Principal Component Analysis

Ananda Biswas

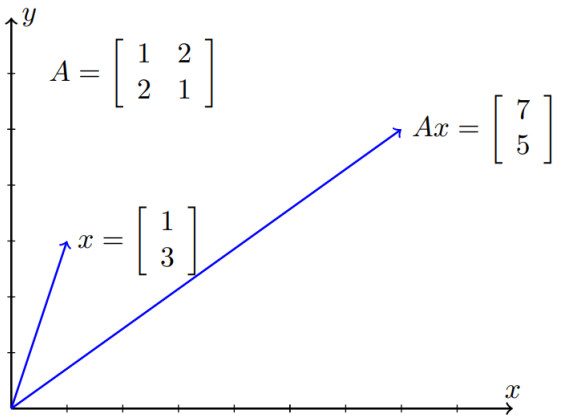
Contents

1 Warming up with Linear Algebra

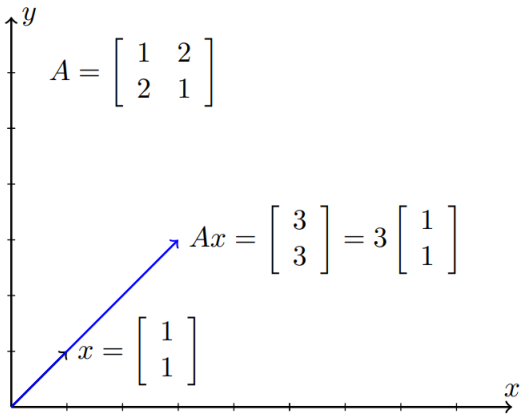
1.1 Eigenvalues and Eigenvectors

☞ What happens when a matrix hits a vector?

✎ The vector gets strayed from its path and becomes a new vector. The vector may also get scaled (elongated or shortened) in the process.



For a given square matrix A, there exists special vectors which refuse to stray from their path. These vectors are called eigenvectors.



In brief, matrix multiplications are linear transformations. Specially for square matrices, after multiplication, there stands few vectors whose respective spans remain unchanged. These vectors themselves are the eigenvectors of the matrix and the relative change in magnitude are the corresponding eigenvalues. It is also justified by the equation $A\underset{\sim}{x} = \lambda\underset{\sim}{x}$ *i.e.* multiplying $\underset{\sim}{x}$ by A is same as multiplying it by a constant λ.

☞ **Definition :** Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be n eigenvalues of an $n \times n$ matrix A. λ_1 is called the **dominant eigenvalue** of A if

$$|\lambda_1| \geq |\lambda_i| \ \ \forall i = 2, 3, \ldots, n.$$

☞ **Definition :** A matrix M is called a **column stochastic matrix** if all the entries are positive and the sum of the elements in each column is equal to 1.

There are also row stochastic matrices and doubly stochastic matrices.

☞ **Theorem :** The largest(dominant) eigenvalue of a stochastic matrix is 1.

☞ **Theorem :** If A is a $n \times n$ square matrix with a dominant eigenvalue, then the sequence of vectors given by $Av_0, A^2v_0, \ldots, A^nv_0, \ldots$ approaches a multiple of the dominant eigenvector of A.

☞ **Theorem :** If A is a square symmetric $n \times n$ matrix, then the solution to the following optimization problem is given by the eigenvector corresponding to the largest eigenvalue of A.

$$\max_{\underset{\sim}{x}} \ \underset{\sim}{x}^T A \underset{\sim}{x}$$

$$\text{s.t. } \|\underset{\sim}{x}\| = 1, \ \underset{\sim}{x} \in \mathbb{R}^n$$

and the solution to

$$\min_{\underset{\sim}{x}} \ \underset{\sim}{x}^T A \underset{\sim}{x}$$

$$\text{s.t. } \|\underset{\sim}{x}\| = 1, \ \underset{\sim}{x} \in \mathbb{R}^n$$

is given by the eigenvector corresponding to the smallest eigenvalue of A.

proof : This is a constrained optimization problem that can be solved using Lagrange Multipliers:

$$L = \underset{\sim}{x}^T A \underset{\sim}{x} - \lambda(\underset{\sim}{x}^T \underset{\sim}{x} - 1)$$

$$\frac{\partial L}{\partial \underset{\sim}{x}} = 2A\underset{\sim}{x} - \lambda(2\underset{\sim}{x}) = 0 \Rightarrow A\underset{\sim}{x} = \lambda \underset{\sim}{x}$$

Hence $\underset{\sim}{x}$ must be an eigenvector of A with eigenvalue λ.

