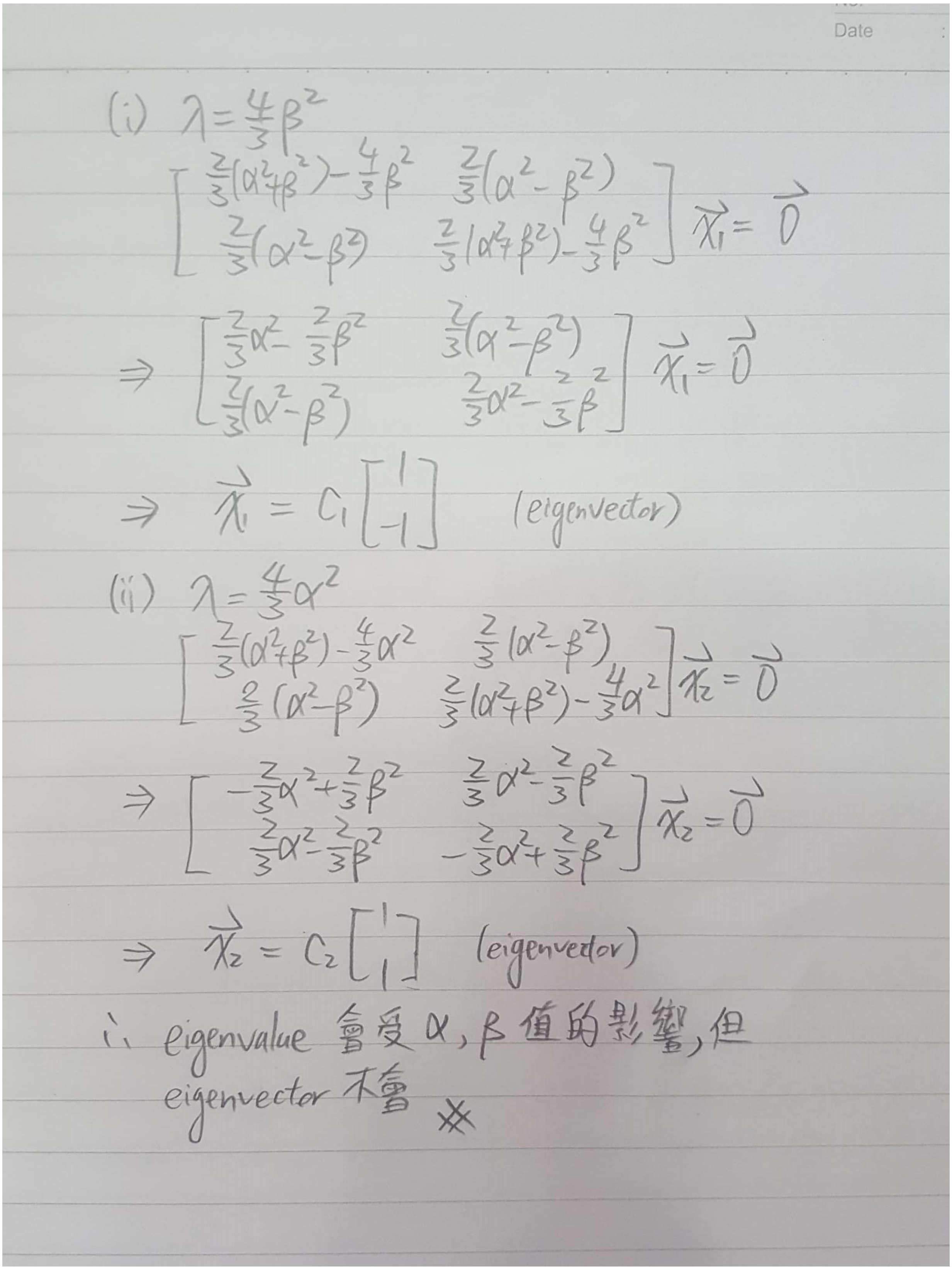
(b) $\det(\mathrm{cov}(x) - \lambda I_2) = 0$

$$\Rightarrow \left(\frac{2}{3}(\alpha^2+\beta^2)-\lambda\right)^2 - \left(\frac{2}{3}(\alpha^2-\beta^2)\right)^2 = 0$$

$$\Rightarrow \frac{2}{3}(\alpha^2+\beta^2)-\lambda = \pm\frac{2}{3}(\alpha^2-\beta^2)$$

$$\Rightarrow \lambda = \frac{2}{3}(\alpha^2+\beta^2) \mp \frac{2}{3}(\alpha^2-\beta^2)$$

$$= \frac{4}{3}\beta^2, \frac{4}{3}\alpha^2$$ (eigenvalue)



