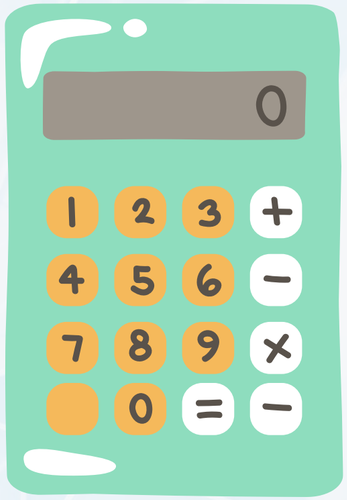
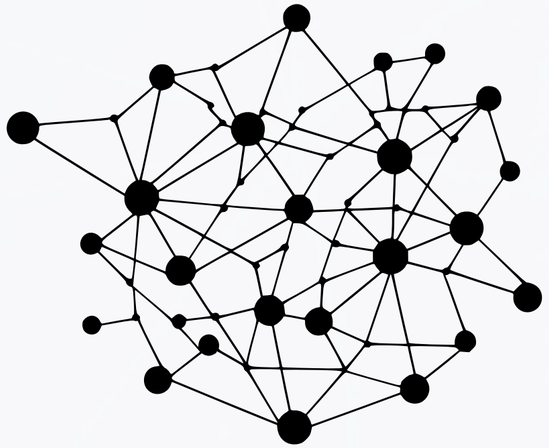
a_i to AI 3.0

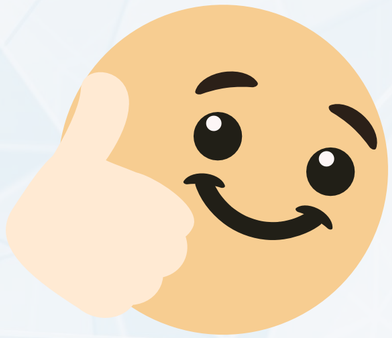


MATH IN AI



What is the role of mathematics in AI/ML?

~~"To make robots so that they can replace human beings."~~



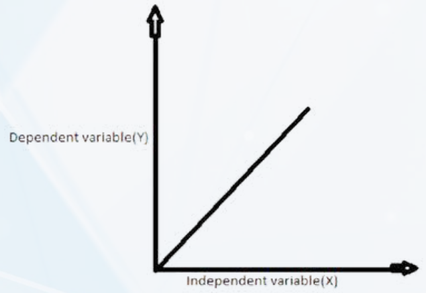
Mathematics provides the tools and frameworks for developing AI tools and algorithms.

Math in PCA

Linear Algebra

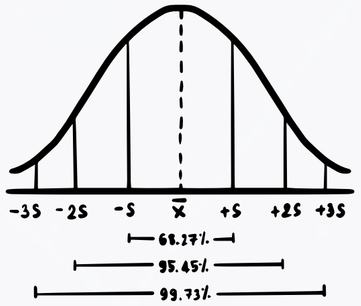
Linear Regression Equation

$$y = a + bx$$



Probability

$$\mathbb{E}[X] = \sum x_i \, p(x_i)$$



Linear Algebra

"AI might look smart, but behind the scenes, it's just Linear Algebra trying not to crash."

Eigen vectors

A non zero vector, that when multiplied by a square matrix, results in a scaled version of itself.

$$Av = \lambda v$$

A is any Matrix, v is eigen vector and λ is the corresponding eigen value.

Eigen Values

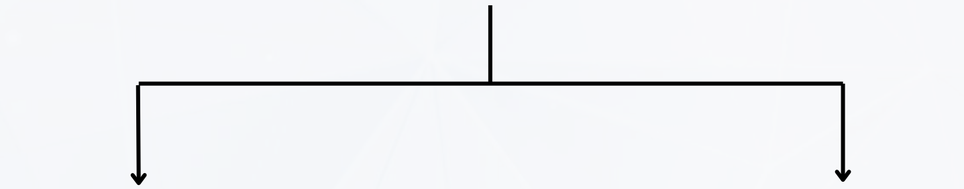
"Examples are the best way to learn concepts."

A symmetric matrix $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

To find the eigen values, we use the equation

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

The eigen values



$$\lambda = 1 \qquad \lambda = 3$$

To find the eigen vectors, use $(A - \lambda I)\, v = 0$

So the eigen vectors which we will obtain are

$$u = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Inferences:

1) Higher is the eigen value, higher is the variance in data along the corresponding eigen vector.

2) Data is most stretched along the direction of such eigen vector.

3) This direction is called First Principal Component (PC1) in PCA.

Basis

Basis for a vector space is a set of linearly independent vectors that span the entire space.

E.g. The standard basic vectors in $\mathbb{R}^2$ are $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

Then any vector w can be written as:

$$w = \alpha_1 e_1 + \alpha_2 e_2$$

where α_1, α_2 are constants

Change of Basis

It is a process of expressing a vector in one coordinate system in terms of another coordinate system.

E.g.

