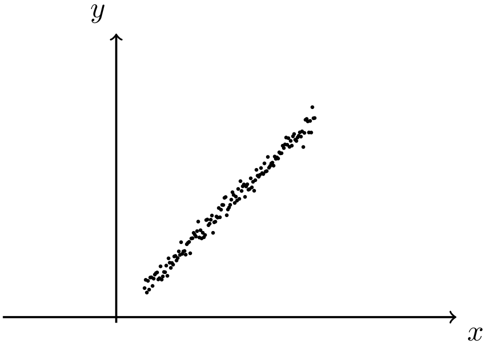
Module 6.4 : Principal Component Analysis and its Interpretations

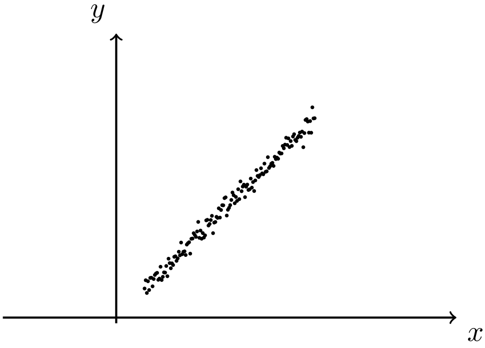
The story ahead...

The story ahead...

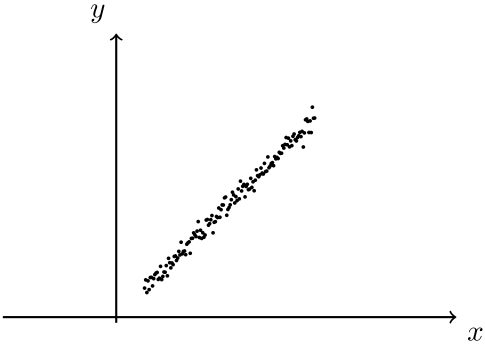
- Over the next few slides we will introduce Principal Component Analysis and see three different interpretations of it



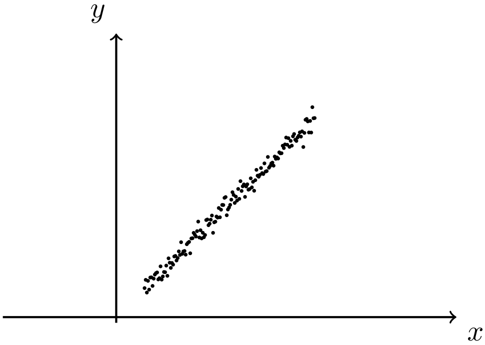
- Consider the following data



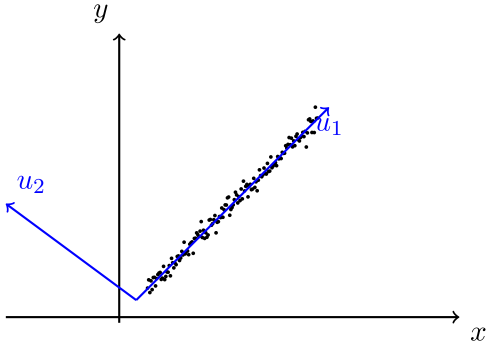
- Consider the following data
- Each point (vector) here is represented using a linear combination of the x and y axes (i.e. using the point's x and y co-ordinates)



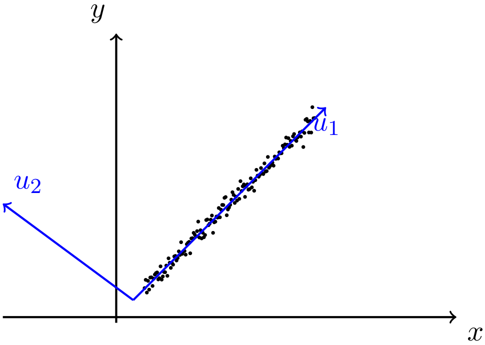
- Consider the following data
- Each point (vector) here is represented using a linear combination of the x and y axes (i.e. using the point's x and y co-ordinates)
- In other words we are using x and y as the basis



