Geometric interpretation of matrix multiplications

Define matrix as function $f, g : \mathbb{R}^2 \to \mathbb{R}^2$ w/ input $[x\ y]^T$

$$f : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad g : \begin{bmatrix} p & q \\ r & s \end{bmatrix}, \quad \text{then} \quad f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} ax+by \\ cx+dy \end{bmatrix}$$

$$g\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} px+qy \\ rx+sy \end{bmatrix}$$

$$f \cdot g\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = f\left(g\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} (ap+br)x + (aq+bs)y \\ (cp+dr)x + (cq+ds)y \end{bmatrix}$$

① $\therefore$ Composite mapping $f \cdot g$ is defined as matrix product

Each element of final vector/matrix is inner product b/w row & column vector

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} a \\ c \end{bmatrix} = a\begin{bmatrix} 1 \\ 3 \end{bmatrix} + c\begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

$\Rightarrow$ Multiplication b/w matrix & vector is ② linear combination of two column vectors

$\Downarrow$

vector space using column vectors = column space

$\Downarrow$

③ Linear transformation

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

Basic vector $\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\} \to \left\{\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ 1 \end{bmatrix}\right\}$

i.e. $\begin{bmatrix} -1 \\ 2 \end{bmatrix}$ (at original coordinate system) is $1\begin{bmatrix} 2 \\ 1 \end{bmatrix} + 1\begin{bmatrix} -3 \\ 1 \end{bmatrix}$

Row vector and inner product

$[2\ \ 1]\begin{bmatrix}3\\4\end{bmatrix} = 2$ $\Rightarrow$ inner product b/w row & column vector

$\swarrow$ function (operator) of column vector $\begin{bmatrix}3\\-4\end{bmatrix}$

$\searrow$ operand for row vector $[2\ \ 1]$

$\therefore [2\ \ 1]\left(\begin{bmatrix}3\\4\end{bmatrix}\right) = 2$

$f : V \to \mathbb{R}$

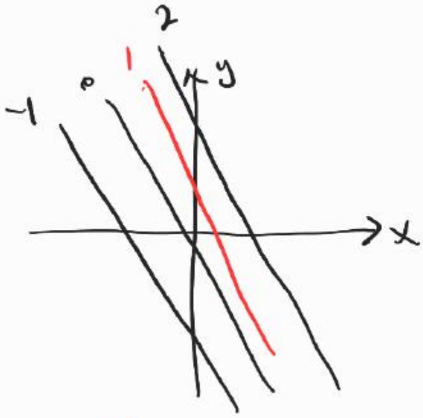
- Visualizing function of row vector

e.g. function $y = x^2$

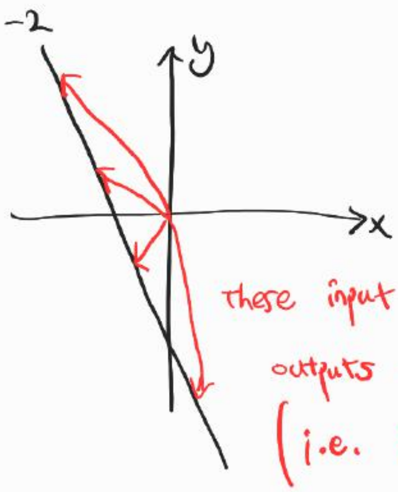
function $[2\ \ 1]\left(\begin{bmatrix}x\\y\end{bmatrix}\right) = 2x+y$

$\rightarrow$ Each output $2x+y$ represents a single line ($\simeq$ contour plot)

e.g. $2x+y=1 \rightarrow y=-2x+1$



$\leftarrow$

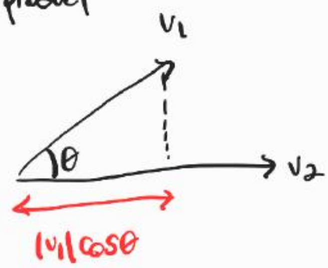


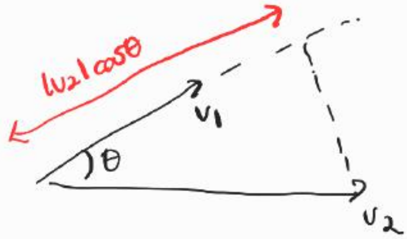
these input vectors $\begin{bmatrix}x\\y\end{bmatrix}$ outputs -2

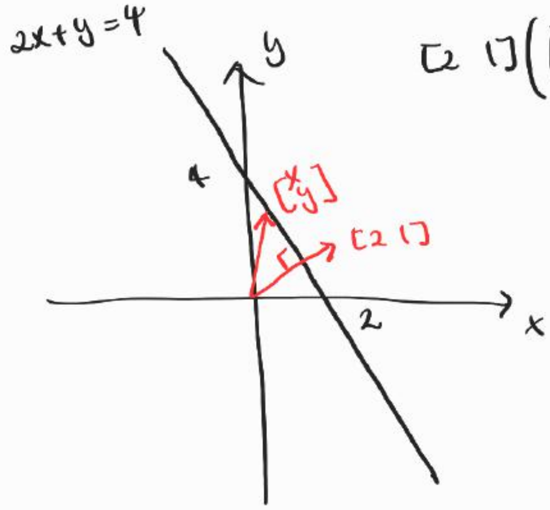
(i.e. $[2\ \ 1]\left(\begin{bmatrix}x\\y\end{bmatrix}\right) = -2$)

- Geometric interpretation of dot product

$v_1 \cdot v_2 = |v_1||v_2|\cos\theta$







$$[2\ 1]\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = 2x+y$$

$$\frac{1}{2}\cdot 2\cdot 4 = \frac{1}{2}\left(\sqrt{2^2+4^2}\right) d \qquad \therefore d = \frac{4}{\sqrt{5}}$$

d = linear projection of any $\begin{bmatrix} x \\ y \end{bmatrix}$ to row vector $[2\ 1]$

$$\therefore v_1 \cdot v_2 = [2\ 1]\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = |v_1| \cdot d = 4$$

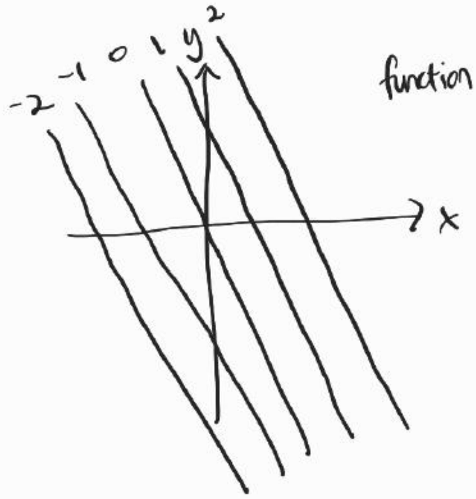
- Linearity of row vector

1) $f(v+w) = f(v) + f(w)$ $\qquad [2\ 1]\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ 4 \end{bmatrix}\right) = [2\ 1]\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) + [2\ 1]\left(\begin{bmatrix} 3 \\ 4 \end{bmatrix}\right)$

2) $f(nw) = nf(w)$ $\qquad [2\ 1]\left(2\begin{bmatrix} 3 \\ 4 \end{bmatrix}\right) = 2[2\ 1]\left(\begin{bmatrix} 3 \\ 4 \end{bmatrix}\right)$

