PCA : Principal Component Analysis

→ introduced by Pearson (1901)

$x^i \in \mathbb{R}^D$, D is often very large

$i \equiv$ sample index

$i = 1, \dots, N$

→ $x^i_k \equiv k^{th}$ component of x^i

→ Data points are centered (assumption)

$\frac{1}{N} \sum_i x^i_k = 0 \quad \forall k$ (mean comes out to 0)

$\langle x_k \rangle = 0$; $\langle x_k \rangle \equiv \frac{1}{N} \sum_i x_k{}^i$

$\langle x \rangle = 0$ (avg of x)

$x \equiv$ vector (often we drop the lower index)

→ PCA performs an explicit dimension reduction.

→ Dimensionality reduction means performing the following transformation

$x^i \in \mathbb{R}^D \longrightarrow y^i \in \mathbb{R}^d$

, $\{d << D\}$

→ $y^i = f_\pi(x^i)$ [f: function of variational parameters]

→ generally $f_\pi()$ is nonlinear, but in the case of PCA, $f_\pi()$ is a linear transformation.

→ So explicitly func^n that transforms original data to reduced representation is a linear transformation of the original data.

→ PCA :- f is a linear function.

→ $y^i = A x^i$; A = matrix, [$y^i, x^i \equiv$ vectors]

→ $\because x^i$ & y^i have vastly diff. no of dimensions, A must be a rectangular matrix.

→ if $x^i \equiv$ row vectors (convention)
$y^i \equiv$ col " (")

$A \rightarrow \underset{\sim}{d}$ rows $\times$ $\underline{D}$ cols
(y^i dim) (x^i dim)

→ in PCA, the idea is that we determine $\underline{A}$, such that, $\{Y\}$ describes a large fraction of the covariance of the original dataset.

→ what is covariance?

(transpose)

① $C_{D\times D} = \langle X \cdot X^t \rangle - \langle X \rangle \langle X^t \rangle$

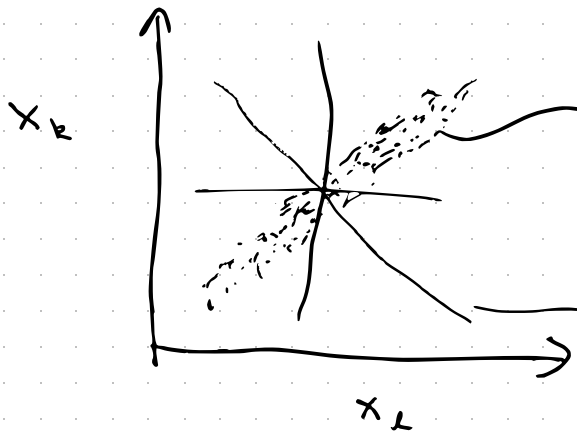
→ (covariance matrix)

→ = 0 if data is centered.

$\therefore C_{k,l} = \langle X_k \cdot X_l \rangle$

↳ ($\{k,l\}$ element of cov. matrix)

$\therefore C_{k,l} = \frac{1}{N} \sum_i x_k^i x_l^i$



C = symmetric matrix

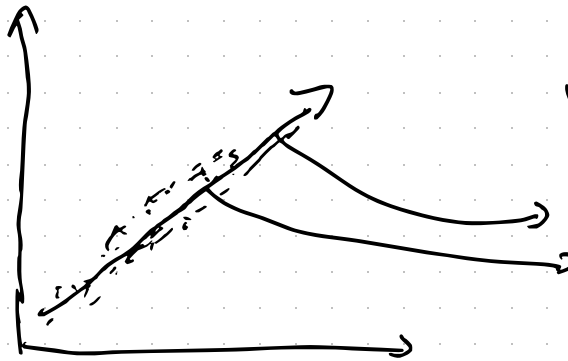
let's say $C = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1.05 \end{pmatrix}$

Target dimensionality reduction from 2 → 1)

$\Rightarrow d = 1$

$\Rightarrow y^i = a x_1^i + b x_2^i$

$\Rightarrow a^2 + b^2 = 1$ (2 coeffs are constricted)



$Y = aX_1 + bX_2$

$Y \Rightarrow a = 1;\ b = 0$

$\text{var}(Y) = \langle Y^2 \rangle$

$\therefore$ we use only a single component here.

→ so this is a v. suitable choice for Y.