Multiplying by $\underset{\sim}{x}^T$:

$$\underset{\sim}{x}^T A \underset{\sim}{x} = \lambda \underset{\sim}{x}^T \underset{\sim}{x} = \lambda \text{ (since } \underset{\sim}{x}^T \underset{\sim}{x} = 1)$$

Therefore, the critical points of this constrained problem are the eigenvalues of A.

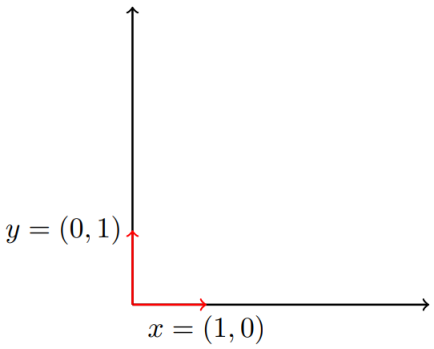
The maximum value is the largest eigenvalue, while the minimum value is the smallest eigenvalue.

1.2 More Linear Algebra

☞ **Definition :** A set of n vectors $v_1, v_2, \ldots, v_n$ is called **linearly independent** if and only if no vector in the set can be expressed as a linear combination of the remaining $n-1$ vectors.

☞ **Definition :** A set of vectors $\in \mathbb{R}^n$ is called a **basis**, if they are linearly independent and every vector $\in \mathbb{R}^n$ can be expressed as a linear combination of these vectors.

☞ Consider the space $\mathbb{R}^2$ and two vectors $\underset{\sim}{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\underset{\sim}{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.



Any vector $\begin{bmatrix} a \\ b \end{bmatrix}$ can be expressed as a linear combination of these two vectors *i.e.*

$$\begin{bmatrix} a \\ b \end{bmatrix} = a \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

And indeed we are used to representing all vectors in $\mathbb{R}^2$ as a linear combination of these two vectors. But there is nothing sacrosanct about this particular choice of $\underset{\sim}{x}$ and $\underset{\sim}{y}$. We could have chosen any 2 linearly independent vectors in $\mathbb{R}^2$ as the basis vectors.

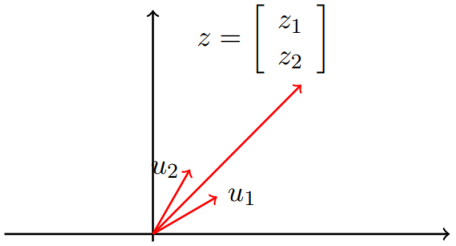
For example, consider the linearly independent vectors, $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\begin{bmatrix} 5 \\ 7 \end{bmatrix}$. See how any vector $\begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^2$ can be expressed as a linear combination of these two vectors.

$$\begin{bmatrix} a \\ b \end{bmatrix} = x_1 \begin{bmatrix} 2 \\ 3 \end{bmatrix} + x_2 \begin{bmatrix} 5 \\ 7 \end{bmatrix}.$$

We can find x_1 and x_2 by solving a system of equations

$$\begin{aligned} a &= 2x_1 + 5x_2 \\ b &= 3x_1 + 7x_2 \end{aligned}$$

In general, given a set of linearly independent vectors $\underset{\sim}{u_1}, \underset{\sim}{u_2}, \ldots, \underset{\sim}{u_n} \in \mathbb{R}^n$, we can express any vector $\underset{\sim}{z} \in \mathbb{R}^n$ as a linear combination of these vectors.