$\therefore$ f is linear operator

- Geometric interpretation of row vector linearity



function f $[2\ 1] : V \to \mathbb{R}$

$\Leftrightarrow$ output of input vector $\begin{bmatrix} x \\ y \end{bmatrix}$

= # of intersecting lines

$\Leftrightarrow$ ↑ length of row vector

= contour lines are closer

Addition of row vectors = creating new contours

(perpendicular to the row vector)

* Row space and column space are dual space → problem hard to solve in row space can be easily substituted into column space

Dual space $V^* = \{f : V \to \mathbb{R} \mid f(ca+b)\ \forall\ a,b \in V\}$

Matrix as linear transformation

As transformation T follows linearity conditions

$$T(a+b) = T(a) + T(b)$$
$$T(ca) = cT(a)$$

$$T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = x\,T\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) + y\,T\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$$

$\Leftrightarrow$ basis vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ transforms into $T\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right), T\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)$

* Geometric features of linearity : 1) grid lines are linear

2) uniform spacing b/w grid lines

- Types of linear transformation

1) shearing $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

2) rotation $\begin{bmatrix} \cos\frac{\pi}{2} & -\sin\frac{\pi}{2} \\ \sin\frac{\pi}{2} & \cos\frac{\pi}{2} \end{bmatrix}$

3) permutation $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

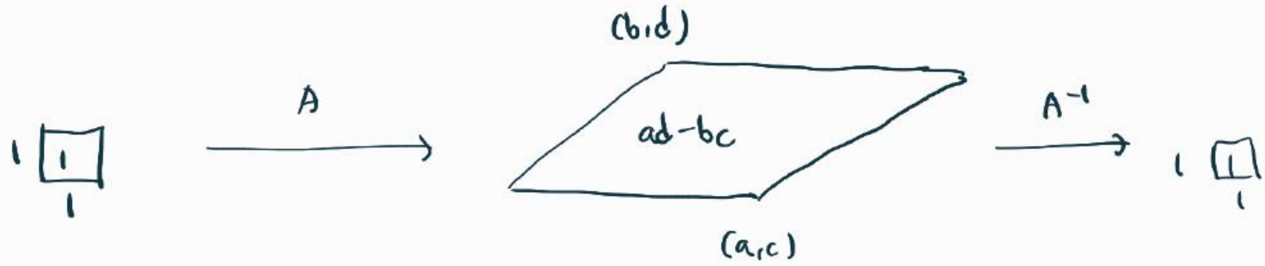
4) projection $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

Geometric interpretation of determinent

$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ where $\det(A) = ad - bc$

det of 2x2 matrix = area of parallelogram of <u>two basis vectors after transformation</u>

$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} a \\ c \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix}$



Eigenvalue and Eigenvector

Normally, matrix A is linear transformation function

$$x \longrightarrow A \longrightarrow \underline{Ax}$$

another vector w/ different length and direction

However, some matrix A doesn't change direction of Ax from x

i.e., $Ax = \lambda x$

$x \rightarrow A \rightarrow Ax = \lambda x$

λ = eigenvalue
x = eigenvector

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

either $A - \lambda I$ or x equals 0

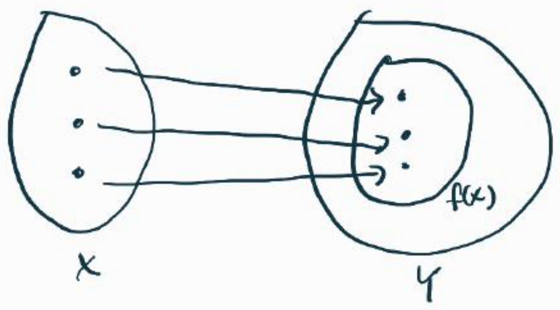
→ trivial solution $x = 0$

→ non-trivial solution $A - \lambda I = 0$ $\Leftrightarrow$ $A - \lambda I$ is non-invertible $\Leftrightarrow$ $\det(A - \lambda I) = 0$

Relationships b/w four fundamental subspaces

Linear transformation A = function ?

Definition of function



$f: X \to Y$

X : domain
Y : codomain
$f(X)$: range

f : subset of Cartesian product $X \times Y$

$\forall x \in X$, if there exists unique $y \in Y$, $y = f(x)$

$\Leftrightarrow$ f is mapping b/w X and Y

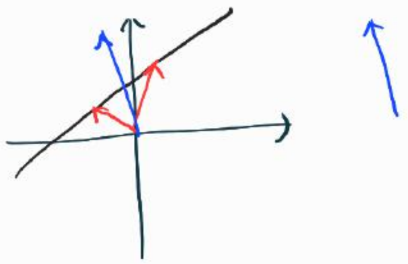
- subspace

* vector space : set of vectors defined by + (addition) and · (scalar multiplication)

Subspace of vector space $\simeq$ subset of set

↳ still requires to follow definition of vector space

e.g.



not on the line = addition rule × holds

- Row / column space

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix} \rightarrow \text{row space} = \text{span}\left\{ \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$$

$$\text{column space} = \text{span}\left\{ \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right\}$$

] They are subspaces!

- null space

set of $\vec{x}$ that satisfies $A\vec{x} = 0$

* Row space $\perp$ null space

- Fundamental theorem of linear algebra

If matrix is a function, how do we define the relationship b/w sets, which is the fundamental meaning of the function?

$$A \in \mathbb{R}^{m\times n}, \quad f: \underbrace{\mathbb{R}^n}_{\substack{\text{domain} \\ \text{n-dim vector space}}} \to \underbrace{\mathbb{R}^m}_{\substack{\text{Codomain} \\ \text{m-dim vector space}}}$$

① Domain $\mathbb{R}^n$ = row space + null space

As row space $\perp$ null space, represent $x \in \mathbb{R}^n$ as linear combination of two vectors inside spaces.

② Range = column space

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = x\begin{bmatrix} a \\ c \end{bmatrix} + y\begin{bmatrix} b \\ d \end{bmatrix} = \text{linear combination of column space}$$

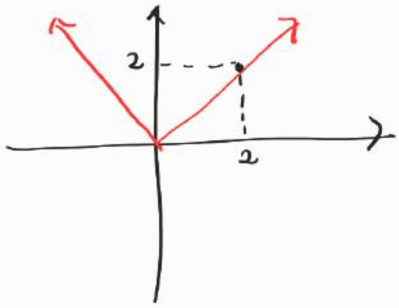
③ Co-domain = column space + left null space

column space $\perp$ left null space

change of basis

Standard basis $(\hat{i}, \hat{j}) = \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) = \mathcal{E}$

e.g. new basis $B = \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right)$



$$\begin{bmatrix} 2 \\ 2 \end{bmatrix}_{\mathcal{E}} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}_B$$

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

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

$\therefore$ For new basis $C = \left\{ \begin{bmatrix} 1 \\ c_1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ c_2 \\ 1 \end{bmatrix} \right\}$ and vector $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}_C$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_{\mathcal{E}} = \begin{bmatrix} 1 & 1 \\ c_1 & c_2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}_C$$

- Coordinate transition

Basis $B = \{v_1, v_2\}$, $C = \{w_1, w_2\}$

$$\begin{cases} w_1 = av_1 + bv_2 \\ w_2 = cv_1 + dv_2 \end{cases}$$

For vector v,

$$\begin{aligned} v &= l_1 w_1 + l_2 w_2 \\ &= l_1(av_1 + bv_2) + l_2(cv_1 + dv_2) \\ &= (al_1 + cl_2)v_1 + (bl_1 + dl_2)v_2 \\ &= k_1 v_1 + k_2 v_2 \end{aligned}$$

$$\therefore \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} = \begin{bmatrix} al_1 + cl_2 \\ bl_1 + dl_2 \end{bmatrix} = \underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{\text{transition matrix}}\begin{bmatrix} l_1 \\ l_2 \end{bmatrix}$$

