Linear Algebra

Vectors:

Vector: A vector is just an arrow in the space (physics) or just a list of numbers (CS). Vectors allow us to graphically represent the data to find certain patterns.

Linear Combination, Span & Basis:

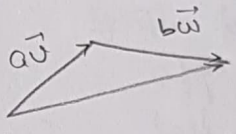
* For an N-dim space, there are N-basis vectors (basis of that coordinate system). These set of N-vectors can be scaled by scalars to represent each vector in the space. There can be many set of basis vectors.

  Eg:- for xy-plane: $\hat{i}, \hat{j}$

* Linear combination of vectors: scaling and adding vectors

$$a\vec{v} + b\vec{w}$$

(a: scalar)



* The set of all possible vectors we can get by the linear combination of those 2 vectors, is called the span of these two vectors.

In 2D space, most pair of vectors span the entire 2D space except when they line up (in this case, they only span a line) or when both vectors are $\vec{0}$. (they span the origin)

* A vector can be represented by just a point at the tip of the vector (tail always sits at origin)

* In a 3D space, span of 2 vectors is a plane passing through the origin. If we use 3 vectors then, the span would be the entire 3D space, except when the 3rd vector lies on the span of the first two vectors. In this case it does not contribute to the span. In this case, we can say that the the three vectors are linearly dependent.

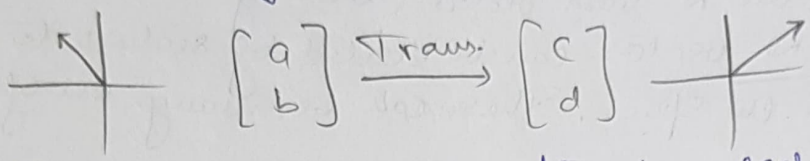
Whenever we have a group of vectors and we can remove some without hurting the span, they are linearly dependent. This means that some of the vectors can be expressed as linear combination of others.

On the other hand, if each vector in a group of vectors adds another dimension to the span, then they are called linearly independent.

Basis of a vector space: A set of linearly independent vectors that span the full space.

Linear Transformation & Matrices:

~~Linear~~ Transformation is basically taking an input vector and move it (transform it) into the output vector.



* Transformations in fact transform each input point (all possible points in space) to another set of points.

* Linear Transformations: after transformation, all the grid lines must remain lines and origin must be fixed at its original location. Even the diagonal lines must remain lines and not curve.

  In linear transformation, grid lines remain parallel and evenly spaced.

* Linear Transformations can be represented by the transformation of the basis vectors (tip of the basis vectors after transformation).

$$\vec{c} = a\hat{i} + b\hat{j} \xrightarrow[\text{Trans}]{\text{Lin}} \vec{c_1} = a\hat{i_1} + b\hat{j_1}$$

($\hat{i_1}$, $\hat{j_1}$ ← transformed)

Thus, we only need the final coordinates of $\hat{i}$ and $\hat{j}$ to find the final position of each point in the input space.

For 2D space we only need 4 numbers to describe linear transformation.

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

($\begin{bmatrix} c \\ d \end{bmatrix}$ ← coordinate of $\hat{j}$; $\begin{bmatrix} a \\ b \end{bmatrix}$ ← coordinate of $\hat{i}$)

In matrix form,

$$\vec{a} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

$$\vec{a_1} = \begin{bmatrix} 1 & 3 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

Eg:- $\vec{a} = -1\hat{i} + 2\hat{j}$

After transformation,

$$\hat{i} = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \quad \hat{j} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

Then,

$$\vec{a_1} = -1\begin{bmatrix} 1 \\ -2 \end{bmatrix} + 2\begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} -1 \\ 2 \end{bmatrix} + \begin{bmatrix} 6 \\ 0 \end{bmatrix}$$

$$\vec{a_1} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = x\begin{bmatrix} a \\ c \end{bmatrix} + y\begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} ax+by \\ cx+dy \end{bmatrix}$$

* Columns of a transformation matrix are the final coordinates of the basis vectors after the transformations.

Matrix Multiplication as Composition:

In case of applying multiple transformations, a transformation matrix can be constructed for the overall effect using the final $\hat{i}$ and $\hat{j}$.

Eg:-

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{\text{Shear}}\left(\underbrace{\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{rotation}}\underbrace{\begin{bmatrix} x \\ y \end{bmatrix}}_{\text{original vector}}\right) = \underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{overall}}\begin{bmatrix} x \\ y \end{bmatrix}$$

$f(g(x))$

Multiplying matrices is eq. to applying one transf. then another.

$$\Rightarrow [\text{Shear}][\text{rotation}] = [\text{overall}]$$

* The transformations are applied from right to left.

Three Dimensional Linear Transformations:

In case of 3D, we have 3 basis vectors $\hat{i}, \hat{j}, \hat{k}$.

