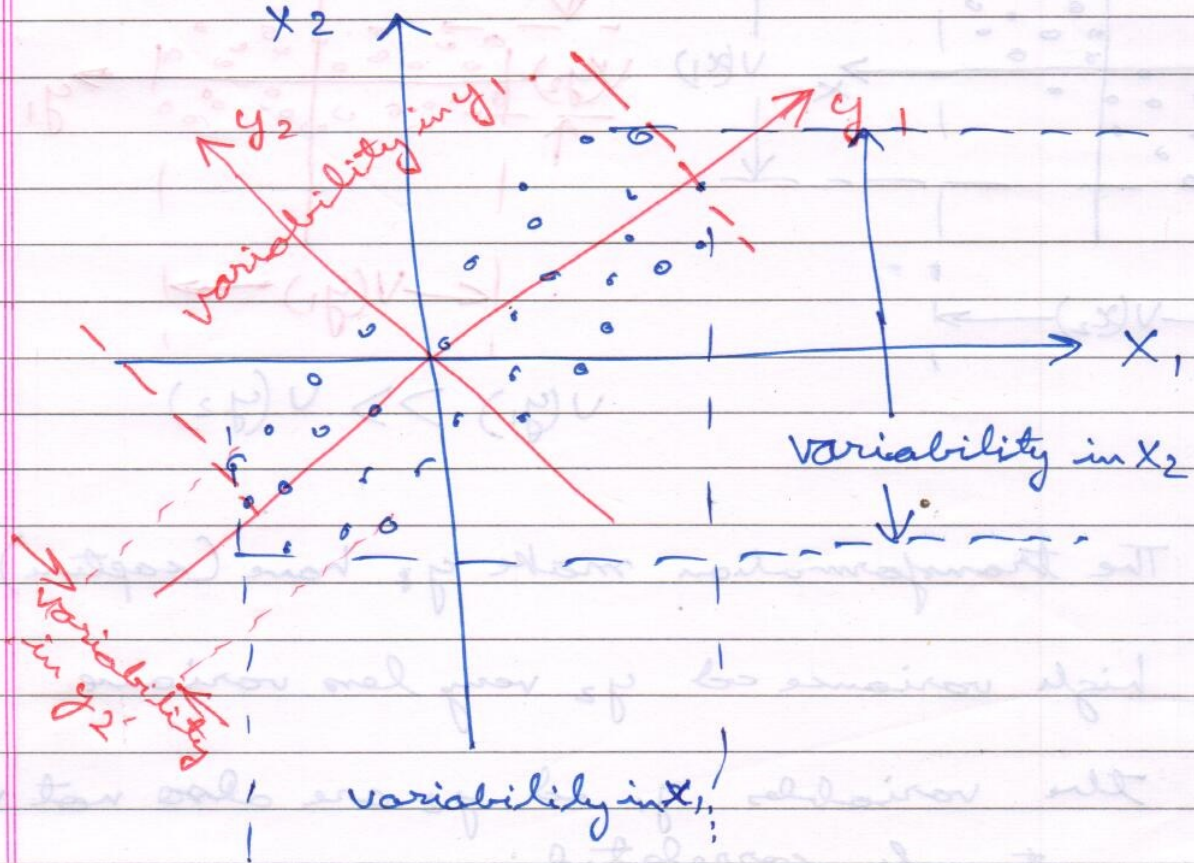
Principal Component Analysis (PCA)

- Linear projection of data so that maximum variance is captured in the lower dimensional representation of data.
- This can help in dimensionality reduction by only considering dimensions which have the highest variance.
- This is also called "Singular value decomposit