Elementary square matrices

· Solving simultaneous equations

$$\begin{cases} 2x + 3y = 1 \\ 4x + 7y = 3 \end{cases}$$

operations ① $r_1 \to kr_1$ Row multiplication

② $r_1 \to r_1 + r_2$ Row addition

③ $r_1 \leftrightarrow r_2$ Row switching

↓

we can use matrix for these computations ⇒ easier for computers to computer

$$\begin{bmatrix} 2 & 3 \\ 4 & 7 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$\left[\begin{array}{cc|c} 2 & 3 & 1 \\ 4 & 7 & 3 \end{array}\right]$$ ← this augmented matrix is operand

① row multiplication $E = \begin{bmatrix} k & 0 \\ 0 & 1 \end{bmatrix}$ $r_1 \to kr_1$ $E^{-1} = \begin{bmatrix} 1/k & 0 \\ 0 & 1 \end{bmatrix}$

② row switching $E = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $r_1 \leftrightarrow r_2$ $E^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

③ row addition $E = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ $r_1 \to r_1 + r_2$ $E^{-1} = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix}$

↓

These operations are called elementary matrices

LU decomposition * only applicable to square matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & -1 \\ 2 & 3 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -3 \\ 0 & 0 & 4 \end{bmatrix}$$

E_1: $r_3 \to r_3 - r_2$ E_2: $r_3 \to r_3 - 2r_1$ E_3: $r_2 \to r_2 - 2r_1$ A U: upper-triangular matrix

$\downarrow$

$$A = E_1^{-1} E_2^{-1} E_3^{-1} U$$
$$= LU$$

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -3 \\ 0 & 0 & -4 \end{bmatrix} = LU$$

* when row switching operation is required, we use PLU decomposition

eg $A = \begin{bmatrix} 0 & 0 & 3 \\ 1 & 1 & 1 \\ 2 & 3 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} LU$

$P: r_1 \leftrightarrow r_3$

- Applications

① solving $Ax = b$

$$Ax = b \xrightarrow{A = LU} L(Ux) = b \xrightarrow{Ux = c} Lc = b \xrightarrow[\text{forward substitution}]{\text{solve } c}$$

$$Ux = c \xrightarrow[\text{backward substitution}]{\text{solve } x} x$$

② Calculate $\det(A)$

$$\det(A) = \det(LU) = \det(L)\det(U) = \prod_{i=1}^{n} l_{i,i} \prod_{j=1}^{n} u_{j,j}$$

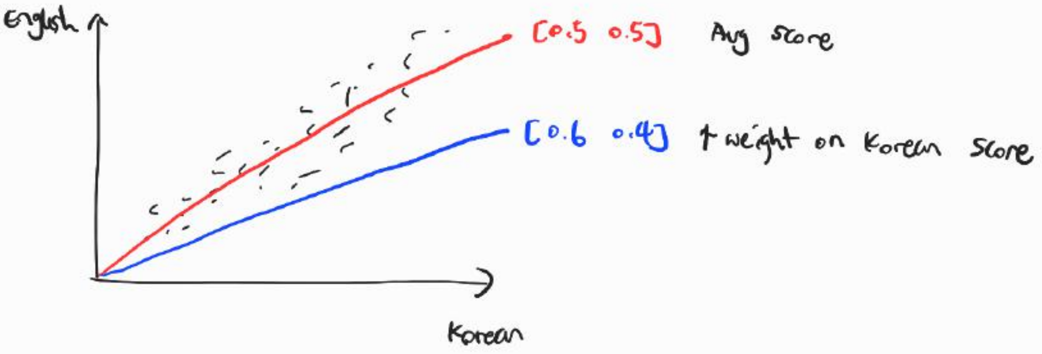
(multiplication of all diagonal elements)

Principal Component Analysis (PCA)

⇒ If you need to reduce dimension of data by projection to a vector, what is the best way to minimize data loss?

e.g. calculating overall test score

English	Korean
80	60
70	65
75	50
⋮	⋮



⇒ Main problem : what is the best vector that gives best results?

Dot product b/w data and vector

e.g. $\begin{bmatrix} 80 & 60 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$

- Covariance matrix

Matrix = linear transformation & function mapping one vector space to another

$$A_{cov} = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$$

→ spread along x-axis

→ spread along y-axis

spread along both x, y-axis

Eigenvector : Principle axis that matrix acts on
Eigenvalue : Stretched magnitude along eigenvector

] ⇒ sort eigenvalue ∴ ↑ eigenvalue = ↑ significance of eigenvector

∴ PCA = projection of data of principle axis = find eigenvector of covariance matrix w/ ↑ eigenvalue

- Calculating covariance matrix

$$X = \begin{pmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_d \\ | & | & & | \end{pmatrix} \in \mathbb{R}^{n \times d}$$ i.e. n samples, d features

$$X^T X = \begin{pmatrix} x_1 \cdot x_1 & \cdots & x_1 \cdot x_d \\ \vdots & \ddots & \\ x_d \cdot x_1 & & x_d \cdot x_d \end{pmatrix}$$ i.e. $(X^T X)_{ij}$ = similarity b/w feature x_i and x_j

$$\Sigma = \frac{X^T X}{n}$$ $\because$ dot product ↑ value as sample size (n) ↑

- Find # of dimensions to reduce

For full rank covariance matrix $\Sigma_{d \times d}$, let eigenvalue $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$,

select m which satisfies $$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{i=1}^{d} \lambda_i} \geq 0.9$$

Eigen-value decomposition

⇒ Decomposing original linear transformation A into V (rotation), Λ (stretch), V^{-1} (rotation)

- Derivation

  Assume there are n independent eigenvectors $(v_1 \cdots v_n)$ and eigenvalues $(\lambda_1 \cdots \lambda_n)$ of matrix $A \in \mathbb{R}^{n \times n}$.

  $Av_i = \lambda v_i$ for $i = 1, 2, \cdots, n$

  $$AV = \begin{bmatrix} | & & | \\ \lambda_1 v_1 & \cdots & \lambda_n v_n \\ | & & | \end{bmatrix} = V \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = V\Lambda$$

  $\therefore A = V\Lambda V^{-1}$

- Geometric interpretation

  e.g. $A = V\Lambda V^{-1}$

  $\therefore \Lambda$ acts as stretching

  $$\begin{bmatrix} 1.2 & -0.5 \\ -1.5 & 1.9 \end{bmatrix} = \begin{bmatrix} \underbrace{\boxed{\begin{matrix} 0.6089 \\ 0.9933 \end{matrix}}}_{v_1} & \begin{matrix} -0.3983 \\ 0.9172 \end{matrix} \end{bmatrix} \begin{bmatrix} \overset{\lambda_1}{\boxed{0.5486}} & 0 \\ 0 & 2.3514 \end{bmatrix} \begin{bmatrix} 1.0489 & 0.4555 \\ -0.9092 & 0.6963 \end{bmatrix}$$

  $\therefore V^{-1}$ acts as inverse rotation

  normalized vector $|v_1| = 1$

  $\therefore V$ acts like rotation (as basis vector length $= 1$)

- EVD of symmetric matrix

  if $A = A^T$, $A = Q\Lambda Q^T$

Meaning of complex eigenvalue and eigenvector

- Eigenvector of rotation matrix

$A(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \rightarrow Ax = \lambda x$: is there vector x that retains its direction after rotation A?