3D transformations transform every point in 3D space to other points in 3D space.

$$\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

$\hat{i}$ $\hat{j}$ $\hat{k}$

(final coordinates)

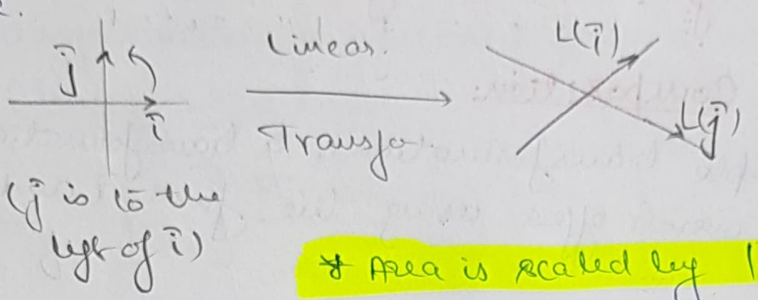
The Determinants:

Considering 2D-space, initially, the area occupied by $\hat{i}$ and $\hat{j}$ is 1 unit. After a linear transformation, the area occupied by transformed $\hat{i}$ and $\hat{j}$ will be scaled. If the new area is '0' then we say that the determinant of the transformation is 0.

Every region's area will be scaled by 0 after the transformation, since any shape can be represented by a collection of squares and grid lines remain parallel and evenly spaced.

The determinant of a 2D transformation is 0, if the 2D space is mapped to a line or a single point.

Depending on whether or not the space is flipped, the determinant can be -ve.



(ĵ is to the left of î)

L(ĵ) is to the right of L(î)

det → -ve

* Area is scaled by |det|

* Imagine ĵ is fixed and î is slowly moving counterclockwise. Initially, det = 1 and slowly reduces as î gets closer to ĵ. When î and ĵ align, det = 0. At this point all the 2D space is squished onto a line (as the span of î and ĵ is that line). As î keeps rotating ccw, det becomes -ve

Determinants in 3D:

In 3D, det (transformation) is amount by which the volume of cube having edges î, ĵ and k̂ is scaled.

In 2D, areas are scaled into parallelograms. In 3D, cubical volumes are turned into parallelopiped whose vol. is the determinant of transformation matrix multiplied by the original cubical area.

* If the transformation squishes the 3D space on a plane, det = 0 as the vol. is now zero (linearly dependent columns of transformation matrix).

* If $\hat{i} \times \hat{j} = \hat{k}$ holds after transformation, det is +ve and if $\hat{i} \times \hat{j} = -\hat{k}$ ⇒ det → -ve.

$$\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

$$\det\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = a \cdot \det\begin{bmatrix} e & f \\ h & i \end{bmatrix} - b \cdot \det\begin{bmatrix} d & f \\ g & i \end{bmatrix} + c \cdot \det\begin{bmatrix} d & e \\ g & h \end{bmatrix}$$

$$\det(M_1 M_2) = \det(M_1)\det(M_2)$$

Inverse Matrices, Column Space and Null Space:

Matrices are used to solve linear system of equations

$$\begin{aligned} 2x + 5y + 3z &= 8 \\ 4x + 0y + 9z &= -3 \\ 2x + 4y + 0z &= -2 \end{aligned} \equiv \begin{bmatrix} 2 & 5 & 3 \\ 4 & 0 & 9 \\ 2 & 4 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 8 \\ -3 \\ -2 \end{bmatrix}$$

This means that we are basically learning the transformation of space so that the input vectors are transformed to output.

$$A\vec{x} = \vec{v}$$

Considering 2D, when $\det(A) \neq 0$, $\vec{x}$ is transformed to $\vec{v}$. If we apply A^{-1} to $\vec{v}$, we get back x.

$$A^{-1}A = \mathbb{I} \text{ (no transformation)}$$

$$A^{-1}A\vec{x} = A^{-1}\vec{v}$$
$$\vec{x} = A^{-1}\vec{v}$$
inverse transformation

* Even in higher dimension, if there exists a transformation that does not squish space into lower dim, that is $\det \neq 0$. There exists a unique transformation A^{-1}, such that performing A and then A^{-1} on a vector leaves that vector unchanged.

* If $\det = 0$, A^{-1} does not exist since it is not possible to transform a vector in lower dim into a higher dim.

* Sol. can still exist when $\det = 0$, the output vector (after transformation) must live on the squished space. Eg:- a line for 2D case.

Rank: The number of dim in the output of a transformation. (No. of lin. independent columns in the matrix)

Eg:- If the transformation squishes the space into a line, the rank of transformation = 1, for a plane, rank = 2.

So, for a 3×3 matrix, rank = 2 implies that the transformation squishes to a plane ($\det = 0$)

Column Space: Set of all possible outputs of a transformation (matrix) whether it is a line, point or plane is called column space of matrix 'A'.

* Since, columns of a matrix are the positions of basis vectors of the transformation, column space is basically the span of columns of the matrix.

Thus Rank is the no. of dimensions in the column space.

Full Rank Matrix: When the rank of a matrix is the same as the no. of columns, it's called full rank matrix. This means the transformation retains the dimensionality.

* $\vec{0}$ (zero vector) is always included in the column space, since linear transformation must keep the origin fixed.
* For full rank transformation, only $\vec{0}$ lands on itself (origin). When the transformation is not full rank (space gets squished), an infinite no. of vectors will land on $\vec{0}$ (origin).

Eg- If a 2D transformation is squished to a line, a whole line of vectors land on origin.

3D ⟶ 2D : line of vectors → origin.
3D ⟶ 1D : plane of vectors → origin.

Null space or kernel: The set of vectors that land on the origin after the transformation is called the null space or kernel.

It is the space of all vectors that become null $\vec{0}$.

* In a set of linear eq. where $A\vec{x} = \vec{0}$, the null space gives all the possible soln. to the eqn.