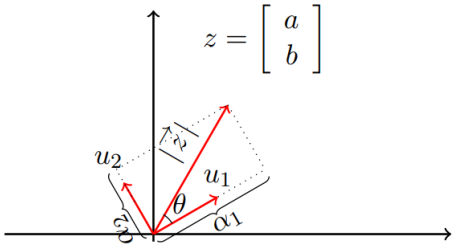
$$\underset{\sim}{z} = \alpha_1 \underset{\sim}{u_1} + \alpha_2 \underset{\sim}{u_2} + \cdots + \alpha_n \underset{\sim}{u_n}$$

$$\Rightarrow \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \alpha_1 \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1n} \end{bmatrix} + \alpha_2 \begin{bmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2n} \end{bmatrix} + \cdots + \alpha_n \begin{bmatrix} u_{n1} \\ u_{n2} \\ \vdots \\ u_{nn} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} u_{11} & u_{21} & \cdots & u_{n1} \\ u_{12} & u_{22} & \cdots & u_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1n} & u_{2n} & \cdots & u_{nn} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}$$

We can now find the α_i's using Gaussian Elimination (Time Complexity: $O(n^3)$).

☞ Now let us see if we have orthonormal basis.



Then $\underset{\sim}{u_i}^T \underset{\sim}{u_j} = 0 \; \forall i \neq j$ and $\underset{\sim}{u_i}^T \underset{\sim}{u_i} = \|\underset{\sim}{u_i}\|^2 = 1$.

Again we have:

$$\begin{aligned} \underset{\sim}{z} &= \alpha_1 \underset{\sim}{u_1} + \alpha_2 \underset{\sim}{u_2} + \cdots + \alpha_n \underset{\sim}{u_n} \\ \Rightarrow \underset{\sim}{u_1}^T \underset{\sim}{z} &= \alpha_1 \underset{\sim}{u_1}^T \underset{\sim}{u_1} + \cdots + \alpha_n \underset{\sim}{u_1}^T \underset{\sim}{u_n} \\ &= \alpha_1 \end{aligned}$$

We can directly find each α_i using a dot product between z and $\underset{\sim}{u_i}$ (time complexity $O(N)$).

The total complexity will be $O(N^2)$.

Also from figure, $\alpha_1 = |\underset{\sim}{z}| \cos\theta = |\underset{\sim}{z}| \cdot \dfrac{\underset{\sim}{z}^T \underset{\sim}{u_1}}{|\underset{\sim}{z}||\underset{\sim}{u_1}|} = \underset{\sim}{z}^T \underset{\sim}{u_1}$.

Similarly, $\alpha_2 = \underset{\sim}{z}^T \underset{\sim}{u_2}$.

So, **an orthonormal basis is the most convenient basis that one can hope for**.

☞ Turns out that the eigenvectors can form a basis. In fact, the eigenvectors of a square symmetric matrix are orthogonal and thus they form a very convenient basis. But why would we want to use the eigenvectors as a basis instead of the more natural co-ordinate axes? We will find the answer in PCA.

1.3 Eigenvalue Decomposition

Let $\underset{\sim}{u_1}, \underset{\sim}{u_2}, \ldots, \underset{\sim}{u_n}$ be the eigenvectors of a square matrix A and let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the corresponding eigenvalues.

Consider a matrix U whose columns are $\underset{\sim}{u_1}, \underset{\sim}{u_2}, \ldots, \underset{\sim}{u_n}$.

Now

$$AU = A \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \underset{\sim}{u_1} & \underset{\sim}{u_2} & \cdots & \underset{\sim}{u_n} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

$$= \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ A\underset{\sim}{u_1} & A\underset{\sim}{u_2} & \cdots & A\underset{\sim}{u_n} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

$$= \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \lambda_1\underset{\sim}{u_1} & \lambda_2\underset{\sim}{u_2} & \cdots & \lambda_n\underset{\sim}{u_n} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

$$= \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \underset{\sim}{u_1} & \underset{\sim}{u_2} & \cdots & \underset{\sim}{u_n} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

$$= U\Lambda$$

where Λ is a diagonal matrix whose diagonal elements are the eigenvalues of A.

$\therefore AU = U\Lambda$.