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

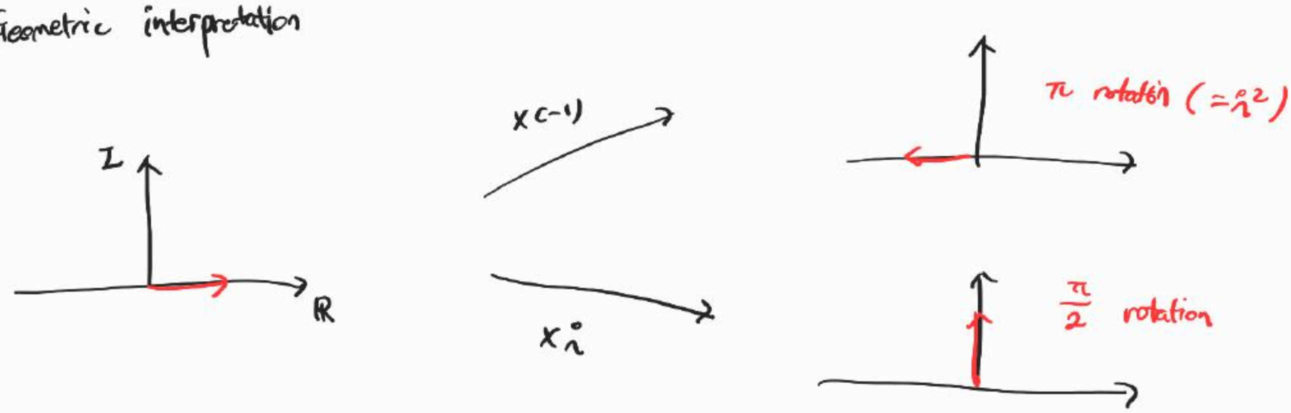
$\det(A - \lambda I) = (\cos\theta - \lambda)^2 + \sin^2\theta = 0$

$\lambda^2 - 2\lambda\cos\theta + 1 = 0$

$\therefore \lambda = \cos\theta \pm i\sin\theta$

if $\lambda = \cos\theta + i\sin\theta$, $x = \begin{bmatrix} i \\ 1 \end{bmatrix}$, if $\lambda = \cos\theta - i\sin\theta$, $x = \begin{bmatrix} -i \\ 1 \end{bmatrix}$

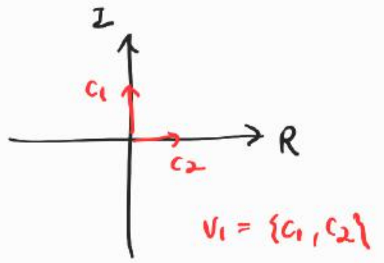
- Geometric interpretation



$\therefore$ Imaginary # multiplications = rotation of vector

- Visualizing complex eigenvector

$x_1 = \begin{bmatrix} i \\ 1 \end{bmatrix}$ $x_2 = \begin{bmatrix} -i \\ 1 \end{bmatrix}$



$\therefore$ two vectors $\equiv$ a complex vector

- Relationships b/w rotation matrix & eigenvector

$\lambda_1 = \cos\theta + i\sin\theta = \exp(i\theta)$ = anti-clockwise θ rad rotation

$\therefore$ scaling complex eigenvector w/ eigenvalue (λx) $\equiv$ rotation of complex eigenvector c_1, c_2 (Ax)

Linear regression

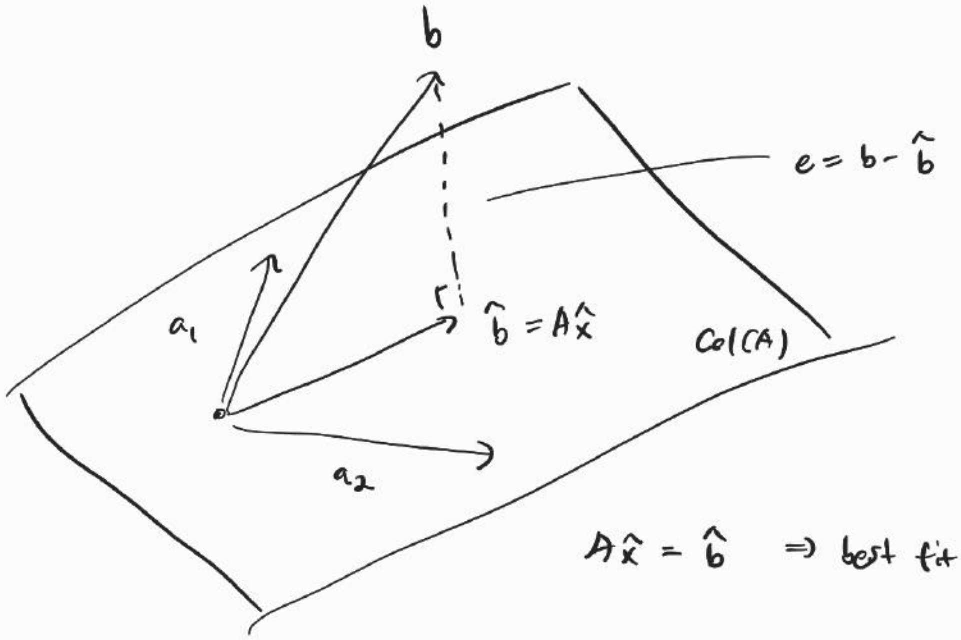
⇒ find the best linear trend line that explains the data (# of data > feature dimension)

$$Ax = b$$

$$\begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} | \\ b \\ | \end{bmatrix} \Rightarrow x_1 \begin{bmatrix} | \\ a_1 \\ | \end{bmatrix} + x_2 \begin{bmatrix} | \\ a_2 \\ | \end{bmatrix} = \begin{bmatrix} | \\ b \\ | \end{bmatrix}$$

How to combine a_1 and a_2 to output b

if $b \notin \text{span}\{a_1, a_2\}$ ($b \notin \text{Col}(A)$), there is no exact solution, thus find "best fit"



$A\hat{x} = \hat{b}$ ⇒ best fit (min error)

$$A \cdot e = \begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix} \cdot e = 0$$

∵ e is perpendicular to any vector in Col(A)

≡ e is in left nullspace

$$A^T e = A^T(b - A\hat{x}) = A^T b - A^T A\hat{x} = 0$$

$$A^T A\hat{x} = A^T b$$

$$\therefore \hat{x} = (A^T A)^{-1} A^T b$$

Geometric meaning of pseudo-inverse

- Definition

For $A \in \mathbb{R}^{m \times n}$, if $m > n$ and column vectors are linearly independent

$$A^+ = (A^TA)^{-1}A^T \quad \text{where } A^TA \text{ is invertible} \quad \rightarrow A^+A = I \text{ (left inverse)}$$

if $m < n$ and —

$$A^+ = A^T(AA^T)^{-1} \quad \text{——} \quad \rightarrow AA^+ = I \text{ (right inverse)}$$

$\Rightarrow$ Matrix of any size can function as inverse matrix

- Mathematical meaning

$$Ax = b$$

$$A^+Ax = A^+b \qquad (A^+ = (A^TA)^{-1}A^T)$$

$$\begin{bmatrix} -1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$$

$\downarrow$ using A^+

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$\downarrow$ BUT

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

- Linear regression & pseudo-inverse

$$Ax = b \;\Rightarrow\; A^+Ax = A^+b \;\Rightarrow\; \hat{x} = (A^TA)^{-1}A^Tb$$