Vector v in $\mathbb{R}^2$ is given by $v = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$

Try to express v in a new basis formed by the vectors

$u = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $w = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

Linear Transformation Equation

$$v_{new} = B^{-1}v$$

where B is the basis matrix formed by u and w as columns

and v is the standard vector

and $v = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$

On solving the equation, we get

$$v_{new} = \begin{pmatrix} 2/3 \\ 5/3 \end{pmatrix}$$

PCA uses linear transformations to reduce the dimensionality of data

A small game

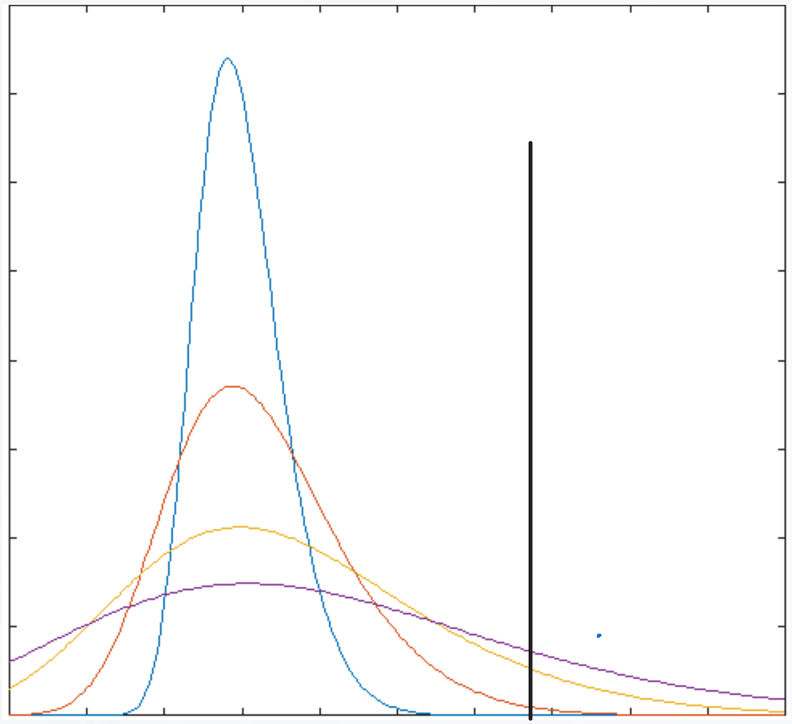
PH 1010 Quiz 2 class average is 10/20.

Is the average enough to judge where you stand in the class?

Was your answer Yes?

If given the marks of every student, what would you do??

P.S: Don't think if you've scored less in the quiz you are out of the race, endsem clutch exists :)

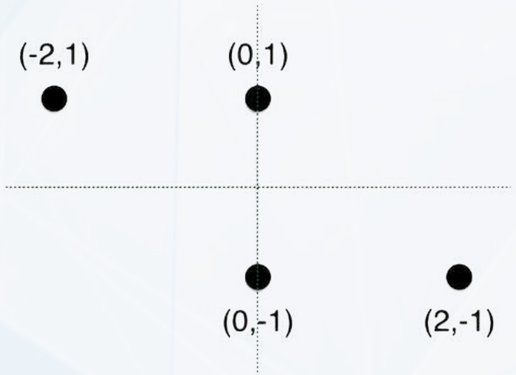


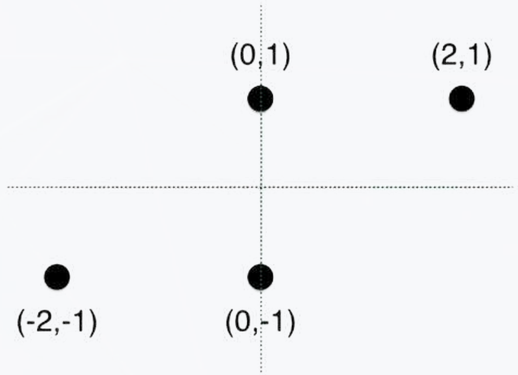
If the black line represents your score, which kind of distribution would give you your best grade?

Variance

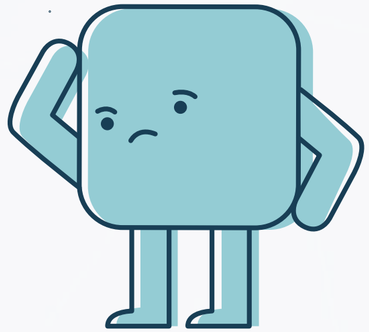
Represents how far raw scores are from the mean.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$





Is variance a sufficient parameter to describe these points completely?

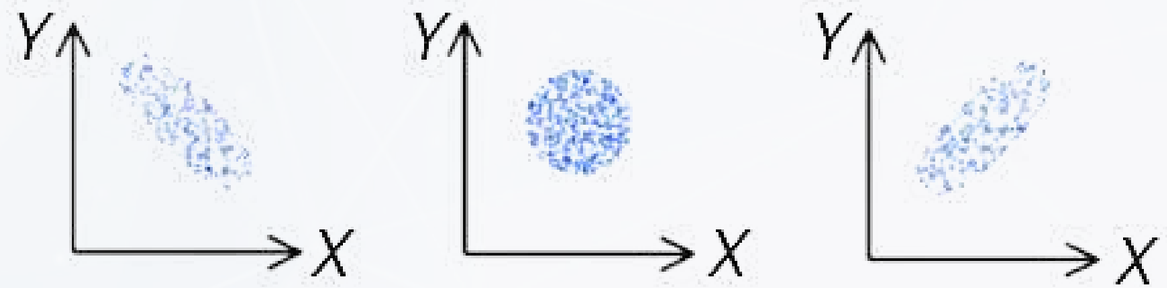


Covariance

Measures how two variables change together.

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

It is worth noting that when Y=X, the formula changes to the formula for variance.



Covariance Matrix

$$\Sigma = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Cov}(X_n, X_n) \end{bmatrix}$$

Question: Is this always a square matrix?