Non Square Matrices (transformation between dimensions)

$\begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix}$ } 3 coordinate, 3D output space

$\underbrace{\begin{bmatrix} i & j \end{bmatrix}}$ 2 basis vectors (2D input space)

Full rank (no. of dimensions in column space equals no. of dim in the input space)

The matrix on the left transforms any plane into some plane in 3D space passing through the origin.

* a×b matrix implies mapping b dim to a dim.

Similarly,

2×3 matrix
3D → 2D

$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$ } 2 coordinates, 2D output

$\underbrace{\begin{bmatrix} i & j & k \end{bmatrix}}$ 3 basis vector, input: 3D

$\begin{bmatrix} a & b \end{bmatrix}$ 2D → 1D.
$\begin{bmatrix} i & j \end{bmatrix}$

i and j have single coordinate, their location on the number line

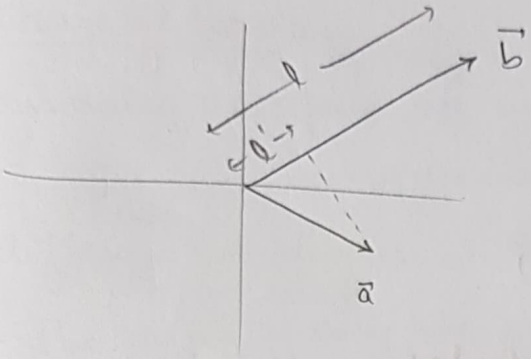
Dot products and duality:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} \cdot \begin{bmatrix} d \\ e \\ f \end{bmatrix} = ad + be + cf$$

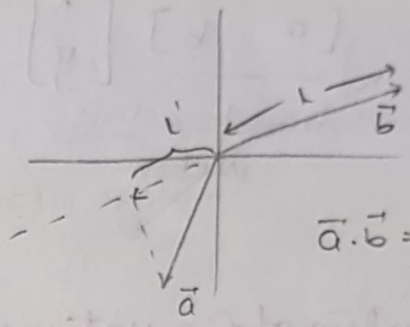
For ⊥ar vectors, length of projection $l' = 0$

$\therefore \vec{a} \cdot \vec{b} = 0$

$\vec{a} \cdot \vec{b} = \vec{b} \cdot \vec{a}$

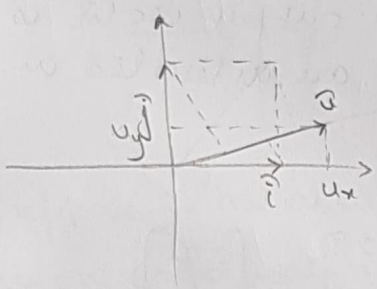


$\vec{a} \cdot \vec{b} = l \times l'$



$\vec{a} \cdot \vec{b} = -l \times l'$

Why dot product can be interpreted as projection?



We can think of the dot product as a transformation mapping a 2D (a any D) to a number line (having unit vector $\hat{u}$).

Then after transformation, both $\hat{i}$ and $\hat{j}$ must land on $\hat{u}$.

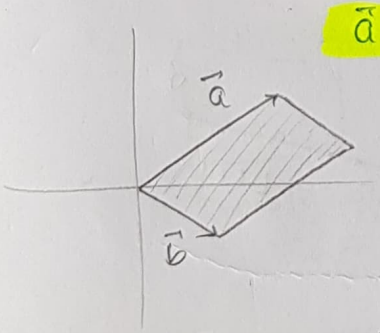
Now, projecting $\hat{i}$ on $\hat{u}$ is same as projecting $\hat{u}$ on $\hat{i}$ and same way for $\hat{j}$.

Thus, the transformation matrix is

$$[u_x \; u_y]$$

Cross Product:

$\vec{a} \times \vec{b}$ = area of the ||gm. enclosed by the vectors



$\vec{b} \times \vec{a} = +ve$

$\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}$

$\hat{i} \times \hat{j} = \hat{k}$

* For 2D, the cross product of 2 vectors is just.

$$\det \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

(vectors)

This is because we are basically scaling the unit area $\hat{i} \times \hat{j}$ after transforming them to this new location.

In reality, cross product of $\vec{a}$ and $\vec{b}$ will be a vector ⊥ar to the ||gm having a length equal to the area of ||gm. and direction based on the cross product

* For any 2 vectors of given length, cross product's magnitude is max when they are orthogonal.

Duality: Any time we have a linear transformation from some space to the number line. The transformation is associated with a unique vector in that space such that:

$$\underbrace{\begin{bmatrix} a & b \end{bmatrix}}_{\text{transformation matrix}} \begin{bmatrix} x \\ y \end{bmatrix} = \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\text{associated vector (dual vector)}} \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

Linear System of Equations and Cramer's Rule:

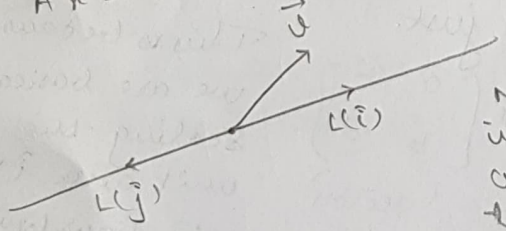
In linear system of eq we find the input vector which when transformed lands on the output vector. Remember that the output vector is a scaled version of the transformed basis vectors. We are interested in finding these scaling parameters.

$$\begin{bmatrix} 3 & 2 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -4 \\ -2 \end{bmatrix} \equiv x \begin{bmatrix} 3 \\ -1 \end{bmatrix} + y \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} -4 \\ -2 \end{bmatrix}$$

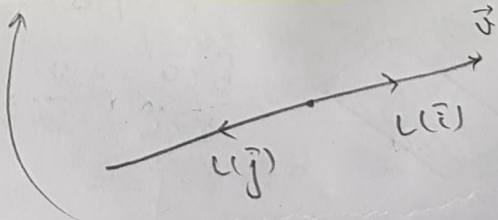
Considering 2D, if the input and output both have 2 dim, unique x and y will exist. (unique soln).