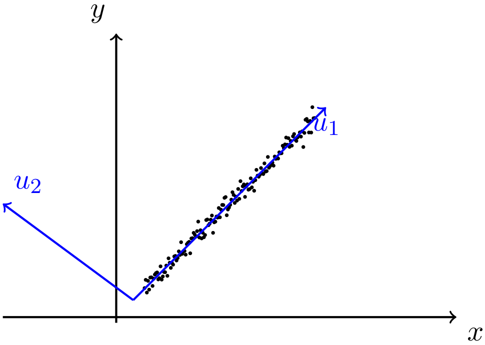
- Consider the following data
- Each point (vector) here is represented using a linear combination of the x and y axes (i.e. using the point's x and y co-ordinates)
- In other words we are using x and y as the basis
- What if we choose a different basis?



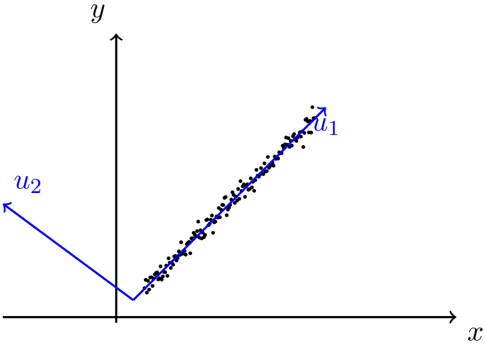
- For example, what if we use u_1 and u_2 as a basis instead of x and y.



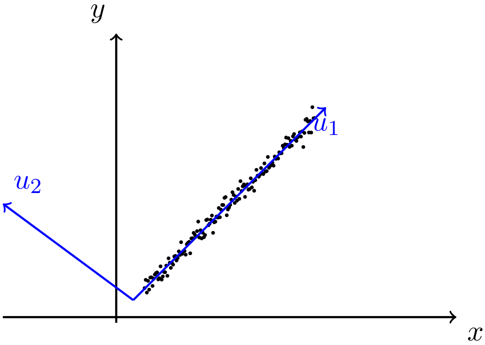
- For example, what if we use u_1 and u_2 as a basis instead of x and y.
- We observe that all the points have a very small component in the direction of u_2 (almost noise)



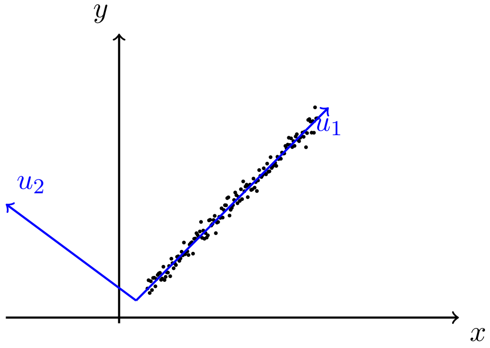
- For example, what if we use u_1 and u_2 as a basis instead of x and y.
- We observe that all the points have a very small component in the direction of u_2 (almost noise)
- It seems that the same data which was originally in $\mathbb{R}^2(x, y)$ can now be represented in $\mathbb{R}^1(u_1)$ by making a smarter choice for the basis



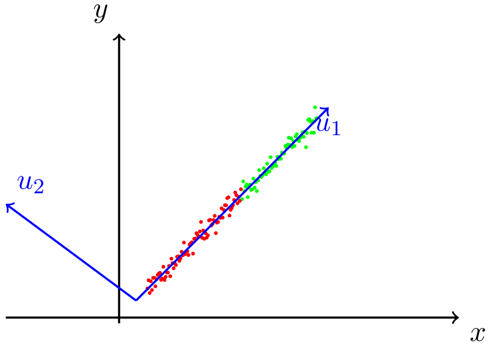
- Let's try stating this more formally



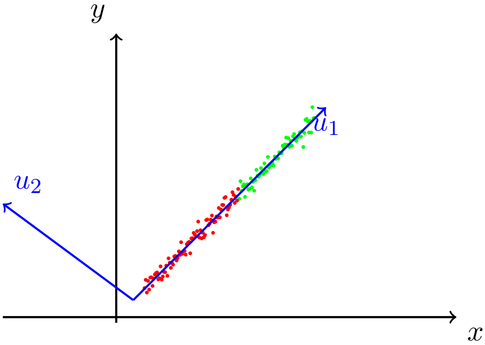
- Let's try stating this more formally
- Why do we not care about u_2?