Question: The Royal Feast

Three friends Atreya, Prad and PK go for a biryani eating competition. The table below shows the number of various biryanis consumed by them.

	Veg	Egg	Chicken
Atreya	3	2	0
Prad	4	5	0
PK	0	5	4

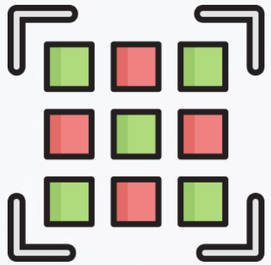
Give the covariance matrix for this question :)

A short tour back to JEE

Symmetric Matrix: $A = A^T$

Orthogonal Matrix: $A^{-1} = A^T$

Diagonal Matrix: A matrix whose non-diagonal elements are 0.



Diagonalization

Any matrix A can be diagonalized if and only if there exists an invertible matrix P such that $P^{-1}AP = D$

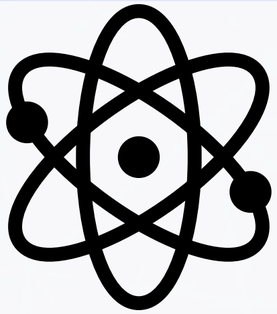
Finding P?

- Find the eigen values of A
- Find the eigen vectors corresponding to the eigen values

Fun Fact:

The matrix D is the diagonal matrix $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ where λ_i denotes the eigen values of the matrix A

An important theorem



Symmetric matrices are orthogonally diagonalizable.

Proof

Let the matrix A be diagonalizable such that $P^{-1}AP = D$
Our assumption is that P is orthogonal, implies $P^{-1} = P^T$

$$\implies D = P^T AP$$

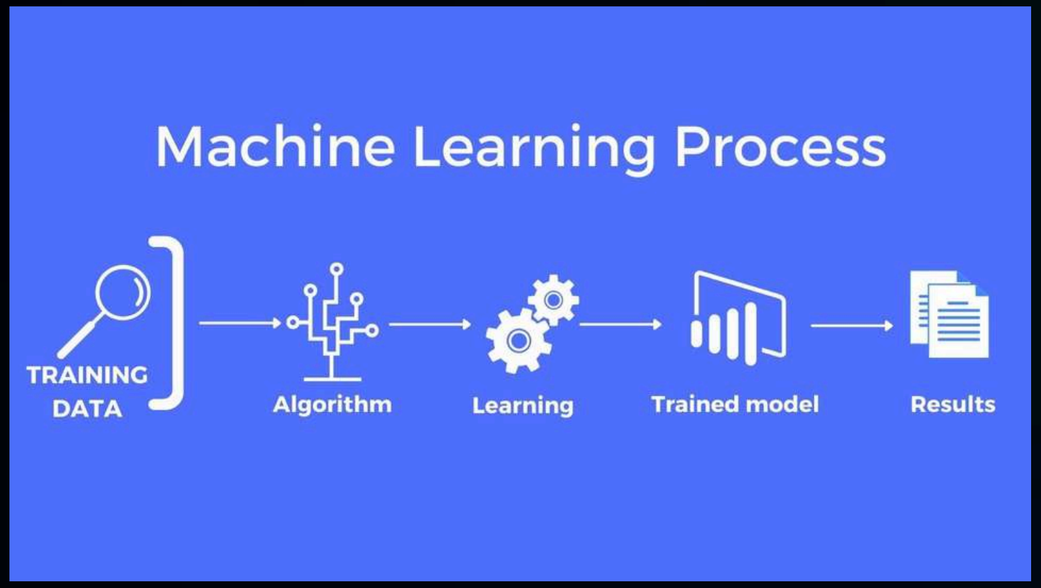
Taking transpose on both sides, we get

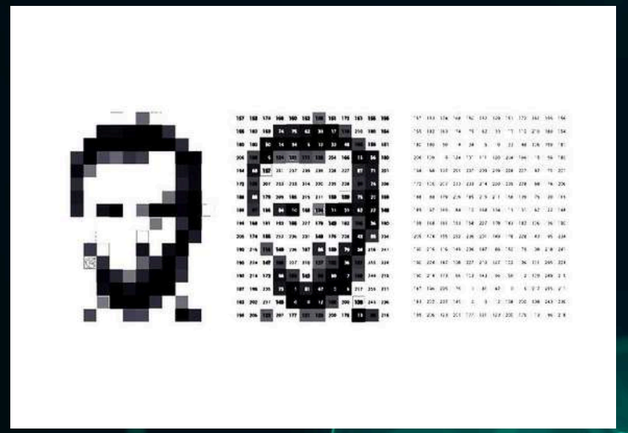
$$\implies D^T = P^T AP$$

We know that D is a diagonal matrix, hence $D = D^T$ and A is orthogonally diagonalizable