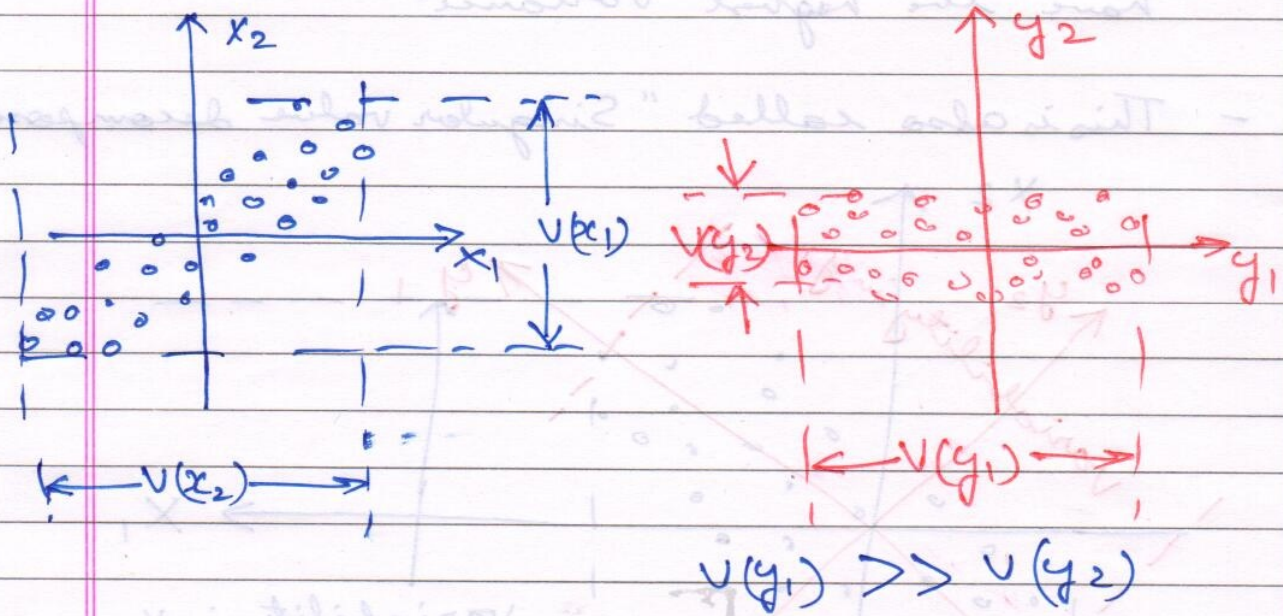


X_1 and X_2 are seen to be strongly correlated, thus

$$Corr(X) = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

where r_{12} is very close to 1.

X Space $\xrightarrow{\text{Transformation}}$ Y space



$V(y_1) >> V(y_2)$

- The transformation make y_1 have (capture) very high variance and y_2 very less variance
- the variables y_1 and y_2 are also not very strongly correlated.

In 2D space (n data points)

X Space $\xrightarrow{\text{Transformation}}$ Y Space

$$X_{n\times 2} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix} \Longrightarrow Y_{n\times 2} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \\ y_{n1} & y_{n2} \end{bmatrix}$$

variance : $V(y_1) > V(y_2)$

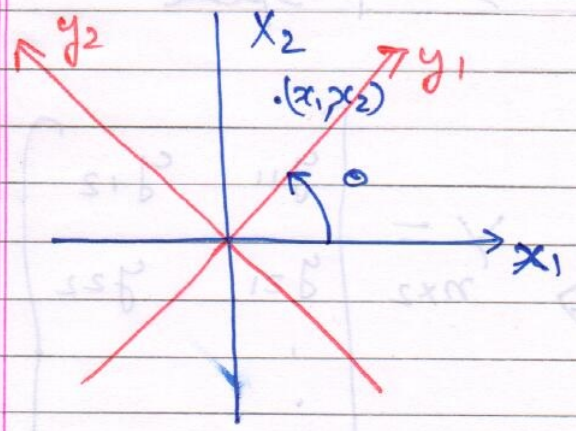
In P-dimensional space (n data points)

$$X_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \Longrightarrow Y_{n\times p} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

For PCA we want:

Variance in Y space : $V(y_1) \geq V(y_2) \geq \cdots \geq V(y_p)$

In 2D Space



The transformation of X space to Y space in 2D is just a rotation.

$$y_1 = x_1 \cos\theta + x_2 \sin\theta$$

and $$y_2 = -x_1 \sin\theta + x_2 \cos\theta$$

$$\Rightarrow \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$Y = A^T X$$

$$A^T = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

$$\Rightarrow A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = A^T \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

↑ Rotation in 3D

$$R_3(\theta) = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \Rightarrow \text{Rotation about } x_3$$

$$R_2(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \Rightarrow \text{Rotation about } x_2$$

$$R_1(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix} \Rightarrow \text{Rotation about } x_1$$

General Rotation in 3D

$A^T = R = R_1(\alpha)\, R_2(\beta)\, R_3(\gamma)$ is a rotation of γ about x_3 followed by a rotation of β about x_2 followed by a rotation of α about x_1