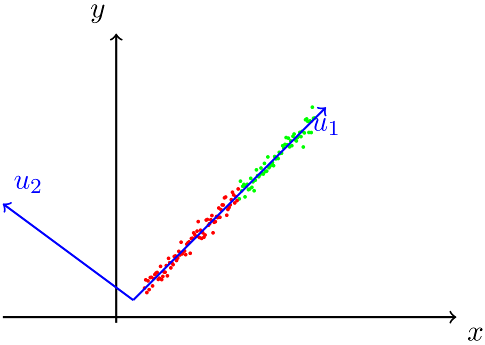
- Let's try stating this more formally
- Why do we not care about u_2?
- Because the variance in the data in this direction is very small (all data points have almost the same value in the u_2 direction)



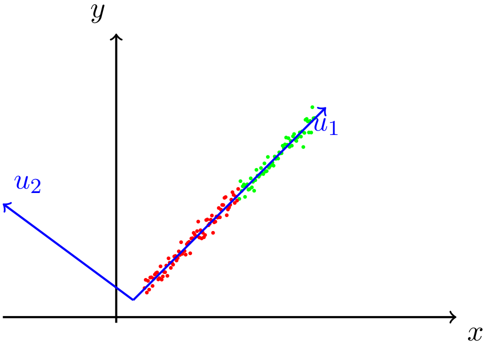
- Let's try stating this more formally
- Why do we not care about u_2?
- Because the variance in the data in this direction is very small (all data points have almost the same value in the u_2 direction)
- If we were to build a classifier on top of this data then u_2 would not contribute to the classifier as the points are not distinguishable along this direction



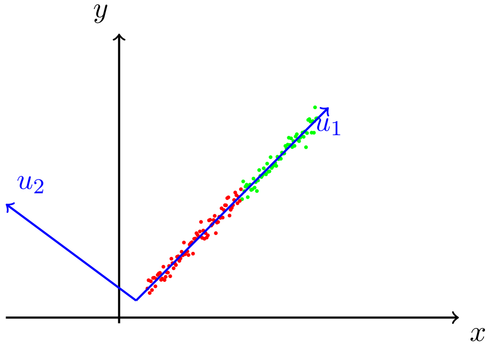
- In general, we are interested in representing the data using fewer dimensions such that



- In general, we are interested in representing the data using fewer dimensions such that the data has high variance along these dimensions



- In general, we are interested in representing the data using fewer dimensions such that the data has high variance along these dimensions
- Is that all?



- In general, we are interested in representing the data using fewer dimensions such that the data has high variance along these dimensions
- Is that all?
- No, there is something else that we desire. Let's see what.

- Consider the following data

x	**y**	**z**
1	1	1
0.5	0	0
0.25	1	1
0.35	1.5	1.5
0.45	1	1
0.57	2	2.1
0.62	1.1	1
0.73	0.75	0.76
0.72	0.86	0.87

x	y	z
1	1	1
0.5	0	0
0.25	1	1
0.35	1.5	1.5
0.45	1	1
0.57	2	2.1
0.62	1.1	1
0.73	0.75	0.76
0.72	0.86	0.87

- Consider the following data
- Is z adding any new information beyond what is already contained in y?

x	**y**	**z**
1	1	1
0.5	0	0
0.25	1	1
0.35	1.5	1.5
0.45	1	1
0.57	2	2.1
0.62	1.1	1
0.73	0.75	0.76
0.72	0.86	0.87

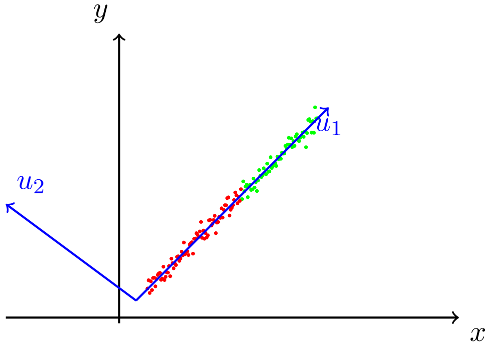
$$\rho_{yz} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(z_i - \overline{z})^2}}$$