But, if transformation squishes the space to 1D, then depending on whether or not the output vector lands on that line, there can be no soln or infinite soln. ($\det = 0$)

$A\vec{x} = \vec{v}$



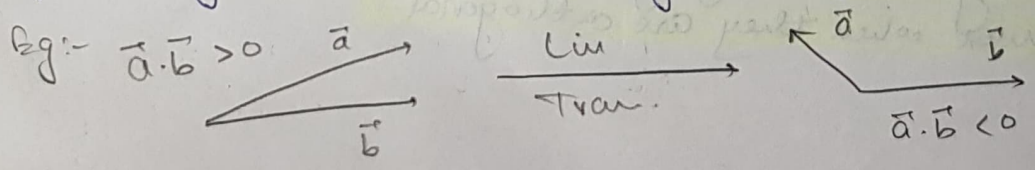
No possible input vector ($\vec{x}$) when passed through the transformation will land on output ($\vec{v}$).



There are ∞ input vectors (corresponding to a line) that will land at output $\vec{v}$ after transformation

*we are concerned with the case when $\det \neq 0$ (unique soln).

*The dot product of two vectors changes after transformation.



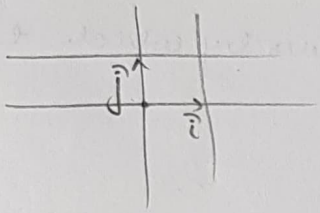
Orthonormal Transformations: Transformation that preserve the dot product of vectors. (rotation)

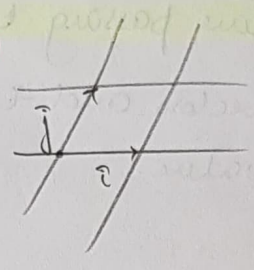
Change of Basis:

Considering 2D space: we use $\hat{i}$ and $\hat{j}$ as basis vectors.

If we use different basis vectors, the same vector will have different coordinates in different basis vector system.

* The origin of any two system must align.

① 

② 

The direction of grid lines and spacing between them can vary depending on the basis.

Suppose in an alternate basis system, the coordinates of a vector is $\begin{bmatrix} -1 \\ 2 \end{bmatrix}$.
In our coordinate system, the alternate bases are $\hat{i} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\hat{j} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$.
Then, in our coordinate system, the vector is:

$$-1\begin{bmatrix} 2 \\ 1 \end{bmatrix} + 2\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \end{bmatrix}$$

$\hat{i}$ $\hat{j}$ — alternate basis — vector in alternate basis system — vector in our system.

Similarly:

$$\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}^{-1}\begin{bmatrix} -4 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

transformation from our coordinate system to alternate

Linear Transformation in Alternate Basis:

Suppose we have a vector in alternate basis ②. We want a transformation to rotate it by 90°. Since, it is difficult to track $\hat{i}$ and $\hat{j}$ in ② after 90° rotation, we can convert to our system ①, rotate and transform back to ②.

$$\underbrace{\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}^{-1}}_{\text{convert to ②}}\underbrace{\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}}_{\text{rotate in ①}}\underbrace{\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}}_{\text{convert to ①}}\underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\text{vector in ②}} \equiv \underbrace{\begin{bmatrix} 1/3 & -2/3 \\ 5/3 & -1/3 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}}_{\text{rotation in ②}}$$

* In maths, this is often seen as $A^{-1}MA$ transformation.
Shift in basis vectors

Eigenvalues & Eigenvectors:

During transformation, most of the vector's span gets changed (their angle/orientation changes).

But, there are some vectors whose span is conserved, only its length gets changed (same happens with all the other vectors on its span).

*The span of a vector is a line passing through that vector.

These vectors are called eigenvectors and the values by which they get scaled is called its eigen value.

Importance of Eigenvectors:

For a spin of 3D space along any arbitrary axis, the vector that is unchanged and has eigenvalue of 1 is the axis of rotation.

$$\lambda = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}$$

Given A, find $\vec{v}$ and λ

$$A\vec{v} = \lambda\vec{v}$$
$$\Rightarrow A\vec{v} = \lambda I\vec{v}$$
$$\Rightarrow (A - \lambda I)\vec{v} = \vec{0}$$

($(A-\lambda I)$: matrix; $\vec{v}$: must be non-zero)

* -ve eigenvalue ⇒ the eigenvector is flipped and then scaled by the trans.

Now, we have $\vec{v}$ being transformed by $(A-\lambda I)$ into a $\vec{0}$ vector. This is only possible if the transformation squishes the space into a lower dim
$\Rightarrow \det = 0$. Thus, $\det(A-\lambda I) = 0$

From, this we get the value of λ.

we can plug each value of λ in $(A-\lambda I)\vec{v} = \vec{0}$ to get values of $\vec{v}$.

* Some transforms may have no eigenvectors. eg: rotation in 2D. The eigenvalues will be imaginary.