Date :

(i) $\lambda = \frac{4}{3}\beta^2$

$$\begin{bmatrix} \frac{2}{3}(\alpha^2+\beta^2)-\frac{4}{3}\beta^2 & \frac{2}{3}(\alpha^2-\beta^2) \\ \frac{2}{3}(\alpha^2-\beta^2) & \frac{2}{3}(\alpha^2+\beta^2)-\frac{4}{3}\beta^2 \end{bmatrix} \vec{x}_1 = \vec{0}$$

$$\Rightarrow \begin{bmatrix} \frac{2}{3}\alpha^2-\frac{2}{3}\beta^2 & \frac{2}{3}(\alpha^2-\beta^2) \\ \frac{2}{3}(\alpha^2-\beta^2) & \frac{2}{3}\alpha^2-\frac{2}{3}\beta^2 \end{bmatrix} \vec{x}_1 = \vec{0}$$

$$\Rightarrow \vec{x}_1 = c_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \text{(eigenvector)}$$

(ii) $\lambda = \frac{4}{3}\alpha^2$

$$\begin{bmatrix} \frac{2}{3}(\alpha^2+\beta^2)-\frac{4}{3}\alpha^2 & \frac{2}{3}(\alpha^2-\beta^2) \\ \frac{2}{3}(\alpha^2-\beta^2) & \frac{2}{3}(\alpha^2+\beta^2)-\frac{4}{3}\alpha^2 \end{bmatrix} \vec{x}_2 = \vec{0}$$

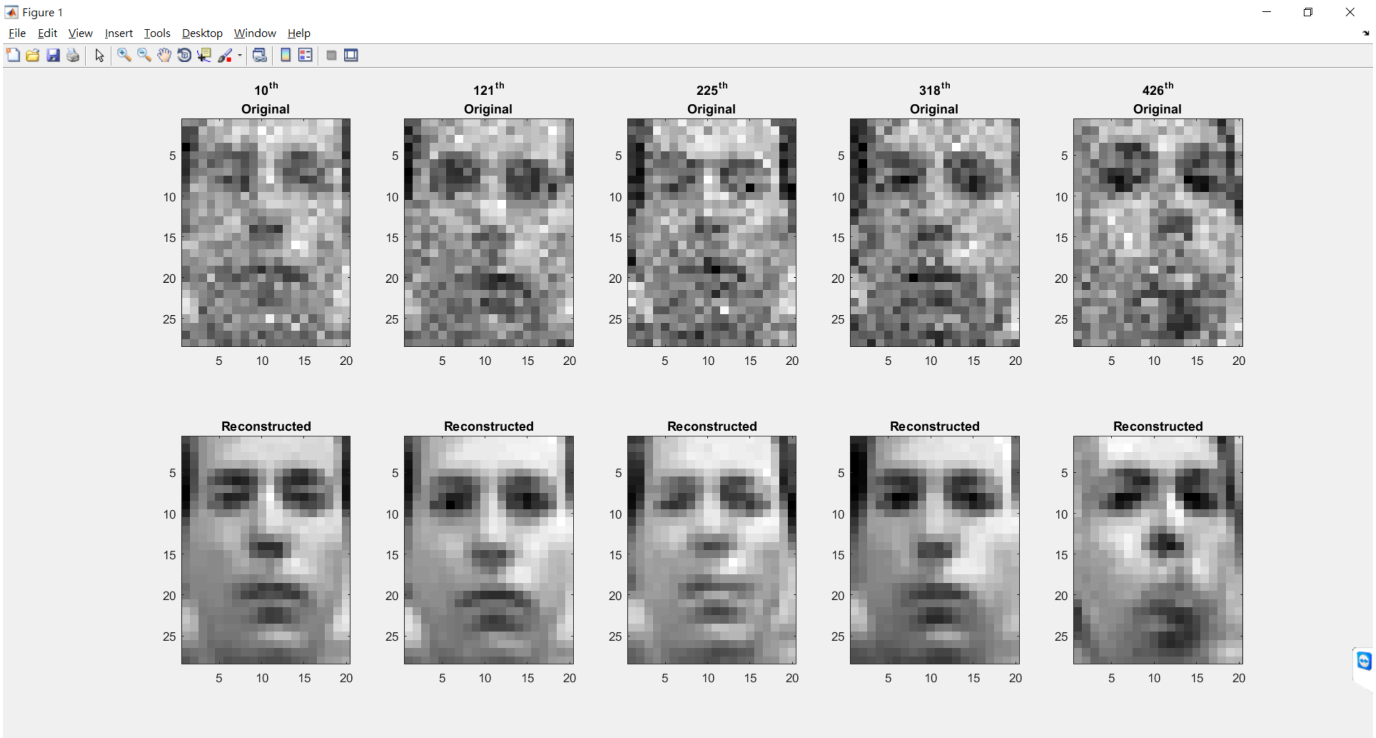
$$\Rightarrow \begin{bmatrix} -\frac{2}{3}\alpha^2+\frac{2}{3}\beta^2 & \frac{2}{3}\alpha^2-\frac{2}{3}\beta^2 \\ \frac{2}{3}\alpha^2-\frac{2}{3}\beta^2 & -\frac{2}{3}\alpha^2+\frac{2}{3}\beta^2 \end{bmatrix} \vec{x}_2 = \vec{0}$$