- Consider the following data
- Is z adding any new information beyond what is already contained in y?
- The two columns are highly correlated (or they have a high covariance)

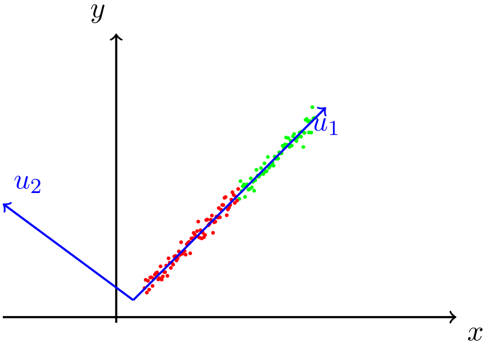
x	y	z
1	1	1
0.5	0	0
0.25	1	1
0.35	1.5	1.5
0.45	1	1
0.57	2	2.1
0.62	1.1	1
0.73	0.75	0.76
0.72	0.86	0.87

$$\rho_{yz} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(z_i - \overline{z})^2}}$$

- Consider the following data
- Is z adding any new information beyond what is already contained in y?
- The two columns are highly correlated (or they have a high covariance)
- In other words the column z is redundant since it is linearly dependent on y.

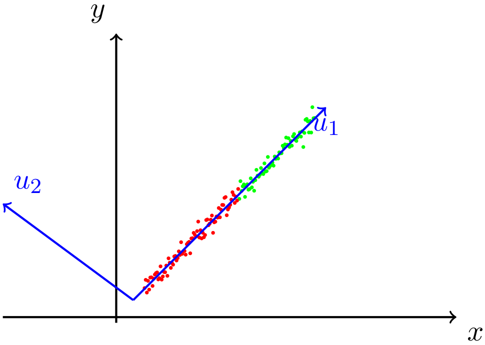


In general, we are interested in representing the data using fewer dimensions such that



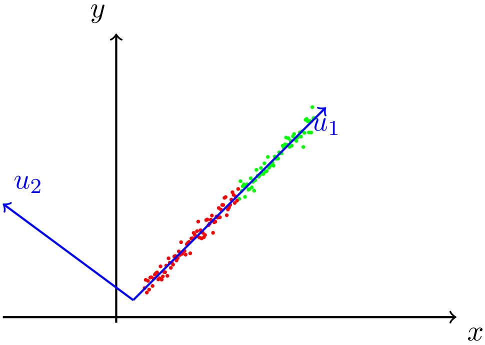
In general, we are interested in representing the data using fewer dimensions such that

- the data has high variance along these dimensions



In general, we are interested in representing the data using fewer dimensions such that

- the data has high variance along these dimensions
- the dimensions are linearly independent (uncorrelated)



In general, we are interested in representing the data using fewer dimensions such that

- the data has high variance along these dimensions
- the dimensions are linearly independent (uncorrelated)
- (even better if they are orthogonal because that is a very convenient basis)

Let $p_1, p_2, \cdots, p_n$ be a set of such n linearly independent orthonormal vectors. Let P be a $n \times n$ matrix such that $p_1, p_2, \cdots, p_n$ are the columns of P.

Let $p_1, p_2, \cdots, p_n$ be a set of such n linearly independent orthonormal vectors. Let P be a $n \times n$ matrix such that $p_1, p_2, \cdots, p_n$ are the columns of P.

Let $x_1, x_2, \cdots, x_m \in \mathbb{R}^n$ be m data points and let X be a matrix such that $x_1, x_2, \cdots, x_m$ are the rows of this matrix. Further let us assume that the data is 0-mean and unit variance.

Let $p_1, p_2, \cdots, p_n$ be a set of such n linearly independent orthonormal vectors. Let P be a $n \times n$ matrix such that $p_1, p_2, \cdots, p_n$ are the columns of P.