Eigenbasis:

Sometimes, the basis vectors can be eigen vectors. Eg: $\hat{i} \rightarrow -\hat{i}$ & $\hat{j} \rightarrow 2\hat{j}$

The transformation matrix will be a diagonal matrix in this case.

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

(Eigen vectors)

Computation using diagonal matrices is easier

Eg:- $\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 9 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 27 & 0 \\ 0 & 6 \end{bmatrix}$

Thus given any linear transformation, ~~we can~~ if we have enough eigen vectors, we can change our basis so that they become eigen vectors. Then we can perform the transformation and revert back to our basis system.

The advantage is that the overall matrix will be diagonal.

$$\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$

new basis — overall transformation matrix

transformation — eigenvectors

* A set of eigenvectors taken as basis are called eigen basis.

Mitesh Khapra (Deep Learning)

Lec 6.2: Linear Algebra Basics:

Basis: Set of linearly independent vectors that span the space

Any vector in that space can be expressed as the linear combination of these basis vectors.

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

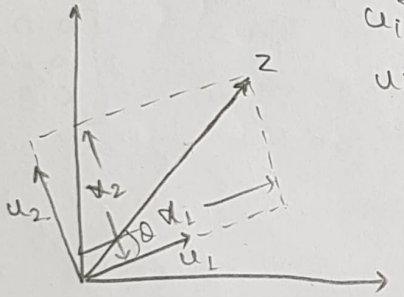
↑ vector ↑ basis ↑

To solve for x_1 and x_2, we solve the system of lin. eqn

$a = 2x_1 + 5x_2$ * Gaussian Elemination is used.

$b = 3x_1 + 7x_2$ $O(N^3)$

Now, consider the special case when the vectors (basis) is an orthonormal basis.



$u_i^T u_j = 0 \quad \forall\, i \neq j$

$u_i^T u_i = \|u_i\|^2 = 1$

For any vector z.

$$z = \alpha_1 u_1 + \alpha_2 u_2 + \dots + \alpha_n u_n$$

$$\Rightarrow u_1^T z = \alpha_1 u_1^T u_1 + \alpha_2 u_1^T u_2 + \dots + \alpha_n u_1^T u_n$$

($u_1^T u_1 = 1$; $u_1^T u_2 = 0$; $u_1^T u_n = 0$)

$$\Rightarrow \alpha_1 = u_1^T z$$

Thus, in case we have orthonormal basis, the coefficients can be found just by dot products. $O(n^2)$

$\alpha_1 = u_1^T z$ ← dot product $O(n)$

$\vdots$

$\alpha_n = u_n^T z$

Total n coefficients. So, $O(n^2)$

Theorem: The eigenvectors of a matrix $A \in R^{n \times n}$ having distinct eigenvalues are linearly independent.

For $n \times n$ matrix, we can have $<= n$ eigen vectors depending on the rank.

Theorem: Eigenvectors of a square symmetric matrix are orthogonal. They can be easily converted to orthonormal and use them as basis.

Lec 6.3: Eigenvalue Decomposition:

Let, $u_1, u_2 \dots u_n$: eigen vectors of a matrix A

$\lambda_1, \lambda_2, \dots \lambda_n$: corresponding eigen values.

$$AU = A\underbrace{\begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \dots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}}_{\text{matrix } U} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ Au_1 & Au_2 & \dots & Au_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \lambda_1 u_1 & \lambda_2 u_2 & \dots & \lambda_n u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

$$= \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \dots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} = U\Lambda \quad (\Lambda: \text{lowda})$$

Now, If U^{-1} exists, then,

$AU = U\Lambda$

Post multiply by U^{-1} ↓

$\Rightarrow AUU^{-1} = U\Lambda U^{-1}$

$\Rightarrow A = U\Lambda U^{-1}$ → eigenvalue decomposition of A.

Pre-multiply by U^T ↓

$U^TAU = U^TU\Lambda$

$\Rightarrow \Lambda = U^TAU$

diagonalization of A.

Now, U^{-1} will exist only if $\det(U) \neq 0$ or when the columns of U are linearly independent.

We know that the set of eigenvectors of a matrix are LD.

Special Case: A is symmetric

$\Rightarrow$ eigen vectors are orthogonal.

Assume, we convert U into orthonormal eigen vectors

Then, $U^TU = \begin{bmatrix} \leftarrow u_1 \rightarrow \\ \leftarrow u_2 \rightarrow \\ \vdots \\ \leftarrow u_n \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \dots & u_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} u_1^Tu_1 & u_1^Tu_2 & \dots & u_1^Tu_n \\ u_2^Tu_1 & u_2^Tu_2 & & \\ \vdots & & & \\ u_n^Tu_1 & & & u_n^Tu_n \end{bmatrix}$

$\Rightarrow U^TU = I \quad \Rightarrow \underline{U^T = U^{-1}}$

Thus, inverse of U is very easy to calculate. (no complexity) $O(1)$

Theorem:

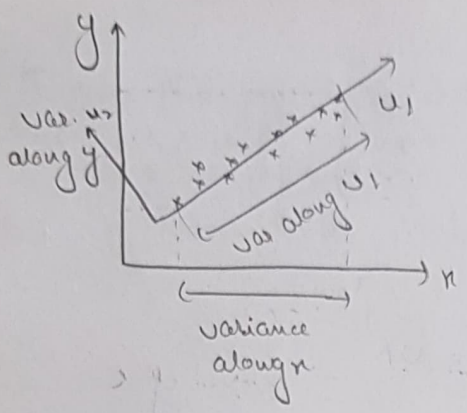
Given $A \in R^{n \times n}$, then Soln. of,

$$\max_x x^TAx \quad \text{such that } \|x\| = 1 \ (x^Tx = 1) \text{ and } x \in R^n.$$

The soln. will be the eigen vector corresponding to the largest eigen value of A. i.e. if λ is the dominant eigen value, then soln. is $x \rightarrow$ corresponding eigen vector.