PCA

Principal component Analysis



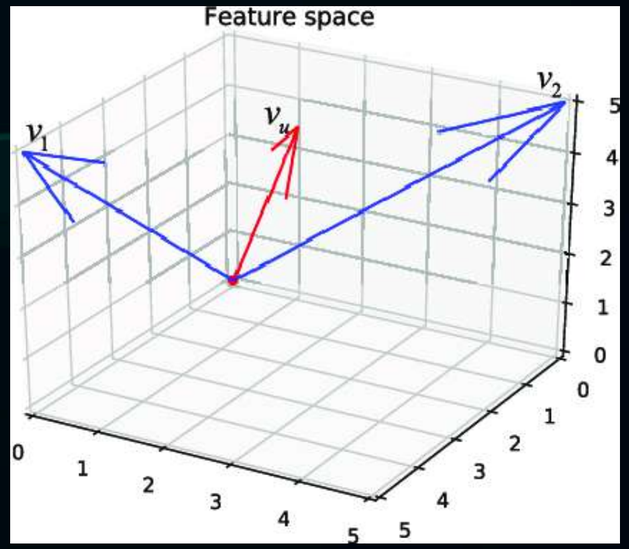


features

observations

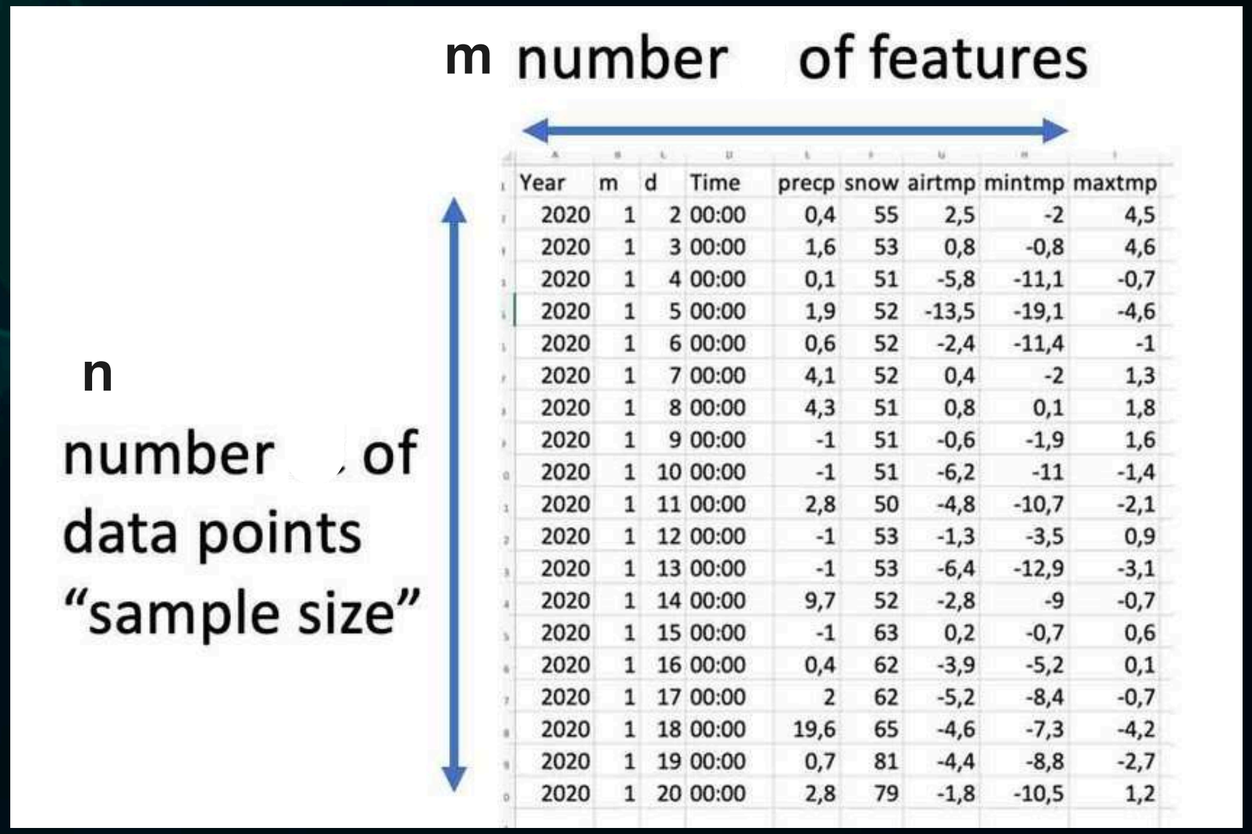
	Blood pressure	Heart rate	height	weight	...	...
Person 1						
Person 2						
Person 3						

```
[3]: {'data':        pixel1  pixel2  pixel3  pixel4  pixel5  pixel6  pixel7  pixel8  pixel9  \
     0          0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
     1          0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
     2          0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
     3          0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
     4          0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
     ...        ...     ...     ...     ...     ...     ...     ...     ...     ...
     69995      0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
     69996      0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
     69997      0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
     69998      0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
     69999      0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0

            pixel10  ...  pixel775  pixel776  pixel777  pixel778  pixel779  \
     0          0.0  ...       0.0       0.0       0.0       0.0       0.0
     1          0.0  ...       0.0       0.0       0.0       0.0       0.0
     2          0.0  ...       0.0       0.0       0.0       0.0       0.0
     3          0.0  ...       0.0       0.0       0.0       0.0       0.0

[ ]:
```