$\hat{x}$ is not an exact solution, but "best-fit" solution

= projection to $Col(A)$

- SVD & pseudo-inverse

$A = U\Sigma V^T$ (U and V are orthogonal matrix i.e. $UU^T = UU^T = I$)

(Σ is diagonal matrix i.e. $\Sigma^T = \Sigma$)

$$A^T = V\Sigma^TU^T = V\Sigma U^T$$

$$\therefore A^TA = V\Sigma U^TU\Sigma V^T = V\Sigma^2V^T$$

$$(A^TA)^{-1} = V(\Sigma^2)^{-1}V^T$$

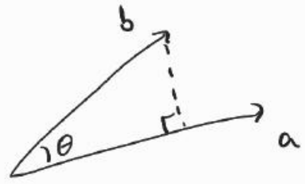
$$A^+ = (A^TA)^{-1}A^T = V(\Sigma^2)^{-1}V^TV\Sigma U^T = V\Sigma^{-1}U^T$$

$$\Sigma^{-1} = \begin{pmatrix} \lambda_1^+ & & \\ & \ddots & \\ & & \lambda^+_{min(n,m)} \end{pmatrix}$$

where $\lambda^+ \begin{cases} \lambda^{-1} & \text{if } \lambda \neq 0 \\ 0 & \text{if } \lambda = 0 \end{cases}$

QR decomposition

- Vector projection



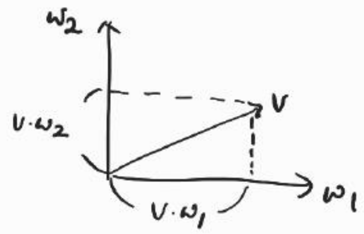
$$\left.\begin{array}{l} comp_a b = |b|\cos\theta \\ a\cdot b = |a||b|\cos\theta \end{array}\right] \quad comp_a b = \frac{a\cdot b}{|a|} \quad (\text{scalar})$$

$$proj_a b = comp_a b \cdot \frac{a}{|a|} = \frac{a\cdot b}{|a|}\cdot\frac{a}{|a|} = \frac{a\cdot b}{a\cdot a}\, a \quad (\text{Vector : multiplied w/ a unit vector})$$

- Grad-Schmidt process

$\Rightarrow$ Convert linearly independent vectors to orthogonal basis

if $\{w_1, w_2\}$ are orthogonal basis, we can represent any $v = (v\cdot w_1)w_1 + (v\cdot w_2)w_2$



Given independent vectors $\{a_1, \cdots, a_k\}$

$u_1 = a_1$

$u_2 = a_2 - \underbrace{proj_{u_1}(a_2)}_{u_1 \text{ component of } a_2}$

$\vdots$

$$u_k = a_k - \sum_{j=1}^{k-1} proj_{u_j}(a_k)$$

Then $q_i = \frac{u_i}{|u_i|}$ $\rightarrow$ unit-vector orthogonal basis $\{q_1, \cdots, q_k\}$

- QR decomposition

$$A = QR = \begin{bmatrix} | & & | \\ q_1 & \cdots & q_n \\ | & & | \end{bmatrix} \begin{bmatrix} a_1\cdot q_1 & a_2\cdot q_1 & \cdots & a_n\cdot q_1 \\ \vdots & & \ddots & \\ & & & \\ a_1\cdot q_n & & \cdots & a_n\cdot q_n \end{bmatrix}$$

for $a_i\cdot q_j$ $(i<j)$, $a_i\cdot q_j = 0$ $\because$ q_j already removed all $i<j$ components during Grad-Schmidt Process

$$\therefore A = \begin{bmatrix} | & | & & | \\ q_1 & q_2 & \cdots & q_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} a_1\cdot q_1 & a_2\cdot q_1 & \cdots & a_n\cdot q_1 \\ & a_2\cdot q_2 & & \vdots \\ 0 & & \ddots & \vdots \\ & & & a_n\cdot q_n \end{bmatrix}$$

SVD decomposition

⇒ For set of orthogonal vectors, what is orthogonal set after linear transformation?

$$A = U\Sigma V^T \quad (A \in \mathbb{R}^{m\times n}, \underbrace{U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}}_{\text{orthogonal}}, \underbrace{\Sigma \in \mathbb{R}^{m\times n}}_{\text{diagonal}})$$

$\therefore U^{-1} = U^T, V^{-1} = V^T$

e.g. 2D case ($A \in \mathbb{R}^{2\times 2}$)

$V = \begin{pmatrix} | & | \\ x & y \\ | & | \end{pmatrix}$ set of orthogonal vectors before transformation

$U = \begin{pmatrix} | & | \\ u_1 & u_2 \\ | & | \end{pmatrix}$ normalized set of orthogonal vectors after transformation

$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$ scaling factor ($\sigma_i|x| = |Ax|$)

$\therefore AV = U\Sigma$ After linear transformation A of V's orthogonal column vectors, is there set of column vectors (U) w/ scaling factor σ?

$A = U\Sigma V^{-1} = U\Sigma V^T$

- A doesn't requires to be a square matrix

  e.g. $A \in \mathbb{R}^{2\times 3}$ ($\because \mathbb{R}^3 \to \mathbb{R}^2$)

  Collapsing 3D to 2D space by one scaling factor $= 0$

$A = U\Sigma V^T = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_m u_m v_m^T$

$\therefore$ SVD decomposes A matrix into different matrix (its weight (size determined by σ_i)

⇒ Able to select few decomposed matrix w/ top this ↓ size but minimize information loss

Non-negative Matrix Factorization

⇒ Decompose a non-negative matrix X into two non-negative matrix H, W

$$\begin{array}{cc} & \text{Data dimension} \\ \text{Sample size} & \begin{bmatrix} \quad X \quad \end{bmatrix} \end{array} = \begin{array}{cc} & \text{feature dim} \\ \text{Sample size} & \begin{bmatrix} \quad W \quad \end{bmatrix} \end{array} \begin{array}{c} \text{Data dim} \\ \begin{bmatrix} \quad H \quad \end{bmatrix} \end{array}$$

Advantage : Can preserve non-negative value feature (e.g. pixel's intensity)

↳ Not assured for other matrix factorization methods e.g. SVD

Can preserve data distribution better ∵ feature × needs to be orthogonal

- How to find W, H

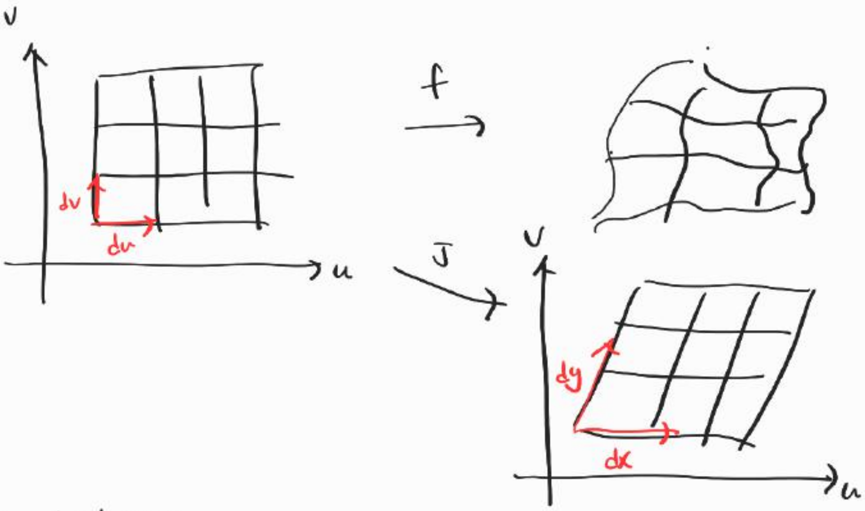
⇒ Iterative update
$$\begin{cases} H := H \circ \dfrac{W^T X}{W^T W H} \\ W := W \circ \dfrac{X H^T}{W H H^T} \end{cases}$$
($\circ$/− is element-wise multiplication / division)