Similarly, to minimise, use minimum (smallest) eigen value.

Lec 5.4 : Principal Component Analysis (PCA):



Consider the data shown.

Here, basis - x & y ($\hat{i}$ and $\hat{j}$)

Here, we can see that the data needs 2 dimensions x and y for accurate representation due to significant variance along both.

Now, suppose, we choose another set of basis u_1 & u_2. Here, we can see that var. along u_1 is very large compared to u_2. Thus, we can neglect u_2 and we can hence represent the data in 1D.

In case of n-dim data, the goal is to choose the ideal set of basis vectors (u in total). Then, rank them, throw away the ones with lowest variance. Don't throw away the lowest variance dim in the original data.

Even after choosing a new set of basis vectors, if all the dim have high variance, then we can't throw away. (no gain)

* We also need to get rid of highly correlated dimensions. (linear dependence)

Eg:- • Salary and income tax
• length in cm and inches.

We need to keep only one of these dim (columns)

Correlation between dim x & y. $\rightarrow$

$$\rho_{xy} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$(x_i - \bar{x})$ $\rightarrow$ zero centering the mean.

N_r will be +ve when $y_i > \bar{y}$ and $x_i > \bar{x}$

$\Rightarrow$ y_i and x_i are correlated

When $y_i > \bar{y}$ & $x_i < \bar{x}$

$N_r \rightarrow$ -ve (uncorrelated)

Goals: • New set of dim having high var.
• linearly independent dim (uncorrelated)
• orthogonal: they form convenient basis.

Let P be a matrix (n×n) containing $P_1, P_2 \dots P_n$. ~~linearly ind~~ (new set of orthonormal basis) along the columns.

Let's say we have m-data points $x_1, x_2 \dots, x_m \in \mathbb{R}^n$. Let X be a matrix (m×n) having $x_1, x_2 \dots, x_m$ along the rows. Let X be 0-mean and unit variance.

Each x_i can be represented using the new basis P.

$$x_i = \alpha_{i_1} P_1 + \alpha_{i_2} P_2 + \cdots + \alpha_{in} P_n$$

Since P is orthonormal,

$$\alpha_{ij} = x_i^T P_j = \begin{bmatrix} \leftarrow x_i^T \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow \\ P_j \\ \downarrow \end{bmatrix} = (1 \times 1) \text{ scalar val.}$$

Transformed data point

$$\hat{x}_i = \underbrace{\begin{bmatrix} \leftarrow x_i^T \rightarrow \end{bmatrix}}_{\text{single row of } x} \underbrace{\begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ P_1 & P_2 & \cdots & P_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}}_{P} = \underbrace{x_i^T P}_{\text{representation of } x_i \text{ in the new basis sys}} \quad (1 \times n)$$

Transforming whole set of m-points;

projection of data onto new basis

$$\hat{X} = XP \quad (m \times n)$$

Now; since columns of X are zero mean, that of $\bar{X}$ are also zero mean.

Theorem:

$X^T X$ is a symmetric matrix

Proof: $(X^T X)^T = X^T (X^T)^T = X^T X$

Covariance Matrix:

For X whose columns are zero mean, $\Sigma = \frac{1}{} X^T X$ is the covariance matrix. Σ_{ij} is the covariance between columns i and j of X.

let C be the covariance matrix of X.

$$C_{ij} = \frac{1}{} \sum_{k=1} (X_{ki} - \mu_i)(X_{kj} - \mu_j)$$

μ_i & μ_j → mean of i^{th} col. and j^{th} col.

Since, $\mu_i = \mu_j = 0$.

$$C_{ij} = \frac{1}{} \sum_{k=1} X_{ki} X_{kj}$$ (dot product)

C → $(n \times n)$ matrix

$$C_{ij} = \frac{1}{} x_i^T x_j = \frac{1}{} (X^T X)_{ij}$$

* In broad terms Covariance matrix gives the correlation between different dimensions wrt the current data.

Since, we are working in the new basis system,

$$\hat{X} = XP$$

Cov. matrix of transformed data (cov. matrix of X)

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

$$= P^T \Sigma P$$

Now, the diagonal of covariance matrix is the variance of data along each dim.

Ideally, we want variance $\neq 0$ $\quad \frac{1}{m}\hat{X}^T\hat{X} = 0 \quad i \neq j$

covariance $= 0$ $\quad \frac{1}{m}\hat{X}^T\hat{X} \neq 0 \quad i = j$

Thus, ideally we want

$$\frac{1}{m}\hat{X}^T\hat{X} = P^T \underline{\Sigma} P = D \leftarrow \text{diagonal matrix}$$

$$\uparrow \frac{1}{m}X^T X$$

Now, $P^T \Sigma P$ is called diagonalization of Σ. Σ is a sq. matrix $(n \times n)$, we want to find the matrix $P\,(n \times n)$ such that $P^T \Sigma P$ is a diagonal matrix. orthogonal $\uparrow$ matrix

Thus, P must be a ~~vector~~, whose columns are ~~eigenvalues~~ eigen vectors of Σ or $X^T X$. — by Eigen value decomposition.