Now, if the columns of U are linearly independent *i.e.* if A has n linearly independent eigenvectors *i.e.* if A has n distinct eigenvalues, then U^{-1} exists and we can write

$$A = U\Lambda U^{-1} \quad \text{[eigenvalue decomposition]}$$
$$U^{-1}AU = \Lambda \quad \text{[diagonalization of } A\text{]}$$

If A is symmetric then the situation is even more convenient. The eigenvectors are orthogonal. Further let's assume, that the eigenvectors have been normalized. $\left[\underset{\sim}{u_i}^T \underset{\sim}{u_i} = 1\right]$
Then,

$$Q = U^T U = \begin{bmatrix} \leftarrow u_1 \rightarrow \\ \leftarrow u_2 \rightarrow \\ \vdots \\ \leftarrow u_n \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \cdots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

Each entry of the matrix, Q_{ij} is given by $\underset{\sim}{u_i}^T \underset{\sim}{u_i}$

$$Q_{ij} = \underset{\sim}{u_i}^T \underset{\sim}{u_i} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

$$\therefore U^T U = \mathbb{I} \text{ (the identity matrix)}$$

So U^T is the inverse of U (very convenient to calculate).

1.4 The story so far ...

- The eigenvectors corresponding to different eigenvalues are linearly independent.
- The eigenvectors of a square symmetric matrix are orthogonal.
- The eigenvectors of a square symmetric matrix can thus form a convenient basis.

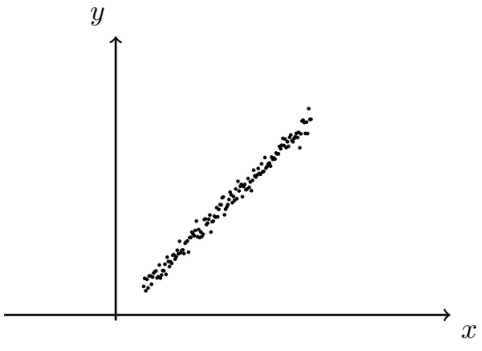
We will put all these to use.

2 Principal Component Analysis

The objective of Principal Component Analysis is to make a new representation of the data. We are also compressing the data and we want the compression to be as loss-less as possible. It is a feature extraction technique.

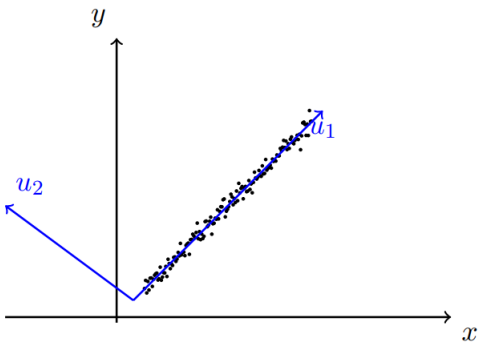
2.1 Interpretation 1

- Consider the following data.



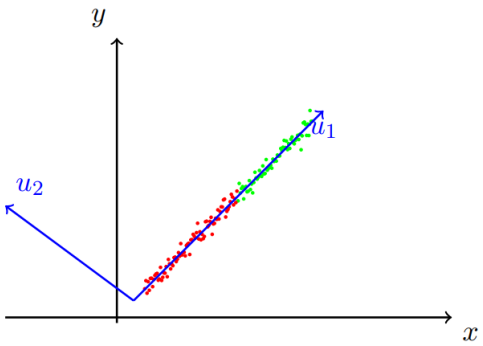
Each point (vector) here is represented using a linear combination of the x and y axes (*i.e.* using the point's x and y co-ordinates). In other words we are using $x-$axis and $y-$axis as the basis. What if we choose a different basis?

For example, what if we use $\underset{\sim}{u_1}$ and $\underset{\sim}{u_2}$ as a basis instead of $x-$axis and $y-$axis. We observe that all the points have a very small component in the direction of $\underset{\sim}{u_2}$ (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.

But why not care about $\underset{\sim}{u_2}$? Because the variance in the data in this direction is very small (all data points have almost the same value in the $\underset{\sim}{u_2}$ direction).



If we were to build a classifier on top of this data then $\underset{\sim}{u_2}$ 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 the data has high variance along these dimensions. But this is not all what we desire.

• 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

Notice that y and z are highly correlated. So z adds no new information beyond what is already contained in y. In other words, z is redundant as it is largely linearly dependent on y.