In general 3D rotation you can always find an axis (direction of unite length) that is not changed. If this invariant axis is $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ then $\underset{3\times1}{X} = \underset{3\times3}{R}\,\underset{3\times1}{X}$

$$\Rightarrow (R - I)X = 0$$

↑ Invariant vector

In a P-dimensional Space

$$\underset{P\times1}{Y} = \underset{P\times P}{A^T}\,\underset{P\times1}{X}$$

In 2D

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \end{bmatrix}$$

$$a_1 = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \qquad a_2 = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}$$

then $a_1^T a_1 = \begin{bmatrix} \cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$

$= \cos^2\theta + \sin^2\theta$

$= 1$

similarly $a_2^T a_2 = 1$

$$A^T A = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$\Rightarrow A^T A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I = A A^T = A A^{-1}$$

$$\Rightarrow A^T = A^{-1}$$

Thus A is an orthogonal transformation

$\therefore$ in p-dimensions

$$A = [a_1 \quad a_2 \quad \dots \quad a_p]$$

$$\therefore y_1 = a_1^T x = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$$

$$y_2 = a_2^T x = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p$$

$$\vdots$$

$$y_j = a_j^T x = a_{j1}x_1 + a_{j2}x_2 + \dots + a_{jp}x_p$$

$$\vdots$$

$$y_p = a_p^T x = a_{p1}x_1 + a_{p2}x_2 + \dots + a_{pp}x_p$$

$$a_i^T a_j = \delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

In PCA we want this transformation such that the variance $V(y_i)$ for y_i is as follows:

$$V(y_1) \geq V(y_2) \geq \dots \geq V(y_p)$$

Thus, in PCA we want a transformation matrix A^T such that we maximize the variance of the transformed space as given below:

$$V(y_1) \geq V(y_2) \geq \cdots \geq V(y_p)$$

---

p-dimensional data with n number of points

This can be represented in two ways

Matrix form

$$X_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n \times p}$$

Set form $X = \{ x_1, x_2 \cdots x_n \}$ where x_k is the element of the set and is a p-dimensional feature vector (ordered plot)

$$x_k = (x_{k1}, x_{k2}, x_{k3} \cdots x_{kp})$$

∴ For the "kth" set

$$\begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{bmatrix}^T = \underbrace{\begin{bmatrix} x_{k1} & x_{k2} & \cdots & x_{kp} \end{bmatrix}}_{1 \times p} \equiv x_k$$

x vector Row representation

k^{th} element of the set representation

Similarly

$$\begin{bmatrix} y_{k1} \\ y_{k2} \\ \vdots \\ y_{kp} \end{bmatrix}^T = \underbrace{\begin{bmatrix} y_{k1} & y_{k2} & \cdots & y_{kp} \end{bmatrix}}_{1 \times p} \equiv y_k$$

$$\therefore \begin{bmatrix} y_{k1} \\ y_{k2} \\ \vdots \\ y_{kp} \end{bmatrix} = A^T \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{bmatrix}$$

By taking Transpose on both the sides:

$$\Rightarrow \begin{bmatrix} y_{k1} \\ y_{k2} \\ \vdots \\ y_{kp} \end{bmatrix}^T = \left(A^T \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{bmatrix} \right)^T$$

$$\Rightarrow \begin{bmatrix} y_{k1} \\ y_{k2} \\ \vdots \\ y_{kp} \end{bmatrix}^T = \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{bmatrix}^T A$$

$$\Rightarrow y_k = x_k A$$

k^{th} element of set representation or row vector representation

$$\therefore \quad Y = A^T X$$

as column vector representation

is same as $y_k = x_k A$ as row vector representation

For PCA the X data set is not only rotated but also translated w.r.t. mean. Thus,

$$\underset{1\times p}{y_k} = \underset{1\times p}{(x_k - \bar{x})} \cdot \underset{p\times p}{A}$$

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{n}\left[\Sigma x_{k1} \;\; \Sigma x_{k2} \cdots \Sigma x_{kp}\right]$$

$$\Rightarrow \quad x_k = y_k A^T + \bar{x}$$

Now variance vector for Y and combination of variance and covariance of Y is

$$\text{variance vector of } Y = \left[\text{var}(Y_1) \;\; \text{var}(Y_2) \cdots \text{var}(Y_p)\right]$$

$$\text{covariance of } Y \equiv \underset{\uparrow}{\underset{p\times p}{\Sigma_y}} = \frac{1}{n-1}\sum_{k=1}^{n} \underset{p\times 1}{y_k^T} \underset{1\times p}{y_k}$$