Thus, we arrive at the soln. that the eigen vectors of $X^T X$ form the ideal basis on which we need to project our data. We ~~also~~ also know that the eigenvectors of sq. symmetric matrix is orthogonal. Thus P is orthogonal.

Thus, we have projected the data along a new set of orthonormal basis where the dimensions are non-redundant (low covariance) and not noisy (high variance).

Now, these new set of basis P can represent the data X perfectly if we use the whole n basis vectors. If we remove some of them, we will get an approximation.

we want top-k dims and remove noisy and redundant dim.

approximation $\hat{X}_i = \sum_{j=1}^{k} a_{ik} P_k$ where $k < n$.

Our goal is to choose the basis such that e is minimized.

$$e = \sum_{i=1}^{m} (x_i - \hat{x}_i)^2 \quad \text{MSE}$$

$\hat{x}_i$: exact representation

$$= \sum_{i=1}^{m} \left[\sum_{j=1}^{n} \alpha_{ij} p_j - \sum_{j=1}^{k} x_{ij} p_j \right]^2$$

x_{ij}: approx repr.

$$= \sum_{i=1}^{m} \left[\sum_{j=k+1}^{n} x_{ij} p_j \right]^2$$

errors due to removing these n-k basis (for each data point)

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

$$= \sum_{i=1}^{m} \left[\alpha_{i,k+1} p_{k+1} + \alpha_{i,k+2} p_{k+2} + \cdots \right]^T \left[\alpha_{i,k+1} p_{k+1} + \alpha_{i,k+2} p_{k+2} + \cdots \right]$$

$$= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \alpha_{ij}^2$$

$$= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \alpha_{ij} \alpha_{il} p_j^T p_l$$

$$= \sum_{i=1}^{m} \sum_{j=k+1}^{n} \alpha_{ij}^2 \, p_j^T p_j + \sum_{i=1}^{m} \sum_{j=k+1}^{n} \sum_{\substack{l=k+1 \\ l \neq j}} \alpha_{ij} \alpha_{il} p_j^T p_l$$

$p_j^T p_j = 1$

$p_j^T p_l = 0$ (orthonormal basis)

$$e = \sum_{i=1}^{m} \sum_{j=k+1}^{n} \alpha_{ij}^2 \quad \text{(dot product)}$$

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

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

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

$\sum_{i=1}^{m} x_i x_i^T$: covariance matrix

$$= \sum_{j=k+1}^{n} p_j^T C p_j$$

$$x_i x_i^T = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{bmatrix} \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{in} \end{bmatrix}$$

$$x_i x_i^T = \begin{bmatrix} x_{i1}^2 & x_{i1}x_{i2} & \cdots & x_{i1}x_{in} \\ \vdots & & & \\ x_{in}x_{i1} & \cdots & & x_{in}^2 \end{bmatrix}$$

Now, first term of $\sum_{i=1}^{m} x_i x_i^T$ is $x_{11}^2 + x_{21}^2 + \cdots + x_{m1}^2$ — variance

Non-diagonal elements will be covariance.

Thus $\sum_{i=1}^{m} x_i x_i^T = m X^T X$

Our goal is

$$\min_{P_{k+1}, P_{k+2}, \ldots, P_n} \sum_{j=k+1}^{n} P_j^T \, \mathrm{wc} \, P_j \quad \text{such that } \overbrace{P_j^T P_j = 1}^{\text{orthonormal basis}} \quad \forall j = k+1, \ldots n.$$

The soln. for this is P_j = eigen vectors corresponding to the n-k smallest eigen values.

Eg:-

Consider a single point in the 2D dataset (3.3, 3)

If the new basis is $u_1 = [1, 1]$, $u_2 = [-1, 1]$

Normalize, $u_1 = \left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$ $u_2 = \left[\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$

$$\alpha_1 = x^T u_1 = [3.3, 3] \begin{bmatrix} 1 \\ 1 \end{bmatrix} = [6.3]/\sqrt{2}$$

$$\alpha_2 = x^T u_2 = 0.3/\sqrt{2}$$

$$x = \alpha_1 u_1 + \alpha_2 u_2$$

(eigenvalues ↑) → throw away the dim with smallest eigenvalue.

$$\hat{x} = \alpha_1 u_1 = [3.15, 3.15]$$

So far we have achieved the goal of low covariance. How do we ensure we retain those dimensions along which the variance is high?

Transformed data along i^{th} dim:

$$\hat{x}_i = X P_i$$

↑ i^{th} basis (eigen vector)

variance along this dim,

$$\frac{\hat{x}_i^T \hat{x}_i}{m} = \frac{1}{m} P_i^T X^T X P_i$$

Now, P_i is an eigen vector of $X^T X$. So, it will only be scaled by $X^T X$ by λ_i

$$= \frac{1}{m} \lambda_i \underbrace{P_i^T P_i}_{1} = \frac{\lambda_i}{m}$$

⇒ variance along a dim ∝ eigen value of that dim.