→ so how do we find Y?

(1) We diagonalize the covariance matrix of $\{X\}$.

This means we are solving the following eigenvalue eqⁿ :-

$$CY = \lambda Y$$

$$\rightarrow C = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1.05 \end{pmatrix} \quad \therefore \begin{cases} \lambda_1 = 1.89 \\ \lambda_2 = 0.12 \end{cases}$$

(2 eigenvalues)

→ 2 eigen-vectors :

$$Y_1 = (0.7\ \ 0.76)$$

$$Y_2 = (-0.71\ \ 0.7)$$

∵ the covariance matrix is symmetric,

→ the left & right eigenvectors are the same

→ " " " " are mutually orthogonal, i.e, they form an orthonormal basis set.

→ λ_1 = Variance along a Y coordinate defined by Y_1

λ_2 = ---- Y coordinate .... Y_2,

→ If Y_1 is Y coord, Var = 1.89

" Y_2 " " " Var = 0.12

so here Y_2 is appropriate choice of reduced dimension

⊙ Qualitative pipeline for PCA :-

Given = dataset

⊙ compute covariance D×D matrix

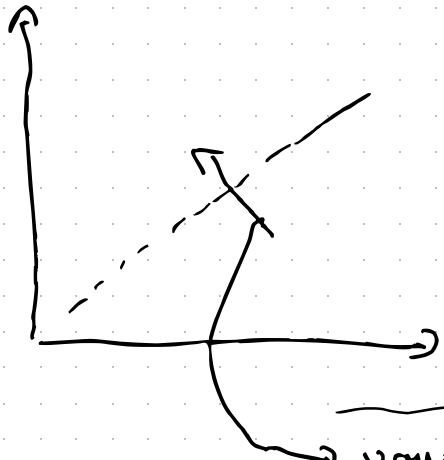
⊙ Find the eigenvalues & eigenvectors

⊙ if we know 'd', choose as Y coordinates the first 'd' eigenvectors of C

[Assume : $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_D$]

↓ eigenvalues are sorted in descending order.

→ Specific Case :- data pts belong to a hyperplane

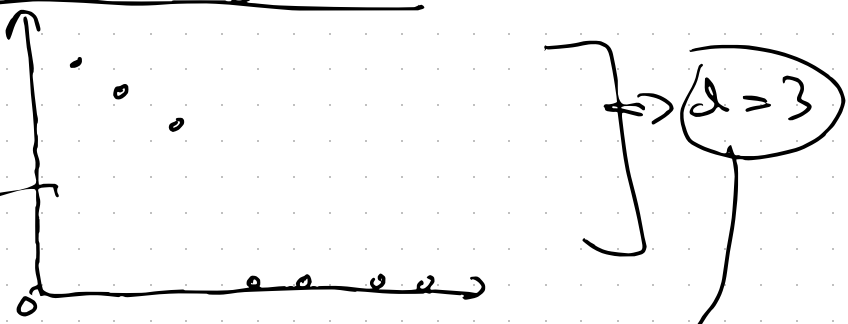


if X^i belongs to a hyperplane of dimension 'd' of the n data pts will be exactly 'c' = 'd'

→ variance along this dir $\hat{=} 0$,
∵ the 2 eigenvectors are orthogonal to each other, & all data points lie on the other eigenvector

→ generally, if $D = 1000$, all data pts are on a hyperplane of dim = 10,
then 990 eigenvalues $= 0$,
10 " $\neq 0$,
→ ∵ the 990 eigenvectors will be orthogonal & have no points

→ plot the eigenvalues :



⇒ implies that the embedding manifold CAN BE a hyperplane

⇒ This is a necessary but not sufficient cond^n for embedding manifold to be a hyperplane