Let $x_1, x_2, \cdots, x_m \in \mathbb{R}^n$ be m data points and let X be a matrix such that $x_1, x_2, \cdots, x_m$ are the rows of this matrix. Further let us assume that the data is 0-mean and unit variance.

We want to represent each x_i using this new basis P.

$$x_i = \alpha_{i1}p_1 + \alpha_{i2}p_2 + \alpha_{i3}p_3 + \cdots + \alpha_{in}p_n$$

Let $p_1, p_2, \cdots, p_n$ be a set of such n linearly independent orthonormal vectors. Let P be a $n \times n$ matrix such that $p_1, p_2, \cdots, p_n$ are the columns of P.

Let $x_1, x_2, \cdots, x_m \in \mathbb{R}^n$ be m data points and let X be a matrix such that $x_1, x_2, \cdots, x_m$ are the rows of this matrix. Further let us assume that the data is 0-mean and unit variance.

We want to represent each x_i using this new basis P.

$$x_i = \alpha_{i1}p_1 + \alpha_{i2}p_2 + \alpha_{i3}p_3 + \cdots + \alpha_{in}p_n$$

For an orthonormal basis we know that we can find these $\alpha_i's$ using

$$\alpha_{ij} = x_i^T p_j = \begin{bmatrix} \leftarrow & x_i & \rightarrow \end{bmatrix}^T \begin{bmatrix} \uparrow \\ p_j \\ \downarrow \end{bmatrix}$$

In general, the transformed data $\hat{x}_i$ is given by

$$\hat{x}_i = \begin{bmatrix} \leftarrow & x_i^T & \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & & \uparrow \\ p_1 & \cdots & p_n \\ \downarrow & & \downarrow \end{bmatrix} = x_i^T P$$

In general, the transformed data $\hat{x}_i$ is given by

$$\hat{x}_i = \begin{bmatrix} \leftarrow & x_i^T & \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & & \uparrow \\ p_1 & \cdots & p_n \\ \downarrow & & \downarrow \end{bmatrix} = x_i^T P$$

and

$$\hat{X} = XP \qquad (\hat{X} \text{ is the matrix of transformed points})$$

Theorem:

If X is a matrix such that its columns have zero mean and if $\hat{X} = XP$ then the columns of $\hat{X}$ will also have zero mean.

Theorem:
If X is a matrix such that its columns have zero mean and if $\hat{X} = XP$ then the columns of $\hat{X}$ will also have zero mean.
Proof: For any matrix A, $\mathbf{1}^T A$ gives us a row vector with the i^{th} element containing the sum of the i^{th} column of A. (this is easy to see using the row-column picture of matrix multiplication).

Theorem:

If X is a matrix such that its columns have zero mean and if $\hat{X} = XP$ then the columns of $\hat{X}$ will also have zero mean.

Proof: For any matrix A, $\mathbf{1}^T A$ gives us a row vector with the i^{th} element containing the sum of the i^{th} column of A. (this is easy to see using the row-column picture of matrix multiplication).

Consider

$$\mathbf{1}^T \hat{X} = \mathbf{1}^T XP = (\mathbf{1}^T X)P$$

But $\mathbf{1}^T X$ is the row vector containing the sums of the columns of X. Thus $\mathbf{1}^T X = 0$. Therefore, $\mathbf{1}^T \hat{X} = 0$.

Hence the transformed matrix also has columns with sum $= 0$.

Theorem:

If X is a matrix such that its columns have zero mean and if $\hat{X} = XP$ then the columns of $\hat{X}$ will also have zero mean.

Proof: For any matrix A, $\mathbf{1}^T A$ gives us a row vector with the i^{th} element containing the sum of the i^{th} column of A. (this is easy to see using the row-column picture of matrix multiplication).

Consider

$$\mathbf{1}^T \hat{X} = \mathbf{1}^T XP = (\mathbf{1}^T X)P$$

But $\mathbf{1}^T X$ is the row vector containing the sums of the columns of X. Thus $\mathbf{1}^T X = 0$. Therefore, $\mathbf{1}^T \hat{X} = 0$.

Hence the transformed matrix also has columns with sum $= 0$.

Theorem:

$X^T X$ is a symmetric matrix.