DIMENSIONALITY REDUCTION

PCA algorithm



Year	m	d	Time	precp	snow	airtmp	mintmp	maxtmp
2020	1	2	00:00	0,4	55	2,5	-2	4,5
2020	1	3	00:00	1,6	53	0,8	-0,8	4,6
2020	1	4	00:00	0,1	51	-5,8	-11,1	-0,7
2020	1	5	00:00	1,9	52	-13,5	-19,1	-4,6
2020	1	6	00:00	0,6	52	-2,4	-11,4	-1
2020	1	7	00:00	4,1	52	0,4	-2	1,3
2020	1	8	00:00	4,3	51	0,8	0,1	1,8
2020	1	9	00:00	-1	51	-0,6	-1,9	1,6
2020	1	10	00:00	-1	51	-6,2	-11	-1,4
2020	1	11	00:00	2,8	50	-4,8	-10,7	-2,1
2020	1	12	00:00	-1	53	-1,3	-3,5	0,9
2020	1	13	00:00	-1	53	-6,4	-12,9	-3,1
2020	1	14	00:00	9,7	52	-2,8	-9	-0,7
2020	1	15	00:00	-1	63	0,2	-0,7	0,6
2020	1	16	00:00	0,4	62	-3,9	-5,2	0,1
2020	1	17	00:00	2	62	-5,2	-8,4	-0,7
2020	1	18	00:00	19,6	65	-4,6	-7,3	-4,2
2020	1	19	00:00	0,7	81	-4,4	-8,8	-2,7
2020	1	20	00:00	2,8	79	-1,8	-10,5	1,2

- Feature space of m dimensions
- PCA overview-

Extracts lesser number of features from existing features while retaining much of the information given by these existing features, thus reducing the dimensions

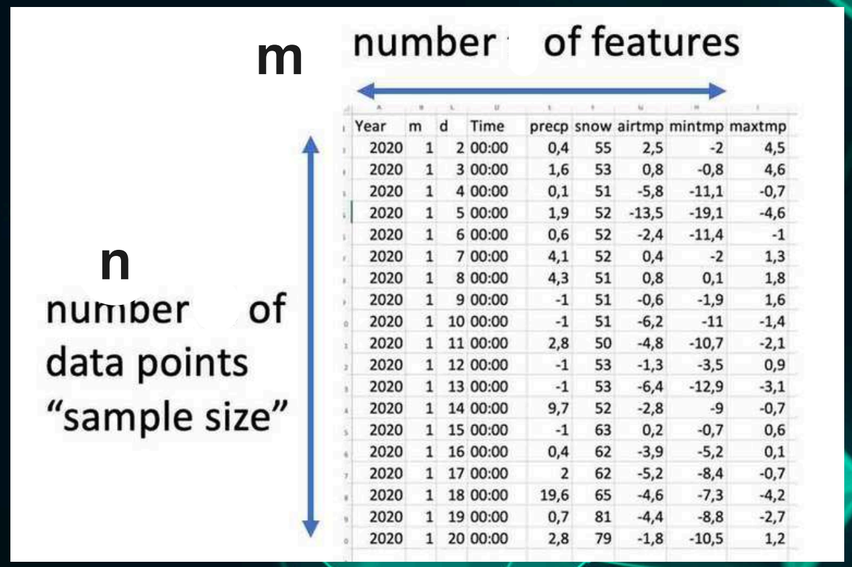
But How?

THE MATH

- All the features can be represented as vectors - like Time as X1, precep as X2 etc.

$$\mathbf{X}_1 = \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,n} \end{bmatrix}^T \quad \mathbf{X}_2 = \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{bmatrix}^T \quad \dots \mathbf{X}_m = \begin{bmatrix} x_{m,1} \\ x_{m,2} \\ \vdots \\ x_{m,n} \end{bmatrix}^T, \mathbf{X}_i \in R^{1 \times n}$$

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{bmatrix} \in R^{m \times n}, \quad \mathbf{X}_i^T \in R^n$$



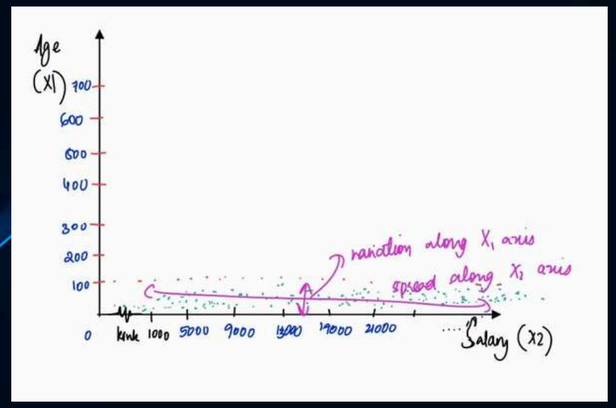
Year	m	d	Time	precp	snow	airtmp	mintmp	maxtmp
2020	1	2	00:00	0,4	55	2,5	-2	4,5
2020	1	3	00:00	1,6	53	0,8	-0,8	4,6
2020	1	4	00:00	0,1	51	-5,8	-11,1	-0,7
2020	1	5	00:00	1,9	52	-13,5	-19,1	-4,6
2020	1	6	00:00	0,6	52	-2,4	-11,4	-1
2020	1	7	00:00	4,1	52	0,4	-2	1,3
2020	1	8	00:00	4,3	51	0,8	0,1	1,8
2020	1	9	00:00	-1	51	-0,6	-1,9	1,6
2020	1	10	00:00	-1	51	-6,2	-11	-1,4
2020	1	11	00:00	2,8	50	-4,8	-10,7	-2,1
2020	1	12	00:00	-1	53	-1,3	-3,5	0,9
2020	1	13	00:00	-1	53	-6,4	-12,9	-3,1
2020	1	14	00:00	9,7	52	-2,8	-9	-0,7
2020	1	15	00:00	-1	63	0,2	-0,7	0,6
2020	1	16	00:00	0,4	62	-3,9	-5,2	0,1
2020	1	17	00:00	2	62	-5,2	-8,4	-0,7
2020	1	18	00:00	19,6	65	-4,6	-7,3	-4,2
2020	1	19	00:00	0,7	81	-4,4	-8,8	-2,7
2020	1	20	00:00	2,8	79	-1,8	-10,5	1,2

