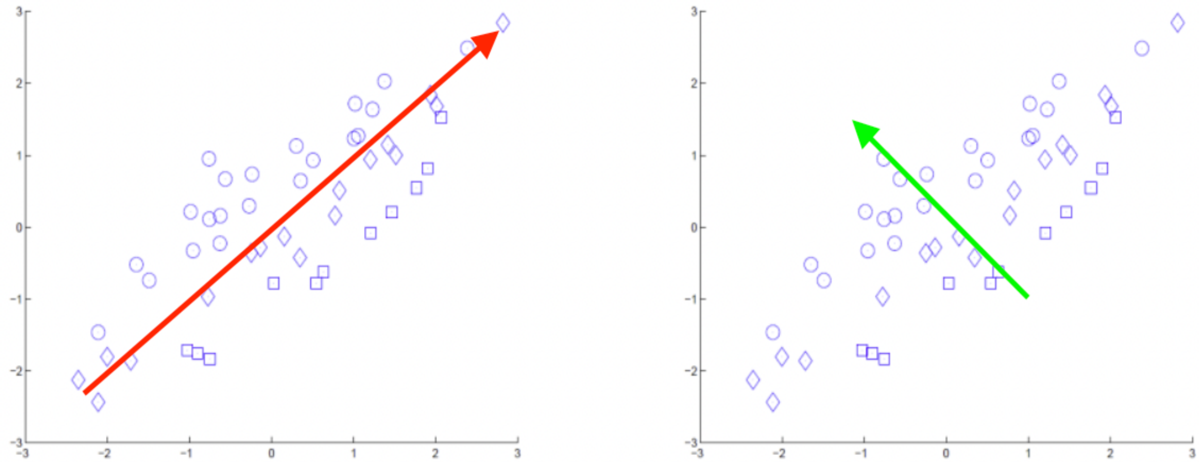
Problem 3: Principal Component Analysis (PCA)

(1)



classmate
Date
Page

(2) 3 data points in 2D space : (2,2), (0,0), (−2,2)

(a) First Principal component :

	x	y
Pt. 1	2	2
Pt. 2	0	0
Pt. 3	−2	−2
Mean	$\bar{x} = 0$	$\bar{y} = 0$

Using unbiased estimation of covariance:

$$var(X) = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})^2$$

$$Cov(X, Y) = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})$$

$\therefore cov(X, X) = var(X) = \frac{1}{3-1}\left((2-0)^2 + (0-0)^2 + (-2-0)^2\right)$

$= \frac{1}{2}(8) = 4.$

$\therefore cov(X, Y) = \frac{1}{3-1}\left[(2-0)(2-0) + (0)(0) + (-2-0)(-2-0)\right]$

$= \frac{1}{2}(8) = 4$

$\therefore cov(Y, X) = cov(X, Y) = 4$

$\therefore cov(Y, Y) = var(Y) = \frac{1}{3-1}\left[(2-0)^2 + (0-0)^2 + (-2-0)^2\right]$

$= \frac{1}{2}(8) = 4$

→ Covariance Matrix,

$$M = \begin{bmatrix} cov(X,X) & cov(X,Y) \\ cov(Y,X) & cov(Y,Y) \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$$

→ Eigen values of covariance matrix,

$$\det(M - \lambda I) = 0$$

$$\therefore \det\left(\begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right) = 0$$

$$\therefore \det \begin{bmatrix} 4-\lambda & 4 \\ 4 & 4-\lambda \end{bmatrix} = 0$$

$$\therefore (4-\lambda)^2 - (4)(4) = 0$$

$$\therefore (4-\lambda)^2 = 16$$

$$\therefore 4-\lambda = \pm 4$$

$$\therefore \lambda_1 = 0 \text{ and } \lambda_2 = 8$$

~~Eigen vector of covariance matrix~~

→ Since $\lambda_2 > \lambda_1$, $\lambda_2 = 8$ will give first principal component.