Theorem:
If X is a matrix such that its columns have zero mean and if $\hat{X} = XP$ then the columns of $\hat{X}$ will also have zero mean.
Proof: For any matrix A, $\mathbf{1}^T A$ gives us a row vector with the i^{th} element containing the sum of the i^{th} column of A. (this is easy to see using the row-column picture of matrix multiplication).
Consider

$$\mathbf{1}^T \hat{X} = \mathbf{1}^T XP = (\mathbf{1}^T X)P$$

But $\mathbf{1}^T X$ is the row vector containing the sums of the columns of X. Thus $\mathbf{1}^T X = 0$. Therefore, $\mathbf{1}^T \hat{X} = 0$.
Hence the transformed matrix also has columns with sum $= 0$.

Theorem:
$X^T X$ is a symmetric matrix.
Proof: We can write $(X^T X)^T = X^T (X^T)^T = X^T X$

Definition:

If X is a matrix whose columns are zero mean then $\Sigma = \frac{1}{m}X^T X$ is the covariance matrix. In other words each entry Σ_{ij} stores the covariance between columns i and j of X.

Definition:

If X is a matrix whose columns are zero mean then $\Sigma = \frac{1}{m}X^TX$ is the covariance matrix. In other words each entry Σ_{ij} stores the covariance between columns i and j of X.

Explanation: Let C be the covariance matrix of X. Let μ_i, μ_j denote the means of the i^{th} and j^{th} column of X respectively. Then by definition of covariance, we can write :

$$
\begin{aligned}
C_{ij} &= \frac{1}{m}\sum_{k=1}^{m}(X_{ki} - \mu_i)(X_{kj} - \mu_j) \\
&= \frac{1}{m}\sum_{k=1}^{m} X_{ki}X_{kj} \qquad (\because \mu_i = \mu_j = 0) \\
&= \frac{1}{m}X_i^T X_j = \frac{1}{m}(X^TX)_{ij}
\end{aligned}
$$

$$\hat{X} = XP$$

$$\hat{X} = XP$$

- Using the previous theorem & definition, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write :

$$\hat{X} = XP$$

- Using the previous theorem & definition, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write :

$$\frac{1}{m}\hat{X}^T\hat{X} = \frac{1}{m}(XP)^T XP$$

$$\hat{X} = XP$$

- Using the previous theorem & definition, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write :

$$\frac{1}{m}\hat{X}^T\hat{X} = \frac{1}{m}(XP)^T XP = \frac{1}{m}P^T X^T XP = P^T\left(\frac{1}{m}X^T X\right)P$$

$$\hat{X} = XP$$

- Using the previous theorem & definition, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write :

$$\frac{1}{m}\hat{X}^T\hat{X} = \frac{1}{m}(XP)^T XP = \frac{1}{m}P^T X^T XP = P^T\left(\frac{1}{m}X^T X\right)P = P^T\Sigma P$$

$$\hat{X} = XP$$

- Using the previous theorem & definition, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write :

$$\frac{1}{m}\hat{X}^T\hat{X} = \frac{1}{m}(XP)^T XP = \frac{1}{m}P^T X^T XP = P^T\left(\frac{1}{m}X^T X\right)P = P^T\Sigma P$$

- Each cell i, j of the covariance matrix $\frac{1}{m}\hat{X}^T\hat{X}$ stores the covariance between columns i and j of $\hat{X}$.

$$\hat{X} = XP$$

- Using the previous theorem & definition, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write :

$$\frac{1}{m}\hat{X}^T\hat{X} = \frac{1}{m}(XP)^T XP = \frac{1}{m}P^TX^TXP = P^T\left(\frac{1}{m}X^TX\right)P = P^T\Sigma P$$

- Each cell i, j of the covariance matrix $\frac{1}{m}\hat{X}^T\hat{X}$ stores the covariance between columns i and j of $\hat{X}$.
- Ideally we want,

$$\left(\frac{1}{m}\hat{X}^T\hat{X}\right)_{ij} = 0 \qquad i \neq j \text{ (covariance} = 0)$$

$$\left(\frac{1}{m}\hat{X}^T\hat{X}\right)_{ij} \neq 0 \qquad i = j \text{ (variance} \neq 0)$$

$$\hat{X} = XP$$

- Using the previous theorem & definition, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write :

$$\frac{1}{m}\hat{X}^T\hat{X} = \frac{1}{m}(XP)^T XP = \frac{1}{m}P^T X^T XP = P^T\left(\frac{1}{m}X^T X\right)P = P^T\Sigma P$$