🎯 So in general, in PCA, we are interested in representing the data using fewer dimensions

(not in the sense of throwing away some dimensions, it is going to be a new set of dimensions; chopping off dimensions ❌, transforming the data ✅) such that

(i) The data has high variance along these dimensions;

(ii) The dimensions are linearly independent (uncorrelated); even better if they are orthogonal because that will be a very convenient basis.

• Now let $\underset{\sim}{p_1}, \underset{\sim}{p_2}, \ldots, \underset{\sim}{p_n}$ be a set of such n linearly independent orthonormal vectors. Let P be a $n \times n$ matrix such that $\underset{\sim}{p_1}, \underset{\sim}{p_2}, \ldots, \underset{\sim}{p_n}$ are the columns of P.

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

We want to represent each $\underset{\sim}{x_i}$ using this new basis P as follows.

$$\underset{\sim}{x_i} = \alpha_{i1}\underset{\sim}{p_1} + \alpha_{i2}\underset{\sim}{p_2} + \alpha_{i3}\underset{\sim}{p_3} + \ldots + \alpha_{in}\underset{\sim}{p_n}.$$

For an orthonormal basis we know that we can find these α'_{ij}s using

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

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

$$\underset{\sim}{\hat{x}_i} = \begin{bmatrix} \leftarrow & \underset{\sim}{x_i} & \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & & \uparrow \\ \underset{\sim}{p_1} & \cdots & \underset{\sim}{p_n} \\ \downarrow & & \downarrow \end{bmatrix} = \underset{\sim}{x_i}^T P = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in})_{1\times n}$$

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.
► **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

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

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

■ **Theorem** : $X^T X$ is a symmetric matrix.
► **Proof :** We can write

$$\left(X^T X\right)^T = X^T \left(X^T\right)^T = X^T X$$

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

► **Explanation** : Let Σ 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:

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

• We have $\hat{X} = XP$. Using the previous theorems & result, we get $\frac{1}{m}\hat{X}^T\hat{X}$ is the covariance matrix of the transformed data. We can write:

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

We know 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

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

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 $P^T\Sigma P = D$ where Σ is a square matrix and P is an orthogonal matrix.

Now the question is which orthogonal matrix satisfies the following condition $P^T\Sigma P = D$.

In other words, which orthogonal matrix P diagonalizes Σ? Answer is a matrix P whose columns are the eigenvectors of $\Sigma = \frac{1}{m}X^T X$ [by Eigenvalue Decomposition].

Thus, the new basis P used to transform X is the basis consisting of the eigenvectors of $\frac{1}{m}X^T X$.

• Why is this a good basis? Because the eigenvectors of $\frac{1}{m}X^T X$ are linearly independent and because the eigenvectors of $\frac{1}{m}X^T X$ are orthogonal.

• 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 alternate interpretation of PCA).

2.2 Interpretation 2

Given n orthogonal linearly independent vectors $\underset{\sim}{p_1}, \underset{\sim}{p_2}, \ldots, \underset{\sim}{p_n}$, we can represent $\underset{\sim}{x_i}$ exactly (exactly only when we use all $\underset{\sim}{p_i}$'s, it will be approximate when we discard some of $\underset{\sim}{p_i}$'s) as a linear combination of these vectors as follows.

$$\underset{\sim}{x_i} = \sum_{j=1}^{n} \alpha_{ij} \underset{\sim}{p_j} \qquad \text{[we know how to estimate } \alpha_{ij}\text{'s but we will come back to that later]}$$

But we are interested only in the top-k dimensions (we want to get rid of noisy & redundant dimensions)

$$\underset{\sim}{\hat{x}_i} = \sum_{j=1}^{k} \alpha_{ij} \underset{\sim}{p_j}$$

So we want to select p_i's such that we minimise the reconstructed error :

$$e = \sum_{i=1}^{m} (\underset{\sim}{x_i} - \underset{\sim}{\hat{x}_i})^T (\underset{\sim}{x_i} - \underset{\sim}{\hat{x}_i})$$