Geometric meaning of Jacobian matrix

$f: \mathbb{R}^n \to \mathbb{R}^m$

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \\ \frac{\partial f_m}{\partial x_1} & & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

$\Rightarrow$ approximates nonlinear transformation in a local region into linear transformation



for local area,

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = J \begin{bmatrix} du \\ dv \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} du \\ dv \end{bmatrix}$$

$dx \cdot dy = |J| \, du \cdot dv$

$$\therefore J = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$

Geometric meaning of Hessian matrix

$f: \mathbb{R}^n \to \mathbb{R}^n$

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \ddots & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & & & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

so H is symmetric matrix

$f(x) = ax^2 + bx + c$: if $f''(x) > 0$ ∪ shape
$f''(x) < 0$ ∩ shape

$\uparrow |f''(x)| \rightarrow \uparrow$ concave/convex

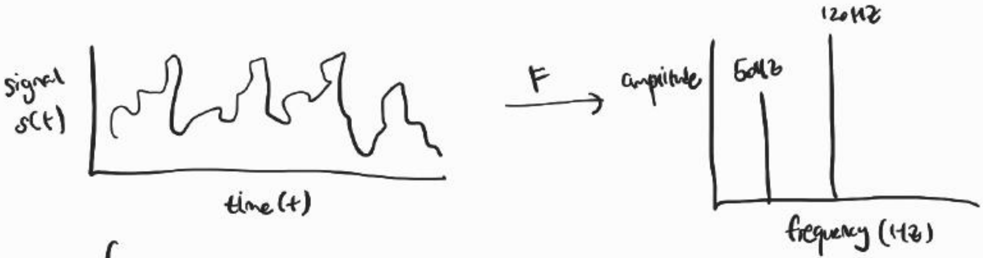
Similarly, H transforms bowl-shaped function more concave/complex

H → find eigenvalue/eigenvector → ↑ eigenvalue ≡ ↑ concave/convex
all eigenvalues > 0 = ∪ shape
mixed eigenvalues' sign ≡ saddle shape function

Linear algebra and Fourier transform

- Fourier transform

Decomposing signal mixed w/ different frequency & amplitude



we can interpret signal as vector (order of numbers)

$x[n] = [x[0], x[1], \cdots, x[N-1]]$ signal discretized every 1 sec

frequency components can also be vector

$X[k] = [X[0], \cdots, X[N-1]]$ discretized every 1Hz

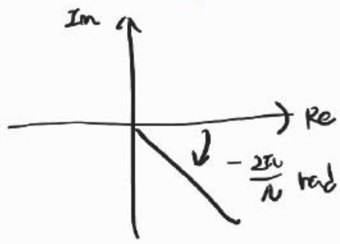
Direct Fourier Transform

$$X[k] = \sum_{n=0}^{N-1} x[n] \exp\left(-j \frac{2\pi k}{N} n\right)$$

$$\begin{bmatrix} X[0] \\ \boxed{X[1]} \\ \vdots \\ X[N-1] \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & w^1 & \cdots & \vdots \\ \vdots & & \ddots & \vdots \\ 1 & w^{N-1} & w^{(N-1)\cdot 2} \cdots & w^{(N-1)(N-1)} \end{bmatrix} \begin{bmatrix} x[0] \\ \vdots \\ x[N-1] \end{bmatrix} \quad \text{where } w = \exp\left(-j\frac{2\pi}{N}\right)$$

Amount of frequency $X[1]$ = similarity (dot product) b/w $\boxed{1 \; w^1 \cdots w^{N-1}}$ and signal

$w = \exp\left(-j\frac{2\pi}{N}\right)$ means



∴ Fourier Transform matrix

= cosine fundamental frequency + i sine fundamental frequency

Circulant matrix and convolution

$\Rightarrow$ Matrix that operates cyclic permutation

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} \rightarrow Px = \begin{bmatrix} x_{n-1} \\ x_0 \\ \vdots \\ x_{n-2} \end{bmatrix} \qquad \therefore P = \begin{bmatrix} 0 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ 0 & & & \\ 0 & \cdots & 1 & 0 \end{bmatrix}$$

- Decomposition of a signal vector

$$\delta \text{ (discrete unit sample function)} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

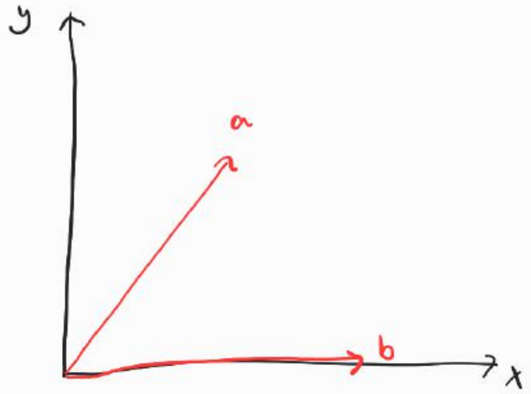
$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} = x_0 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + x_{n-1} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = x_0\delta + x_1 P\delta + \cdots + x_{n-1}P^{n-1}\delta$$

$$= (x_0 I + x_1 P + \cdots + x_{n-1}P^{n-1})\delta$$

$$= \underbrace{\begin{bmatrix} x_0 & x_{n-1} & \cdots & x_1 \\ x_1 & \ddots & & \vdots \\ \vdots & & \ddots & \\ x_{n-1} & \cdots & & x_0 \end{bmatrix}}_{\text{circulant matrix}} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Correlation and inner product

For data X and Y, correlation $r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{Y}}{s_Y} \right)$



$a \cdot b = |a||b| \cos\theta$

$= (|a|\cos\theta)(|b|)$

$=$ (Projection of a onto b) (length of b)

$=$ how much of the change of a can be explained by b

similarly, $\text{proj}_b a = \frac{a \cdot b}{|a|}$

$=$ how much — b — explained by a

$\therefore$ a and b explains each other $= \frac{a \cdot b}{|a||b|} = \cos\theta$

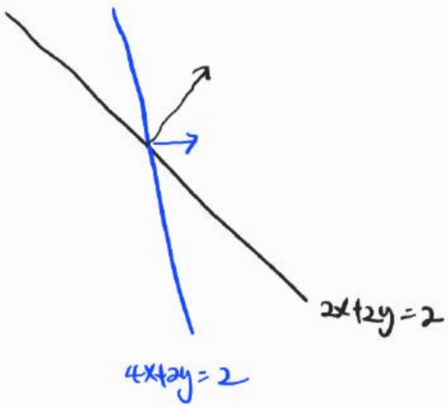
if $a = x_i - \bar{x}$, $b = y_i - \bar{Y}$,

$r = \frac{a \cdot b}{|a||b|}$ $\quad \therefore$ how much $x_i - \bar{x}$ and $y_i - \bar{Y}$ explains each other?