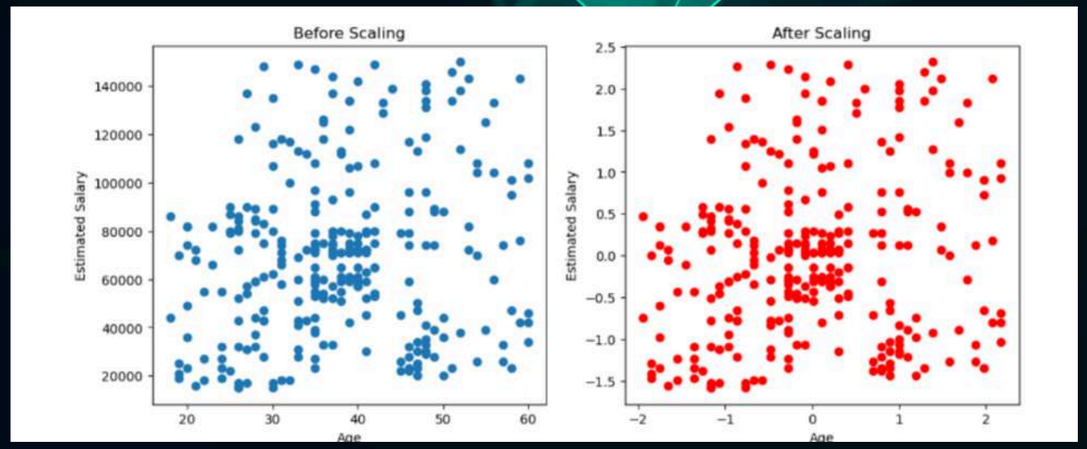
- **X** is the matrix that has the feature vector as its rows and datapoints as it’s columns, it is the data matrix.
- Question - How many observations are seen in X1 ? Is it the same as Xm and the remaining feature vectors?
- The columns are the observations.
- **Standardization - The data is made zero centered in each of the feature vectors , the mean is made to be zero.**

The math… **Standardization…**

- **As we will see in the upcoming slides, this algorithm is heavily influenced by the variance of the data or how it spreads**
- **So now if one feature (X1) takes on very high values and another (X2) takes on very small values , the spread of the data on X1 axis will be very high when compared to X2 axis and this will affect the PCA algorithm in an unwanted way**
- **Which is why we standardize the data - all variables will be transformed to the same scale**
- **And now each feature has mean = 0 and std. deviation as 1.**
-

$$Z = \frac{x - \mu}{\sigma}$$





- **If we didn’t standardize, salary would contribute more in this algorithm just because it takes on larger values**

COVARIANCE MATRIX

- Recollect the covariance formula -

The formula for the covariance $\mathrm{Cov}(X, Y)$ between two sets X and Y of n observations each is given by:

$$\mathrm{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, respectively.

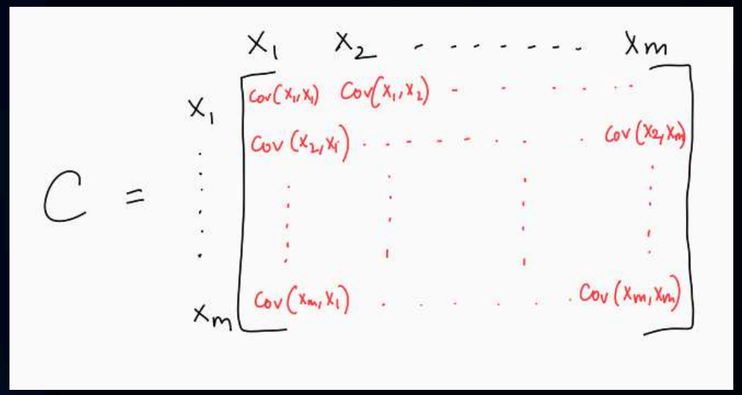
- In this formula, X can be thought of as [x1, x2, x3.....xn] and Y as [y1, y2, y3, ...yn]
- Similarly we can compute the covariance between two feature vectors like X_1 and X_2, as they too are a set of observations and between the same feature vector as well
- So, the new formulas of Covariance will be :

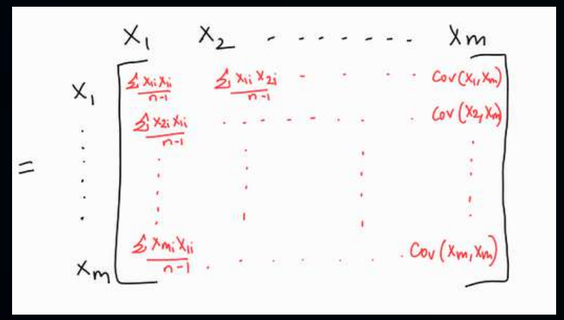
$$\mathbf{Cov}(X_i, X_j) = \frac{1}{n-1} \sum_{k=1}^{n} x_{ik} x_{jk} = \mathbf{C}_{ij}$$

$$\mathbf{Cov}(X_i, X_i) = \frac{1}{n-1} \sum_{k=1}^{n} x_{ik} x_{ik} = \mathbf{C}_{ii} \quad (i = j)$$

- Now, we make a matrix, where each element represents the covariance between all the feature vectors. What can we say about diagonal elements??