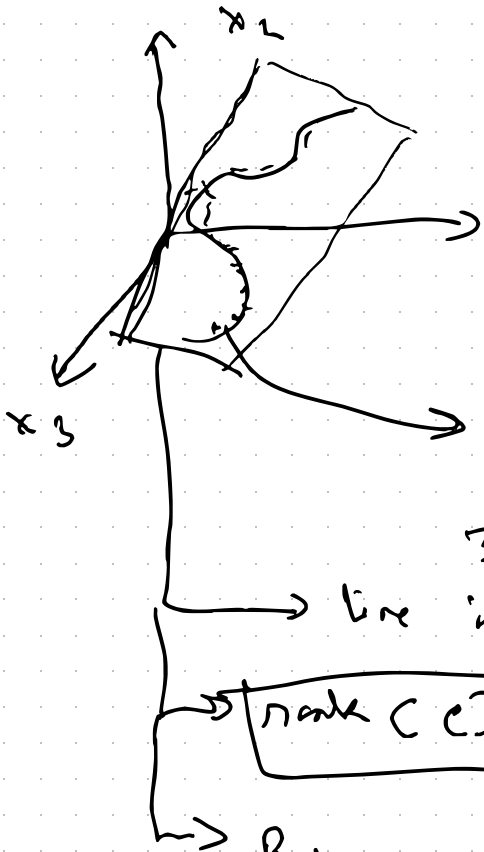
⇒ sufficient cond^n : $\text{rank}(C) = ID$

↳ (reason in next page)

→ $\text{rank}(C) \cong r_e$

→ This only implies $(r_e \geq ID)$

⇒ Another case:



→ all data pts. lie on a 2D line in 3D space.

→ line is on a hyperplane

→ $\boxed{\text{rank}(C) = 2}$

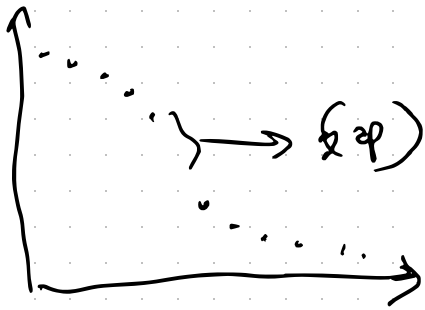
→ But, embedding manifold is NOT a hyperplane, ∵ here it is a line.

→ Here, rank = 2, ID = 1, so, not a hyperplane.

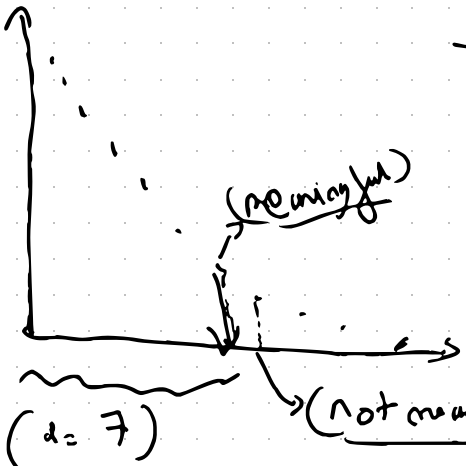
→ Real-world scenario (1) :

(1) General rule in $P(D)$:

gap in eigenvalues $\{\lambda_i\}$



→ Real-world scenario (2) :



→ (NO gap)

⇓

data points are not constrained in a 'd' hyperplane

→ If we say here (by truncating at $d = 7$), then we say that $d > 7$ are ignored & we can't put a meaningful upper-bound on the error.

→ This is because we don't have a gap, so we choose arbitrarily a point to cut-off the eigenvalues.

⇒ "Total" variance of X = $\mathrm{Tr}\, C$ (Trace)

$$\mathrm{Tr}\, C = \sum_{i=1}^{D} \lambda_i$$

(sum of eigenvalues)

if $n_c = d$,

$$\mathrm{Tr}\, C = \sum_{i=1}^{d} \lambda_i$$

let's say truncation level = d :

$$\eta(d) = \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i}$$

→ $\eta(d) \equiv$ fraction of variance described by truncated description.

→ a common way of choosing $'d'$ is to reach a pre-defined value of $\eta(d)$

↳ for eg, obtain $'d'$ s.t. $\eta(d) \geq 0.95$

→ Formal PCA derivation.

$x^i \in \mathbb{R}^D$, $y^i \in \mathbb{R}^d$

Y :- 1) $Y = AX$, $y^i_k = \sum_{l=1}^{D} A_{kl} x^i_l$

2) $\sum_l A_{kl} A_{ml} = \delta_{km}$

3) Trace of the covariance matrix of Y is maximal.