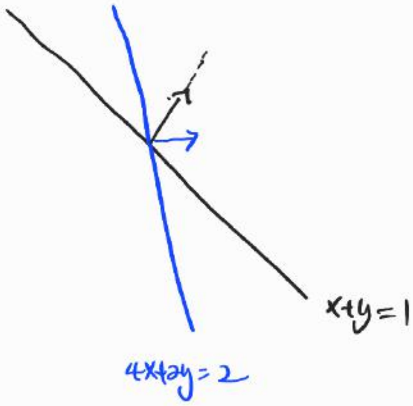
Geometric interpretation of Gauss-Jordan Elimination

⇒ transforming normal vectors of line equation into unit vectors parallel to each other

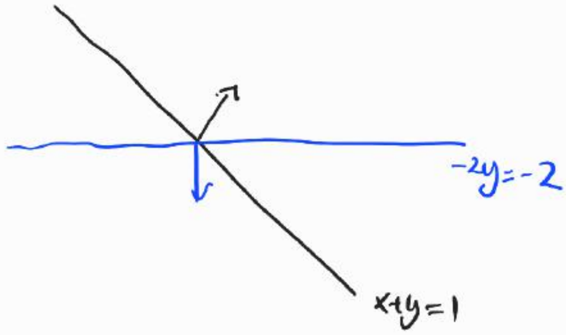
$$\begin{cases} 2x+2y=2 \\ 4x+2y=2 \end{cases}$$



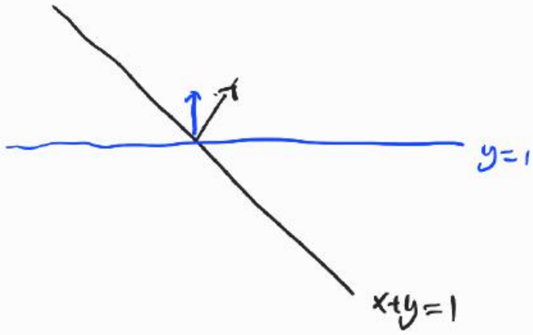
(i) Scaling $2x+2y=2 \rightarrow x+y=1$



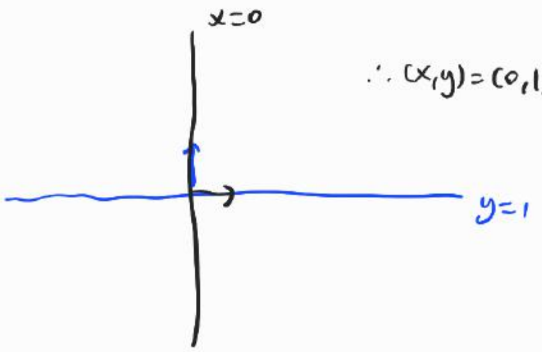
(ii) subtraction $4x+2y=2 \rightarrow -2y=-2$



(iii) Scaling $-2y=2 \rightarrow y=1$



(iv) Subtraction $x+y=1 \rightarrow x=0$



$\therefore (x,y)=(0,1)$

Wronskian function

Suppose functions $f_1(x), f_2(x), \cdots, f_n(x)$ possesses at least $n-1$ derivatives, then if determinent

$$w(f_1, f_2, \cdots f_n) = \begin{vmatrix} f_1 & f_2 & \cdots & f_n \\ f'_1 & f'_2 & \cdots & f'_n \\ \vdots & & & \vdots \\ f_1^{(n-1)} & & \cdots & f_n^{(n-1)} \end{vmatrix} \neq 0$$

$f_1, f_2, \cdots f_n$ are linearly independent

* $w \neq 0 \not\Rightarrow f_1, f_2, \cdots, f_n$ are linearly dependent

- Proof by contradiction

  suppose $w \neq 0$ and functions are linearly dependent and

$$c_1 f_1 + c_2 f_2 + \cdots + c_n f_n = 0$$

$$\vdots$$

$$c_1 f_1^{(n-1)} + c_2 f_2^{(n-1)} + \cdots + c_n f_n^{(n-1)} = 0$$

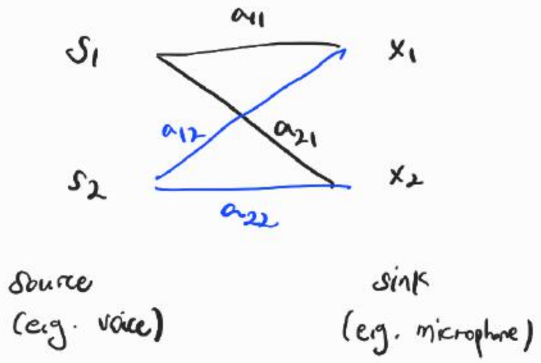
$$\underset{A}{\begin{pmatrix} f_1 & f_2 & \cdots & f_n \\ f'_1 & f'_2 & \cdots & f'_n \\ & & & \\ f_1^{(n-1)} & f_2^{(n-1)} & \cdots & f_n^{(n-1)} \end{pmatrix}} \underset{X}{\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}} = \underset{b}{\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}} \qquad AX = b$$

if $\det(A) = W$, $W \neq 0$, A^{-1} exists $\therefore X = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = 0$

$c_1 f_1 + \cdots + c_n f_n = 0 \rightarrow (c_1, c_2, \cdots, c_n) = (0, 0, \cdots, 0)$ is the only solution

$\therefore f_1, f_2, \cdots, f_n$ are linearly independent ※

Independent Component Analysis (ICA)



$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t)$

$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t)$

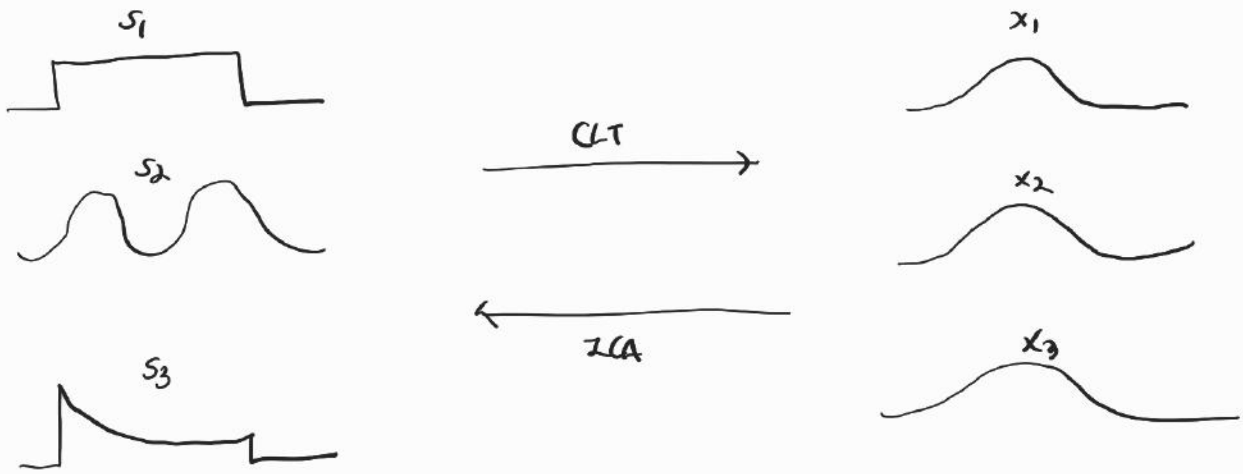
$X = As$ (A : mixing matrix)