COVARIANCE MATRIX





$$= \frac{1}{n-1} \begin{pmatrix} \mathbf{x_1}\mathbf{x_1}^T & \mathbf{x_1}\mathbf{x_2}^T & \cdots & \mathbf{x_1}\mathbf{x_m}^T \\ \mathbf{x_2}\mathbf{x_1}^T & \mathbf{x_2}\mathbf{x_2}^T & \cdots & \mathbf{x_2}\mathbf{x_m}^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x_m}\mathbf{x_1}^T & \mathbf{x_m}\mathbf{x_2}^T & \cdots & \mathbf{x_m}\mathbf{x_m}^T \end{pmatrix} \in \mathbb{R}^{m \times m}$$

$$\mathbf{C_X} = \frac{1}{n-1}\mathbf{X}\mathbf{X}^T$$

X is the data matrix.
So this is the formula for
THE COVARIANCE MATRIX OF X

- Is there anything special about this matrix? What is it's use?
- Till now I just explained how we can represent the data and it's features and their relationships in a matrix

FEATURE EXTRACTION...

- PCA produces a new set of features by linearly combining the existing features.
- And we can represent the data in a new matrix Y using the new features .

$$Y = PX$$

- *This relation stems from the fact that we are only doing linear transformations on X1, X2, X3.... or the rows of X.*
- **X is a m x n matrix, P is taken to be a m x m matrix giving Y as an m x n matrix**
- So P is the matrix that compresses the data into a simpler form while retaining as much as info possible and we have to find this P----> Transformation matrix

okay, but where is the compression or dimensionality reduction happening?

LET US THINK ABOUT THE NEW FEATURE SPACE...

- These new features axes or basis, are collectively called the Principal components:
- The rows of P ---> the principal components

HOW WOULD WE LIKE THEM TO BE?

- The new features should be unrelated to each other, Why?
- We'd want the new features to keep as much as info possible from the original feature vectors .

- Information - the variation in the data, the spread of the data in the original feature space
- So we'd want our principal components capturing the maximum variance in the data or in other words align themselves in the direction of maximum variance.

- Where do you think these two requirements would reflect?

COVARIANCE MATRIX OF Y :

- Since we want every other feature to be unrelated - every element apart from diagonal elements to be zero.
- Now let us see how we derive the Principal components mathematically :

$$\mathbf{C_Y} = \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T = \frac{1}{n-1}(\mathbf{PX})(\mathbf{PX})^T = \frac{1}{n-1}(\mathbf{PX})(\mathbf{X}^T\mathbf{P}^T) = \frac{1}{n-1}\mathbf{P}(\mathbf{X}\mathbf{X}^T)\mathbf{P}^T$$

$$\text{i.e.} \quad \mathbf{C_Y} = \frac{1}{n-1}\mathbf{P}\mathbf{S}\mathbf{P}^T \quad \text{where} \quad \mathbf{S} = \mathbf{X}\mathbf{X}^T$$

- S is a symmetric matrix , of shape m x m. So it is orthogonally hence, orthonormally diagonalizable, hence

$$\mathbf{S} = \mathbf{E}\mathbf{D}\mathbf{E}^T$$

- Where E is a matrix whose columns are the orthonormal Eigen Vectors of S and D is a diagonal matrix with eigen values of S as entries .

where:
- E is the matrix of eigenvectors:

$$E = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \end{bmatrix}$$

Each column e_i is an eigenvector of S.
- D is the diagonal matrix of corresponding eigenvalues:

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues corresponding to the eigenvectors $e_1, e_2, \ldots, e_n$.

- **Now, if we choose $P = E^T$, Then rows of P are the eigen vectors , Thus**

The covariance matrix C_y can be written as:

$$C_y = \frac{1}{n-1}\mathbf{PSP}^T$$

$$= \frac{1}{n-1}\mathbf{PEDE}^T\mathbf{P}^T$$

$$= \frac{1}{n-1}\mathbf{D}$$

- as the eigen vectors are orthonormal, their magnitude is 1
- So now, from Cy we find that the eigen values are equal to the variance of the data, more the EV more the Variance in that direction.
- Rearrange the eigen values in D, with the largest one coming first, this means that even the eigen vectors in E and P are rearranged. - Not an issue at all and **we take the first r PCs (first r eigen vectors)**

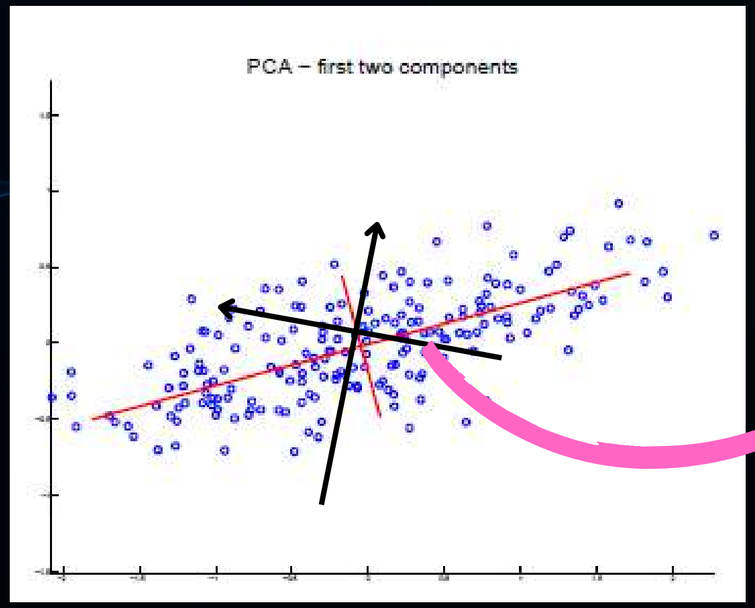
Thus dimensions are reduced.