$$\Rightarrow v_y = \frac{1}{n-1} \sum_{k=1}^{n} y_k^T y_k$$

$$= \frac{1}{n-1} \sum_{k=1}^{n} \left[(x_k - \bar{x}) \cdot A\right]^T \left[(x_k - \bar{x}) \cdot A\right]$$

$$= \frac{1}{n-1} \sum_{k=1}^{n} A^T (x_k - \bar{x})^T \cdot (x_k - \bar{x}) \cdot A$$

$$= A^T \left[\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^T \cdot (x_k - \bar{x}) \right] \cdot A$$

$$\Rightarrow v_y = A^T C A$$

↑ Covariance matrix of X

$$c_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} \left[x_{ki} - \frac{1}{n} \sum_{l=1}^{n} x_{li}\right] \cdot \left[x_{kj} - \frac{1}{n} \sum_{l=1}^{n} x\right.$$

$i, j = 1, 2 \ldots p$

Now we want to maximize v_y by choosing the optimal A.

However, we cannot choose any A, since if it has to have the following constrain:

$$A^T A = I$$

To solve this optimization problem with a constrain, we construct a Lagrange function:

$$L = A^T C A - \lambda (A^T A - I)$$

and maximize the value of L.

$$\Rightarrow \quad \frac{\partial L}{\partial A} = 0$$

$$\Rightarrow \quad CA - \lambda A = 0$$

$$\Rightarrow \quad (C - \lambda I) \cdot A = 0$$

This is an Eigenvalue Eigenvector problem

The Eigenvalue problem can be traced back to solving the following DE:

$$\frac{dy}{dt} = \underbrace{A}_{n\times n \text{ matrix}} y \Big\} \; n \text{ linear DEs}$$

This will have a solution of type

$$y(t) = e^{\lambda t} X$$

$$\Rightarrow \lambda e^{\lambda t} X = A e^{\lambda t} X$$

$$\Rightarrow \lambda X = A X$$

$$\Rightarrow \underbrace{(A - \lambda I)}_{n\times n} \underbrace{X}_{n\times 1} = 0$$

This gives n eigenvalues of λ and n eigenvectors

Eg. $y_1' = 5y_1 + y_2$
$y_2' = 3y_1 + 3y_2$ $\Big\}$ $A = \begin{bmatrix} 5 & 1 \\ 3 & 3 \end{bmatrix}$

this gives two eigenvalues and two eigenvectors

$\lambda_1 = 6 \quad x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$; $\lambda_2 = 2 \quad x_2 = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$

Thus general solution of the DE is

$$y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = c_1 e^{\lambda_1 t} x_1 + c_2 e^{\lambda_2 t} x_2$$

$$\Rightarrow y(t) = c_1 e^{6t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 e^{2t} \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

If we put the B.Cs we get

at $t = 0$; $y(0) = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ -3 \end{bmatrix}$

Now note if we know eigenvector for $AX = \lambda_1 X$ then we also know the eigenvector for $A^2 X = \lambda_2 X$

$\Rightarrow$ This is because $A^2 X = \lambda_2 X$

$$\Rightarrow A(\lambda_1 X) = \lambda_2 X$$

$$\Rightarrow \lambda_1^2 X = \lambda_2 X$$

$\Rightarrow$ they have the same eigenvector by $\lambda_2 = (\lambda_1)^2$

In the same manner for the equation

$$(A + cI)X = \lambda_3 X$$

$$\lambda_3 = (\lambda_1 + c)$$

and similarly $A^n X = \lambda_n X$ has same eigenvector but $\lambda_n = (\lambda_1)^n$.

Note that in eigenvalue problem X is a $n \times 1$ matrix (or a vector)

i.e. $(A - \lambda I) X = 0$

($n \times n$, $n \times n$, $n \times 1$)

whereas in PCA it has a square matrix form

$$(C - \lambda I) A = 0$$

($p \times p$, $p \times p$, $p \times p$)

Thus, A is a concatenation of the eigenvectors of C $\Rightarrow A = (a_1, a_2 \ldots a_p)$

The variance in Y corresponds to the eigenvalues of $\lambda_1, \lambda_2 \dots \lambda_p$.