$$\Rightarrow \vec{x}_2 = c_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{(eigenvector)}$$

∴ eigenvalue 會受 α, β 值的影響, 但 eigenvector 不會 ※

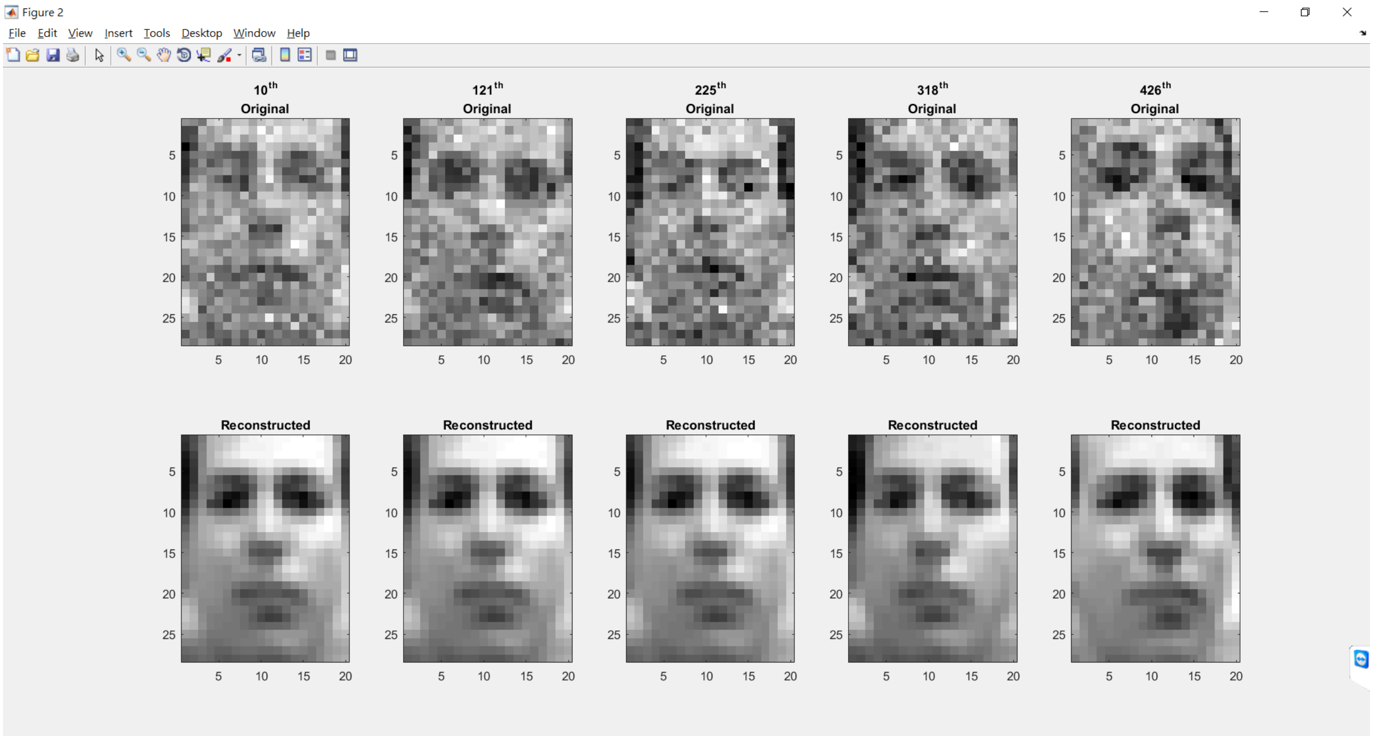
Problem 3 (Principal component analysis)