Now,

$$e = \sum_{i=1}^{m} (\underset{\sim}{x_i} - \underset{\sim}{\hat{x}_i})^T (\underset{\sim}{x_i} - \underset{\sim}{\hat{x}_i})$$

$$= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \alpha_{ij} \underset{\sim}{p_j} - \sum_{j=1}^{k} \alpha_{ij} \underset{\sim}{p_j} \right)^T \left(\sum_{j=1}^{n} \alpha_{ij} \underset{\sim}{p_j} - \sum_{j=1}^{k} \alpha_{ij} \underset{\sim}{p_j} \right)$$

$$= \sum_{i=1}^{m} \left[\left(\sum_{j=k+1}^{n} \alpha_{ij} p_j \right)^T \left(\sum_{j=k+1}^{n} \alpha_{ij} p_j \right) \right]$$

$$= \sum_{i=1}^{m} \left(\alpha_{i,k+1} \cdot \underset{\sim}{p_{k+1}} + \alpha_{i,k+2} \cdot \underset{\sim}{p_{k+2}} + \ldots + \alpha_{i,n} \cdot \underset{\sim}{p_n} \right)^T \left(\alpha_{i,k+1} \cdot \underset{\sim}{p_{k+1}} + \alpha_{i,k+2} \cdot \underset{\sim}{p_{k+2}} + \ldots + \alpha_{i,n} \cdot \underset{\sim}{p_n} \right)$$

$$= \sum_{i=1}^{m} \left(\alpha_{i,k+1} \cdot \underset{\sim}{p_{k+1}}^T + \alpha_{i,k+2} \cdot \underset{\sim}{p_{k+2}}^T + \ldots + \alpha_{i,n} \cdot \underset{\sim}{p_n}^T \right) \left(\alpha_{i,k+1} \cdot \underset{\sim}{p_{k+1}} + \alpha_{i,k+2} \cdot \underset{\sim}{p_{k+2}} + \ldots + \alpha_{i,n} \cdot \underset{\sim}{p_n} \right)$$

$$= \sum_{i=1}^{m} \left(\sum_{j=k+1}^{n} \alpha_{ij} \underset{\sim}{p_j}^T \cdot \alpha_{ij} \underset{\sim}{p_j} \right) + \sum_{i=1}^{m} \left(\sum_{j=k+1}^{n} \sum_{L=k+1, L \neq k}^{n} \alpha_{ij} \underset{\sim}{p_j}^T \cdot \alpha_{iL} \underset{\sim}{p_L} \right)$$

$$= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \alpha_{ij} \underset{\sim}{p_j}^T \underset{\sim}{p_j} \alpha_{ij} + \sum_{i=1}^{m} \sum_{j=k+1}^{n} \sum_{L=k+1, L \neq k}^{n} \alpha_{ij} \underset{\sim}{p_j}^T \underset{\sim}{p_L} \alpha_{iL}$$

$$= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \alpha_{ij}^2 \qquad (\because \underset{\sim}{p_j}^T \underset{\sim}{p_j} = 1,\ \underset{\sim}{p_i}^T \underset{\sim}{p_j} = 0 \ \forall i \neq j)$$

$$= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \left(\underset{\sim}{x_i}^T \underset{\sim}{p_j} \right)^2$$

$$= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \left(\underset{\sim}{p_j}^T x_i \right) \left(x_i^T \underset{\sim}{p_j} \right)$$

$$= \sum_{j=k+1}^{n} \underset{\sim}{p_j}^T \left(\sum_{i=1}^{m} x_i \underset{\sim}{x_i}^T \right) \underset{\sim}{p_j}$$

$$= \sum_{j=k+1}^{n} \underset{\sim}{p_j}^T m\Sigma \underset{\sim}{p_j} \qquad \left[\because \frac{1}{m} \sum_{i=1}^{m} x_i \underset{\sim}{x_i}^T = \frac{X^T X}{m} = \Sigma \right]$$

So we want to minimize

$$\min_{\underset{\sim}{p_{k+1}}, \underset{\sim}{p_{k+2}}, \ldots, \underset{\sim}{p_n}} \sum_{j=k+1}^{n} \underset{\sim}{p_j}^T m\Sigma \underset{\sim}{p_j} \qquad \text{s.t. } \underset{\sim}{p_j}^T \underset{\sim}{p_j} = 1 \quad \forall j = k+1, k+2, \ldots, n$$

The solution to the above problem is given by the eigenvectors corresponding to the smallest eigenvalues of Σ (Proof : refer to theorems in Section 1).