- Each cell i, j of the covariance matrix $\frac{1}{m}\hat{X}^T\hat{X}$ stores the covariance between columns i and j of $\hat{X}$.
- Ideally we want,

$$\left(\frac{1}{m}\hat{X}^T\hat{X}\right)_{ij} = 0 \qquad i \neq j \text{ (covariance} = 0)$$

$$\left(\frac{1}{m}\hat{X}^T\hat{X}\right)_{ij} \neq 0 \qquad i = j \text{ (variance} \neq 0)$$

In other words, we want

$$\frac{1}{m}\hat{X}^T\hat{X} = P^T\Sigma P = D \qquad \text{[where D is a diagonal matrix]}$$

- We want,

$$P^T \Sigma P = D$$

- We want,

$$P^T \Sigma P = D$$

- But Σ is a square matrix and P is an orthogonal matrix

- We want,

$$P^T \Sigma P = D$$

- But Σ is a square matrix and P is an orthogonal matrix
- Which orthogonal matrix satisfies the following condition?

- We want,

$$P^T \Sigma P = D$$

- But Σ is a square matrix and P is an orthogonal matrix
- Which orthogonal matrix satisfies the following condition?

$$P^T \Sigma P = D$$

- We want,

$$P^T \Sigma P = D$$

- But Σ is a square matrix and P is an orthogonal matrix
- Which orthogonal matrix satisfies the following condition?

$$P^T \Sigma P = D$$

- In other words, which orthogonal matrix P diagonalizes Σ?

- We want,

$$P^T \Sigma P = D$$

- But Σ is a square matrix and P is an orthogonal matrix
- Which orthogonal matrix satisfies the following condition?

$$P^T \Sigma P = D$$

- In other words, which orthogonal matrix P diagonalizes Σ?
- **Answer:** A matrix P whose columns are the eigen vectors of $\Sigma = X^T X$ [By Eigen Value Decomposition]

- We want,

$$P^T \Sigma P = D$$

- But Σ is a square matrix and P is an orthogonal matrix
- Which orthogonal matrix satisfies the following condition?

$$P^T \Sigma P = D$$

- In other words, which orthogonal matrix P diagonalizes Σ?
- **Answer:** A matrix P whose columns are the eigen vectors of $\Sigma = X^T X$ [By Eigen Value Decomposition]
- Thus, the new basis P used to transform X is the basis consisting of the eigen vectors of $X^T X$

- Why is this a good basis?

- Why is this a good basis?
- Because the eigen vectors of $X^T X$ are linearly independent (**proof : Slide 19 Theorem 1**)

- Why is this a good basis?
- Because the eigen vectors of X^TX are linearly independent (**proof : Slide 19 Theorem 1**)
- And because the eigen vectors of X^TX are orthogonal ($\because$ X^TX is symmetric - **saw proof earlier**)

- Why is this a good basis?
- Because the eigen vectors of $X^T X$ are linearly independent (**proof : Slide 19 Theorem 1**)
- And because the eigen vectors of $X^T X$ are orthogonal ($\because$ $X^T X$ is symmetric - **saw proof earlier**)
- This method is called Principal Component Analysis for transforming the data to a new basis where the dimensions are non-redundant (low covariance) & not noisy (high variance)

- Why is this a good basis?
- Because the eigen vectors of $X^T X$ are linearly independent (**proof : Slide 19 Theorem 1**)
- And because the eigen vectors of $X^T X$ are orthogonal ($\because$ $X^T X$ is symmetric - **saw proof earlier**)
- This method is called Principal Component Analysis for transforming the data to a new basis where the dimensions are non-redundant (low covariance) & not noisy (high variance)
- In practice, we select only the top-k dimensions along which the variance is high (this will become more clear when we look at an alternalte interpretation of PCA)