(a) Compute reconstructed images using the top 10 eigenvectors and plot the 10th, 121st, 225th, 318th, and 426th original and reconstructed images

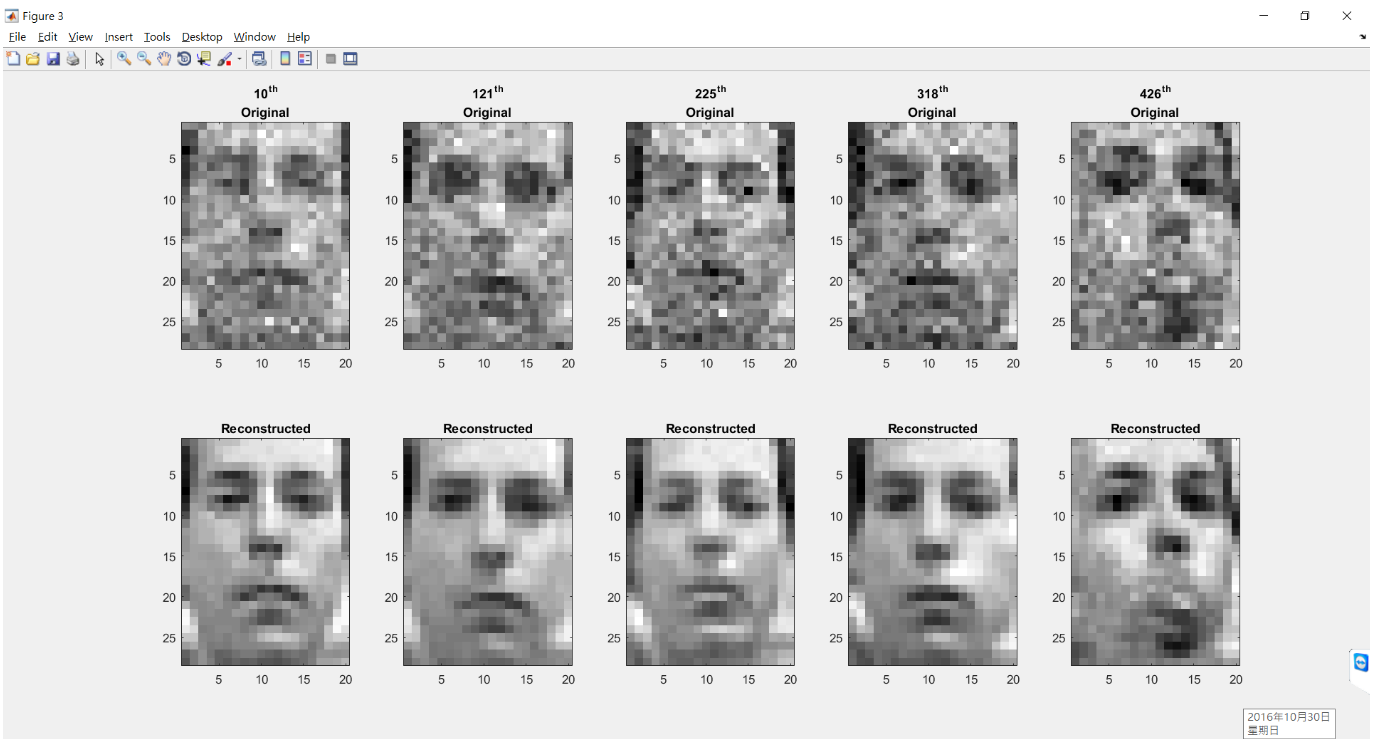


(b) Repeat part (a), assuming the correct intrinsic dimensionality of the data is 2 and 30