Thus we select $\underset{\sim}{p_1}, \underset{\sim}{p_2}, \ldots, \underset{\sim}{p_n}$ as eigenvectors of Σ and retain only top-k eigenvectors to express the data [or discard the eigenvectors $\underset{\sim}{p_{k+1}}, \underset{\sim}{p_{k+2}}, \ldots, \underset{\sim}{p_n}$].

Here the key idea was to minimize the error in reconstructing $\underset{\sim}{x_i}$ after projecting the data on to a new basis.

♠ A quick recap :

- The eigenvectors of a matrix with distinct eigenvalues are linearly independent.
- The eigenvectors of a square symmetric matrix are orthogonal.
- PCA exploits this fact by representing the data using a new basis comprising only the top-k eigenvectors.
- The $n - k$ dimensions which contribute very little to the reconstruction error are discarded. These are also the directions along which the variance is minimum (we shall establish this in yet another interpretation of PCA).

2.3 Interpretation 3

We started off with the following wishlist.

We are interested in representing the data using fewer dimensions such that

- the dimensions have low covariance
- the dimensions have high variance

So far we have paid a lot of attention to the covariance. But what about variance? Have we achieved our stated goal of high variance along dimensions? To answer this question we will see yet another interpretation of PCA.

The i^{th} dimension of the transformed data $\hat{X}$ is given by

$$\underset{\sim}{\hat{X}_i} = X \underset{\sim}{p_i}$$

The variance along this dimension is given by

$$\begin{aligned}
\frac{\hat{\underset{\sim}{X}}_i^T \hat{\underset{\sim}{X}}_i}{m} &= \frac{1}{m} \underset{\sim}{p_i}^T X^T X \underset{\sim}{p_i} \\
&= \underset{\sim}{p_i}^T \underbrace{\frac{1}{m} X^T X \underset{\sim}{p_i}} \\
&= \underset{\sim}{p_i}^T \underbrace{\Sigma \underset{\sim}{p_i}} \\
&= \underset{\sim}{p_i}^T \lambda_i \underset{\sim}{p_i} \qquad [\because \underset{\sim}{p_i} \text{ is an eigenvector of } \Sigma] \\
&= \lambda_i \underbrace{\underset{\sim}{p_i}^T \underset{\sim}{p_i}}_{=1} \\
&= \lambda_i
\end{aligned}$$

- Thus the variance along the i^{th} dimension (i^{th} eigenvector of Σ) is given by the corresponding eigenvalue.
- Hence, we did the right thing by discarding the dimensions (eigenvectors) corresponding to lower eigenvalues!

♠ A Quick Summary : We have seen 3 different interpretations of PCA.

- It ensures that the covariance between the new dimensions are minimized.
- It picks up dimensions such that the data exhibit high variance across these dimensions.
- It ensures that the data can be represented using less number of dimensions.

♠ Note : The total variability caused by the initial feature set *i.e.* $\underset{\sim}{X_1}, \underset{\sim}{X_2}, \ldots, \underset{\sim}{X_n}$ is same as the total variability caused by the transformed feature set *i.e.* $\hat{\underset{\sim}{X}}_1, \hat{\underset{\sim}{X}}_2, \ldots, \hat{\underset{\sim}{X}}_n$. How ?

$$\begin{aligned}
\text{Total variation due to } \underset{\sim}{X_1}, \underset{\sim}{X_2}, \ldots, \underset{\sim}{X_n} &= \text{sum of the principal diagonal elements of } \Sigma \\
&= \text{trace } (\Sigma) \\
&= \text{sum of the eigenvalues of } \Sigma \\
&= \text{Total variation due to } \hat{\underset{\sim}{X}}_1, \hat{\underset{\sim}{X}}_2, \ldots, \hat{\underset{\sim}{X}}_n.
\end{aligned}$$

♠ Terminology : $\hat{\underset{\sim}{X}}_1$ is called the **First Principal Component**, $\hat{\underset{\sim}{X}}_2$ is called the **Second Principal Component** and so on ...

♠ The algorithm for PCA is an Unsupervised Learning Algorithm.