→ Eigen vector of covariance matrix,

→ for $\lambda_1 = 0$, $\begin{bmatrix} 4-\lambda & 4 \\ 4 & 4-\lambda \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$

$\therefore \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$

$\therefore \begin{bmatrix} 4x + 4y \\ 4x + 4y \end{bmatrix} = 0$

$\therefore x + y = 0 \Rightarrow x = 1, y = -1$ or $x = -1, y = 1$ and so on

→ for $\lambda_2 = 8$, $\begin{bmatrix} 4-\lambda & 4 \\ 4 & 4-\lambda \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$

$\therefore \begin{bmatrix} -4 & 4 \\ 4 & -4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$

$\therefore \begin{bmatrix} -4x + 4y \\ 4x - 4y \end{bmatrix} = 0$

$\therefore \left.\begin{matrix} -x + y = 0 \\ x - y = 0 \end{matrix}\right\} \Rightarrow x = y.$

$\Rightarrow \therefore x = y = 1$ or $x = y = 2$ and so on

$\therefore$ Eigen vector for λ_2, $U_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

→ Since λ_2 gives first principal component, we will normalize eigen vector v_2.

$$\therefore e_2 = \begin{bmatrix} \dfrac{1}{\sqrt{(1)^2+(1)^2}} \\ \dfrac{1}{\sqrt{(1)^2+(1)^2}} \end{bmatrix}$$

$$\therefore e_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}$$

⇒ First principal component, $v = \left[\dfrac{\sqrt{2}}{2}, \dfrac{\sqrt{2}}{2}\right]^T$

or $v = \left[-\dfrac{\sqrt{2}}{2}, -\dfrac{\sqrt{2}}{2}\right]^T$ is also right.

(b) Project the given data points into 1D subspace and find new coordinates.

→ Let the new coordinates in 1D subspace be P_1, P_2, P_3.

→ Now, $P_1 = e_2^T \begin{bmatrix} x_1 - \bar{x} \\ y_1 - \bar{y} \end{bmatrix}$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 2-0 \\ 2-0 \end{bmatrix}$$

$$= \frac{1}{\sqrt{2}}(2) + \frac{1}{\sqrt{2}}(2)$$

$$= \sqrt{2} + \sqrt{2} = 2\sqrt{2}$$

→ $P_2 = e_2^T \begin{bmatrix} x_2 - \bar{x} \\ y_2 - \bar{y} \end{bmatrix}$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$= 0$$

→ $P_3 = e_2^T \begin{bmatrix} x_3 - \bar{x} \\ y_3 - \bar{y} \end{bmatrix}$

$$= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} -2 \\ -2 \end{bmatrix}$$

$$= \frac{1}{\sqrt{2}}(-2) + \frac{1}{\sqrt{2}}(-2) = -\sqrt{2} - \sqrt{2} = -2\sqrt{2}$$

classmate

Date

Page

$\Rightarrow$ New coordinates,

$$P_1 = 2\sqrt{2}, \quad P_2 = 0, \quad P_3 = -2\sqrt{2}$$

$\rightarrow$ Variance of above data,

$$var(P) = \frac{1}{N-1} \sum_{n=1}^{N} \left(P_n - \bar{P}\right)^2$$

$\rightarrow$ mean is 0.

$$\therefore var(P) = \frac{1}{3-1} \left[(2\sqrt{2})^2 + (0)^2 + (-2\sqrt{2})^2 \right]$$

$$= \frac{1}{2} \left[8 + 8 \right]$$

$$\therefore Var(P) = 8$$

(c) Covariance Matrix of original data points,

$$M = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$$

→ variances are on the diagonal, so the overall variability is $4+4 = \boxed{8}$

→ we know the eigen values are $\lambda_1 = 0$ and $\lambda_2 = 8 \Rightarrow \lambda_2 > \lambda_1$.

$$\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{0}{0+8} = 0 \qquad \Bigg| \qquad \frac{\lambda_2}{\lambda_1+\lambda_2} = \frac{8}{0+8} = 1 = 100\%$$

→ So λ_2 will give the first principal component and will be responsible for explaining the total variance.

→ Cumulative explained variance of first principal component is total variance of the data set by combining variances of individual variances.

→ Thus as 100% of variance is captured by λ_2, no other variance is left to capture