PCs ---- Perpendicular to each other and oriented along max variance

BUT WAIT!

$$P = E^T$$

- Amongst the eigen vectors, max Eigen value corresponds to eigen vector along which max variance exists, but how does that mean are the directions that correspond to max variance amongst all the set of directions possible?? What makes Eigen vectors so special?



WHY NOT THIS?

The Proof :

Take vector $w \rightarrow$ along which 'X' has max variance

$\Rightarrow$ Var $(X \cdot w)$ is high

$\text{Var}(X \cdot w) = w^T \cdot C_x w$, where $w^T \cdot w = \|w\|^2 = 1$

The w for which the RHS is max is the vector we are finding

and it is maximum when $Cw = \lambda w$.

which turns out to be satisfied by Eigen vectors.

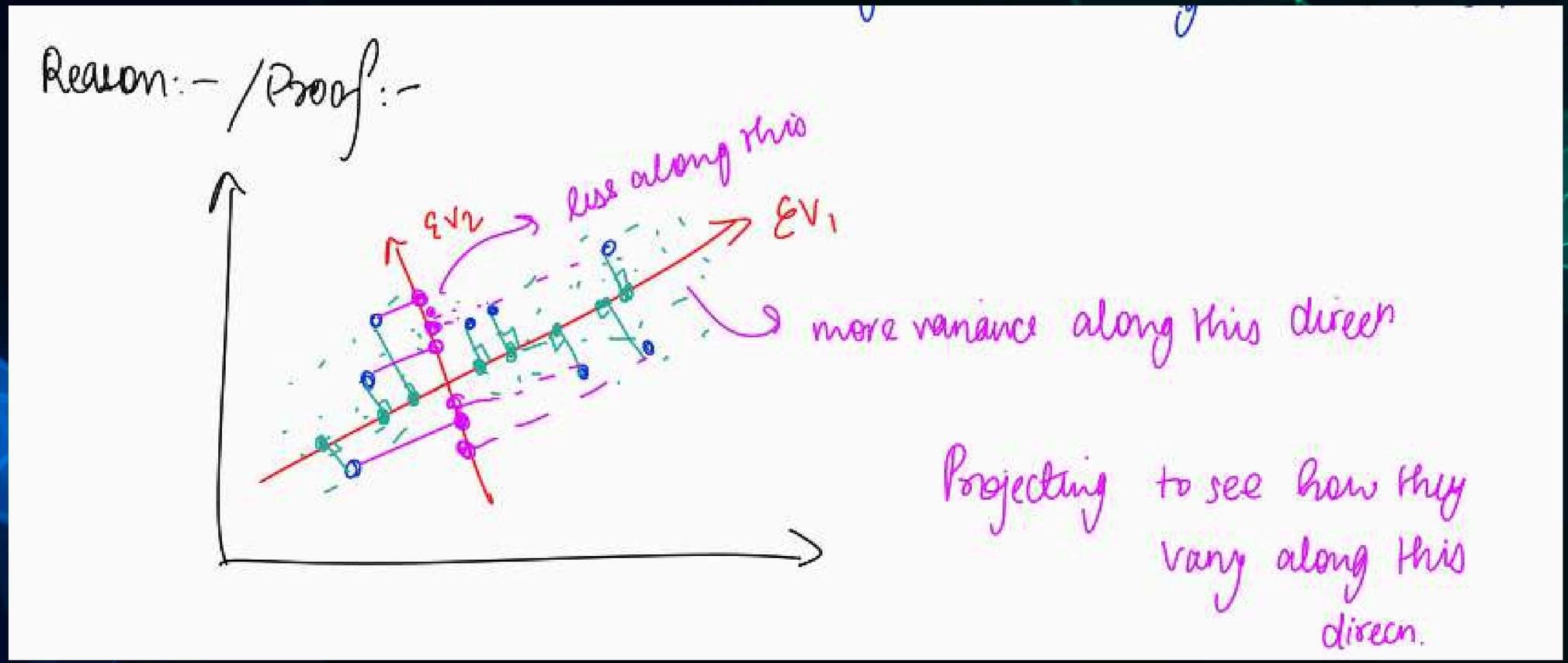
$$\therefore \quad \text{Var}(Xw) = w^T C_x w$$

$$= w^T \cdot \lambda w$$

$$= \lambda \|w\|^2 = \lambda$$

$$\therefore \boxed{\text{Var}(X \cdot w) = \lambda}$$

Visually,


Reason:- /Proof:-
EV2
Rise along this
EV1
more variance along this direcn
Projecting to see how they vary along this direcn.

Recapping,

1. Standardize the data, to prevent certain features with larger range from getting an unfair dominance in the analysis
2. Find the data matrix (X) and Cx (Covariance matrix) given as k S where S = XX^T
3. Find P = E^T, where E is the eigen vector matrix of S
4. Arrange P in the order of Lamda in D.
5. Feel happy that you've got your PCs
6. Find the data (Y) now when transformed from the original feature space to the PC space.
7. Feel happy that you've done PCA

Our Socials

MATH CLUB

AI CLUB