This is because

$$CA = \lambda A$$

$$\Rightarrow \quad \lambda = A^T C A = v_y$$

Thus, if we want to capture 95% variance, then we can reduce the number of dimensions to q such that

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq \frac{95}{100}$$

For example take $p = 2$ and $n = 4$ such that $X = \{(1,1), (2,1), (2,2), (3,2)\}$

then $\bar{X} = \left(2, \frac{3}{2}\right)$ is the mean matrix

and the covariance matrix is

$$C = \frac{1}{3}\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

Thus, the eigenvalues are

$$\lambda_1 = 0.8727$$

and $\lambda_2 = 0.1273$

This corresponds to

$$a_1 = \begin{bmatrix} -0.85065 \\ -0.52571 \end{bmatrix}$$

$$a_2 = \begin{bmatrix} 0.52571 \\ -0.85065 \end{bmatrix}$$

and $A = [a_1 \quad a_2]$

$$\Rightarrow A = \begin{bmatrix} -0.85065 & 0.52571 \\ -0.52571 & -0.85065 \end{bmatrix}$$

∴ Transformation of X by A^T rotation matrix would lead to Y as given by

$$Y = A^T(X - \bar{X})$$ In column vector representation

$$y_k = (x_k - \bar{x}) \cdot A$$ In row vector representation

$$\Rightarrow X = \{(1,1), (2,1), (2,2), (3,2)\}$$

will transform to

$$y_{n=1} = \begin{bmatrix} -0.85065 & -0.52571 \\ 0.52571 & -0.85065 \end{bmatrix} \begin{bmatrix} 1-2 \\ 1-1.5 \end{bmatrix}$$

$\uparrow A^T$ $\quad$ $[X - \bar{X}]$

$$= \begin{bmatrix} 1.1135 \\ -0.10039 \end{bmatrix}$$

$$y_{n=2} = \begin{bmatrix} A^T \end{bmatrix} \begin{bmatrix} 2-2 \\ 1-1.5 \end{bmatrix} = \begin{bmatrix} 0.2628 \\ 0.42533 \end{bmatrix}$$

similarly $y_{n=3} = \begin{bmatrix} -0.26286 \\ -0.42533 \end{bmatrix}$ and $y_{n=4} = \begin{bmatrix} -1.1135 \\ 0.10039 \end{bmatrix}$

∴ $X = \{(1,1), (2,1), (2,2), (3,2)\}$ is transformed to

$Y = \{(1.1135, -0.10039), (0.26286, 0.42533), (-0.26286, -0.425$
$/ (-1.1135, 0.10039)\}$

Without dimensionality Reduction

$X \xrightarrow{A^T} Y$

$A^T = \begin{bmatrix} -0.85065 & -0.525 \\ -0.52571 & -0.8506 \end{bmatrix}$

$\{(1,1),(2,1),(2,2),(3,2)\}$

$\{(1.1135, -0.10039),$
$(0.26286, 0.425533),$
$(-0.26286, -0.425533),$
$(-1.1135, -0.10039)\}$

Mean X

$= (2, 1.5)$

Mean Y

$= (0, 0)$

Covariance of X

$= \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{bmatrix}$

Covariance of Y

$= \begin{bmatrix} 0.87266 & 0 \\ 0 & 0.12732 \end{bmatrix}$

$= \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$

Covariance is a diagonal vecto

$X \xleftarrow{A} Y$

$\{(1,1),(2,1),(2,2),(3,2)\}$

The inverse transform results in the same X data.

With dimensionality reduction

Since the variance of Y_1 is 0.87266 and Y_2 is 0.12732, thus if we take only Y_1 dimension then we get 0.87266 i.e. 87% of variance.

Thus we do not take $A = [a_1 \quad a_2]$

but instead $A' = [a_1 \quad 0] = \begin{bmatrix} -0.85065 \\ -0.52571 \end{bmatrix}$

$$\Rightarrow X \xrightarrow{(A')^T} Y'$$

$$(A')^T = \begin{bmatrix} -0.85065 & -0.52571 \\ 0 & 0 \end{bmatrix}$$

$$X \longrightarrow Y'$$

$$X = \begin{Bmatrix} (1,1) \\ (2,1) \\ (2,2) \\ (3,2) \end{Bmatrix} \qquad Y' = \begin{Bmatrix} (1.11351, 0) \\ (0.26286, 0) \\ (-0.26286, 0) \\ (-1.11351, 0) \end{Bmatrix}$$

The inverse transform of Y' with A' gives

$$X' = \begin{Bmatrix} (1.0528, \ 0.91462) \\ (1.7764, \ 1.36181) \\ (2.2236, \ 1.63819) \\ (2.9472, \ 2.08538) \end{Bmatrix}$$