So, for high variance, just retain the dim with large eigen values

PCA Practical Example:

Consider we have millions of human face images each $100 \times 100 = 10k$ dim. Without compression, we need to store $10K \times$ number of images values.

using PCA, we can just use around 100 dim to represent each face.

we have $X \in R^{m \times 10K}$ (m = no. of images)

Now, we retain those eigen vectors corresponding to top 100 eigen values. Each eigenvector has a length 10k (10k dim). We can reshape each eigenvector as 100×100 image. Since, we have taken top 100 eigen vectors, each will be connected in some way to the faces.

It turns out, these eigen vectors will look somewhat like a face. These are called eigen faces. We can scale and add them to represent any face in our database. (linear combination)

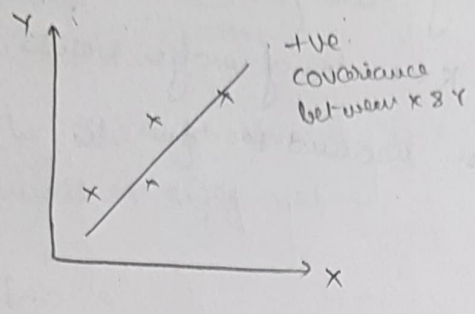
To find the α values, take dot product of face vector with the eigen vector.

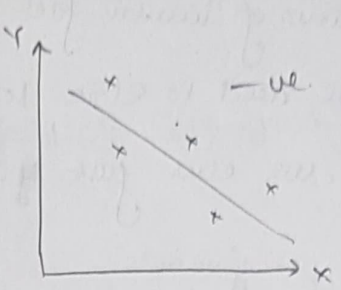
using this technique, we only need to store $100 \times 10k$ values plus 100 α values for each face.

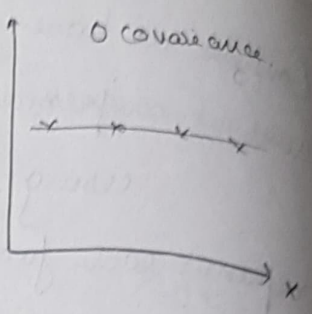
* A very important interpretation of PCA is that it only keeps relevant data and throws away the rest. So, two diff. faces of the same person under different lighting condition will be very far in the original space due to this irrelevant info. It will come closer in the new space.

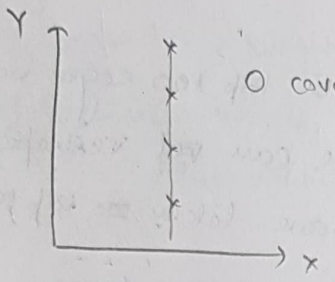
Covariance:

It tells us the relationship between two dimensions of a data

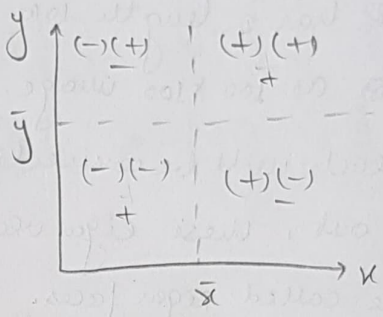








$$Cov = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$



Data points lying in '+' contribute to +ve covariance and vice versa

* We only care about the sign of covariance and not the magnitude.

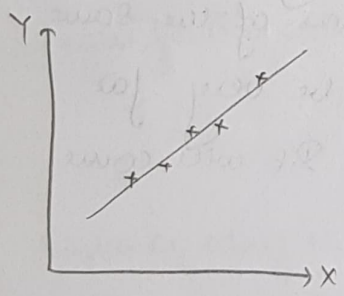
Covariance between the same dim is the same as variance

$$Var = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

* Correlation values are scale dependent and hence should not be used. (magnitude)

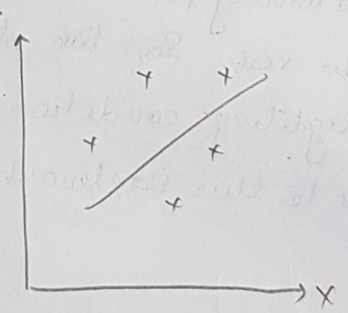
Correlation:

It tells the strength of the relationship between two dimensions of data



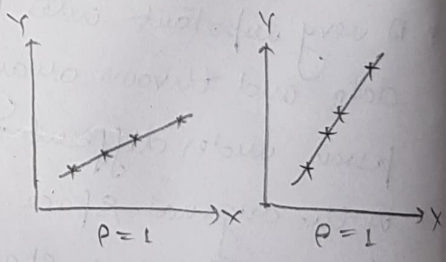
Data lies almost on a straight line
Strong correlation.
($\rho \approx 1$)

Given a value for X or Y we can very accurately predict the value for the other.



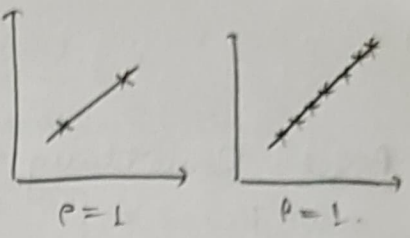
Data lies far from the trend line
weak correlation
$0 < \rho << 1$

Given a value of X we can't predict Y with accuracy.



If a straight line passes through all data points $\rho = 1$. Does not depend on slope or scale on the axes.

Correlation does not depend upon no. of data points. Thus, we should not have confidence in correlation when we have a small data set.



Correlation magnitude is scale independent.

$$\text{Correlation: } \rho = \frac{\text{Covariance}(X,Y)}{\sqrt{\text{variance}(X)}\sqrt{\text{variance}(Y)}}$$

For -ve slope, ρ is -ve.