$\downarrow$

$s = A^{-1}x = Wx$ ($A^{-1} = W$: unmixing matrix)

ICA : find W (from sink to source) w/o knowing A

- Central Limit Theorem (CLT) $\longleftrightarrow$ ICA



e.g. $s \sim \text{Uniform}[0,1] \rightarrow A=2 \longrightarrow x \sim \text{Uniform}[0,2]$

$X = AS$

$S = A^{-1}X = WX$

$p_X(x) = |W|\, p_S(WX)$

Geometric interpretation of Cramer's Rule

* $\det\left(\begin{bmatrix} a & kb \\ c & kd \end{bmatrix}\right) = k\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right)$

$\det([A_1, A_2, \cdots, k_1B_1 + k_2B_2, \cdots, A_n])$ ($A\cdot, B\cdot$ are col vectors)

$= k_1\det([A_1, \cdots, B_1, \cdots, A_n]) + k_2\det([A_1, \cdots, B_2, \cdots, A_n])$

- Cramer's Rule

$Ax = b \Rightarrow x_i = \dfrac{\det(A_i^{rep})}{\det(A)}$ where $A_i^{rep} = \begin{bmatrix} a_{11} & \cdots & b_1 & \cdots & a_{1n} \\ \vdots & & \ddots & & \vdots \\ a_{1n} & \cdots & b_n & \cdots & a_{nn} \end{bmatrix}$

- proof

$$Ax = x_1\begin{bmatrix} | \\ A_1 \\ | \end{bmatrix} + x_2\begin{bmatrix} | \\ A_2 \\ | \end{bmatrix} + \cdots + x_n\begin{bmatrix} | \\ A_n \\ | \end{bmatrix} = b$$

$$A_i^{rep} = [A_1, A_2, \cdots, b, \cdots, A_n]$$

$$\det(A_i^{rep}) = \det([A_1, A_2, \cdots, b, \cdots A_n]) = \det([A_1, \cdots, x_1A_1 + x_2A_2 + \cdots + x_nA_n, \cdots A_n])$$

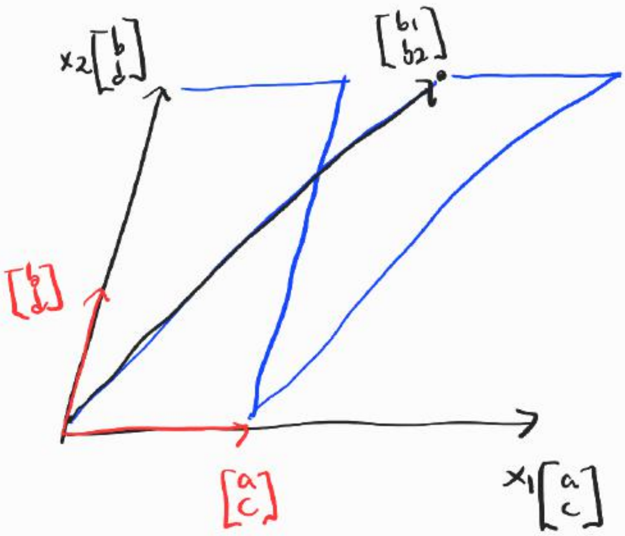
$$= \sum_{j=1}^{n} x_j \det([A_1, A_2, \cdots, A_j, \cdots, A_n])$$

$= x_i\det([A_1, A_2, \cdots A_i, \cdots A_n])$ ∵ linearly dependent col vector det $= 0$

$= x_i \det A$

- Geometric interpretation

$Ax = b \quad x_1\begin{bmatrix} a \\ c \end{bmatrix} + x_2\begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$



$$\det\left(\begin{bmatrix} a & x_2b \\ c & x_2d \end{bmatrix}\right) = \det\left(\begin{bmatrix} a & b_1 \\ c & b_2 \end{bmatrix}\right)$$

$$\therefore x_2 = \frac{\det\left(\begin{bmatrix} a & b_1 \\ c & b_2 \end{bmatrix}\right)}{\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right)}$$

Cholesky decomposition

- LU decomposition of symmetric matrix

Hypothesis : If A is symmetric matrix, $A = LL^T = L^TL$

$$|Lx|^2 = (Lx)^T(Lx) = x^T(L^TL)x$$
$$= x^TAx \quad (\text{if } L^TL = A)$$

$\therefore x^TAx \geq 0 \rightarrow$ A is semi-positive matrix

If (i) A is semi-positive matrix (ii) A is symmetric matrix (iii) A is square matrix

$A = LL^T = L^TL \Rightarrow$ cholesky factorization

Positive definite matrix

$x^T A x > 0$

A is positive → A x reverse the direction, only changes the magnitude

$a \cdot b = a^T b = |a||b| \cos\theta$ → $a^T b > 0$ if $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$

$x^T(Ax)$: Dot product b/w original x and linearly transformed Ax

= Angle difference b/w x and Ax is $[-\frac{\pi}{2}, \frac{\pi}{2}]$

- positive definite matrix and eigenvalue

$Ax = \lambda x$

$x^T A x = x^T \lambda x = \lambda |x|^2 > 0 \qquad \therefore \lambda > 0$

∴ All eigenvalues are positive

- positive definite matrix and Hessian matrix

If H is positive definite matrix, f is convex downwards (have local minimum)

Mahalanobis distance

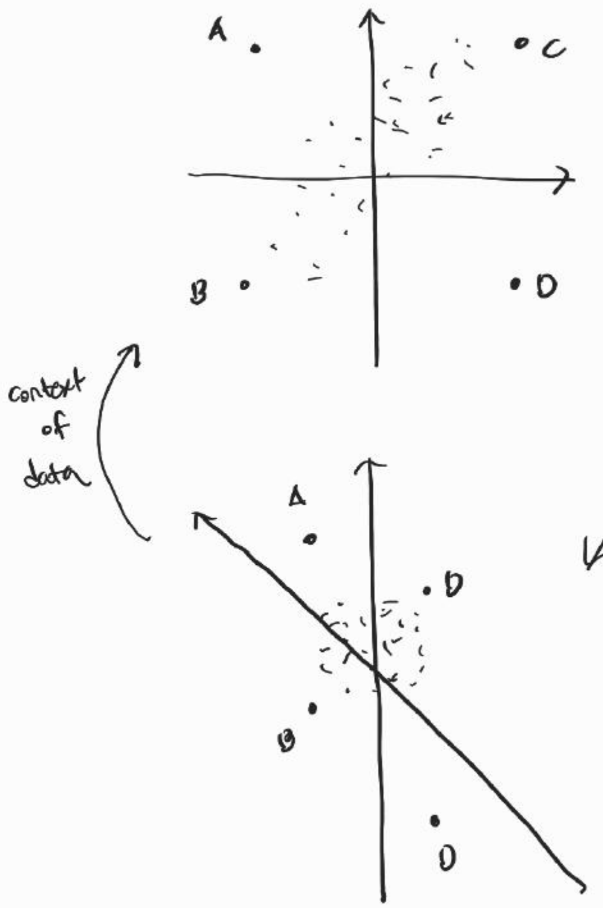
⇒ Contextual relative distance

Euclidean distance : d(AD) = d(BC)

Mahalanobis distance : d(AD) > d(BC)

$$d_E = \sqrt{(x-y)(x-y)^T}$$

$$d_M = \sqrt{(x-y)\,\Sigma^{-1}(x-y)^T}$$



Normalizing the context