dimensionality of the data : 2



dimensionality of the data : 30

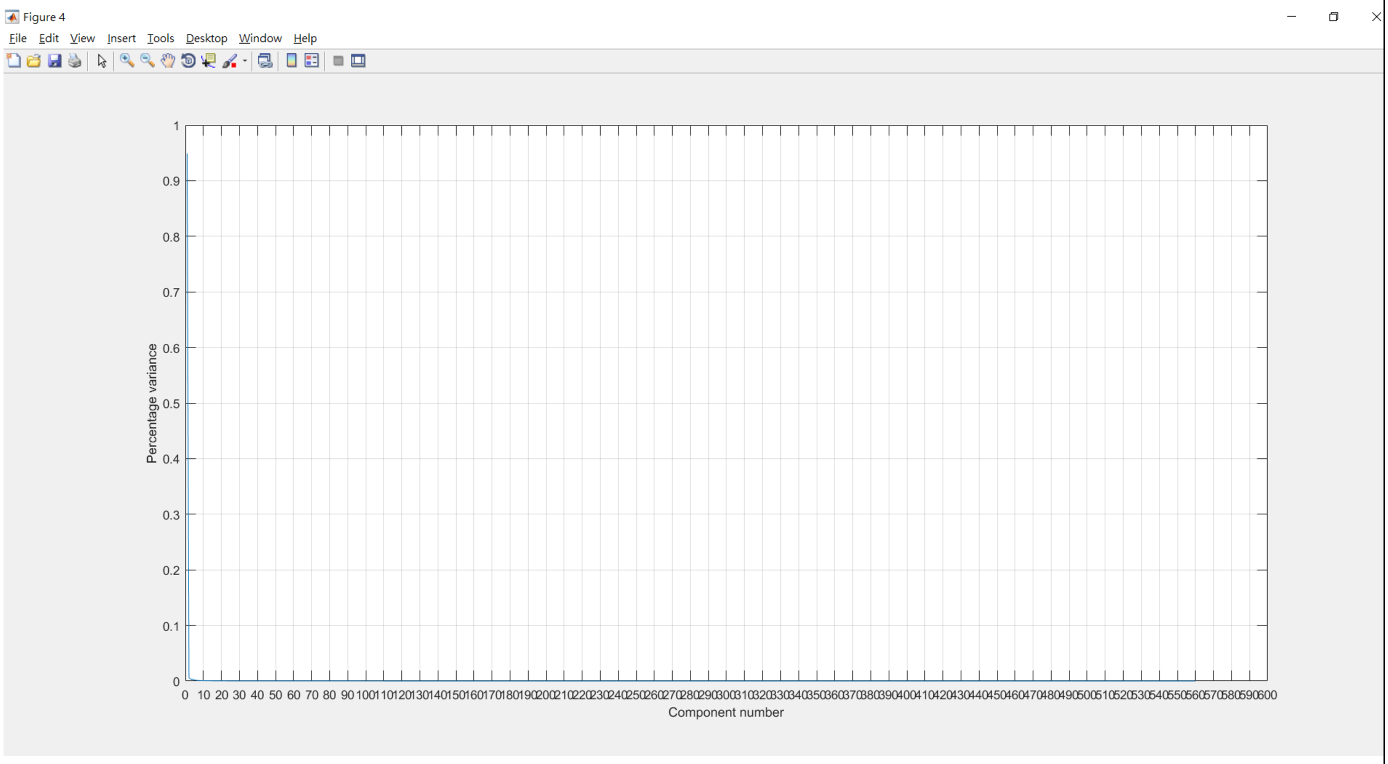


Discussion:

從取兩個eigenvector的圖可以看到，由於特徵太少，雖然還是能描述一個大概的人臉輪廓，但每個人臉都長得一樣!

而從取三十個eigenvector的圖我們發現，不同人臉的特徵有出來，但是取太多eigenvector的結果就是會將一些雜訊也取出，使得人臉較為模糊。

(c) Determine the best intrinsic dimensionality of the dataset



Discussion:

將每個PC的percentage of variance和component number的關係畫出，可以發現在component number = 之後的variance百分比都很小，可以視為添加的雜訊，因此我認為最好的intrinsic dimensionality of the dataset大約是4左右，如下圖所示。

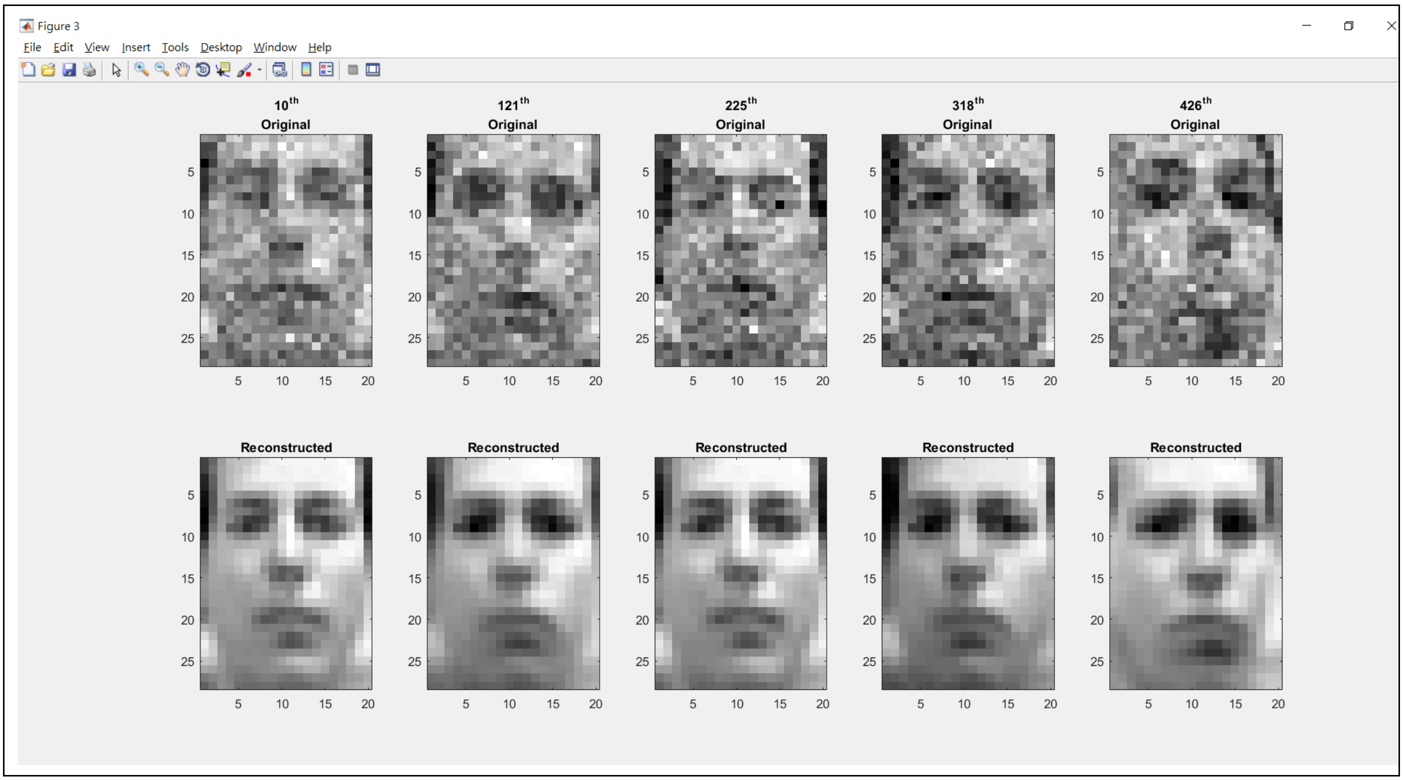


Figure 3
File Edit View Insert Tools Desktop Window Help
10th
Original
121th
Original
225th
Original
318th
Original
426th
Original
Reconstructed
Reconstructed
Reconstructed
Reconstructed
Reconstructed
5
10
15
20
25
5
10
15
20