(if "d" is given)

→ $Tr((cov(Y)) = \langle Y, Y^T \rangle$

$= \left\langle \sum_{np} \underbrace{A_{ln} x_n}_{Y_l} \underbrace{A_{pm} x_p}_{Y_m} \right\rangle$

$= \sum_{np} A_{ln} A_{pm} C_{np}$

$\because Y = AX$

$\langle Y_l Y_m \rangle = \sum_{kh} A_{ln} A_{mk} C_{hk}$ ← $C_{np} = \langle x_n x_p \rangle$

$cov(Y)$ → $Tr((cov(Y)) = \sum_l \langle Y_l Y_m \rangle$

(i) Condition :- all coeffs of transformation matrix are normalized so that eigenvectors necessarily form an orthonormal basis set.

$$\sum_{n} A_{ln}^2 = 1 \quad \forall l$$

(for normalization)

→ PCA Loss function :-

→ rename ... $Y^i = A x^i$

$Tr(A^t C A)$

$$L(A) = \left(\sum_{lkh} A_{lh} A_{lk} C_{hk} \right) + \sum_{l=1}^{d} \lambda_l \left(\sum_k A_{lk}^2 - 1 \right)$$

(due to normalization)

(we have (d) normalization cond^ns)

(Lagrange multipliers)

→ sum over diff. Lagrange multipliers

→ globally convex func^n of A ; ∴ it has a single maxima.

$$\frac{\partial L}{\partial A_{hk}} = 0 \rightarrow \forall h, k$$

$$\Rightarrow \sum_i A_{hi} C_{ik} - \lambda_k A_{hk} = 0$$

$$\boxed{AC = \lambda A}$$ (in matrix formulation)

⊙ The coefficients of the linear transformation defining $Y = AX$ with the cond^n that the trace of the covariance of Y is maximal are the 'd' eigenvectors of the matrix 'C'

11/11

⇒ TL;DR :- Recap of PCA from last lec :-

⊙ PCA (from last lecture → contd.)

$x^i \in \mathbb{R}^D \longrightarrow y^i \in \mathbb{R}^\delta$

(dataset) $(\delta << D)$ (ideally)

$y^i = A x^i$ ( ) (Linear transformation allows PCA)

A ↓ rectangular matrix $(\delta \times D)$

↳ each row of this matrix is an eigenvector of the covariance matrix.

→ $C = \langle x x^t \rangle$

(sum of convention let's define $x^i \equiv$ col vector)

↳ $(D \times D)$

→ $\underline{A}$ has first 5 eigenvectors of the $\underline{C}$.

→ A is chosen s.t. cov(Y) is as large as possible.

$$\begin{aligned} cov(Y) &= \langle Y Y^t \rangle \\ &= \langle A X X^t A^t \rangle \\ &= A \cdot C \cdot A^t \\ &\downarrow \\ C &= cov(X) \end{aligned}$$

→ $Y = AX$

→ A can be taken out of $\langle \rangle$ as it doesn't depend on the index

$$\therefore \boxed{cov(Y) = A \cdot C \cdot A^t}$$

⇒ Goal: minimize $Tr(A C A^t)$, under the constraint that each row of A matrix is normalized.

$$\therefore \boxed{\sum_i A_{kl}^2 = 1 \quad \forall k}$$

→ To impose these constraints we use the Lagrange Multipliers term (λ_k)

↓

$$\sum_{kL} \lambda_k \left(A^2_{kL} - 1 \right)$$

→ $\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & & \\ 0 & & \lambda_3 & \\ \vdots & & & \ddots \end{pmatrix} \equiv (5 \times 5)$ diagonal matrix

(5×5)

↓

all the diagonal elements are Lagrange multipliers. It's a (5×5) matrix.

$$\therefore \sum_{kL} \lambda_k \left(A^2_{kL} - 1 \right)$$

$$= Tr \left(\Lambda A A^t \right)$$

⇒ Loss function to be minimized:

$$L(A) = Tr \left(ACA^t - \Lambda AA^t \right)$$

$$\frac{\partial L}{\partial A^t} = 0 \Rightarrow \frac{\partial}{\partial A^t} Tr\left(ACA^t - \Lambda AA^t\right) = 0$$

$$\Rightarrow \frac{\partial}{\partial A^t}\left(ACA^t - \Lambda AA^t\right) = 0$$

$$\rightarrow AC - \Lambda A = 0$$

$$\Rightarrow AC = \Lambda A$$

(doing this directly gives us the eigenvalue eq^n)

$\Rightarrow$ A must contain eigenvectors of C

→ We want to maximize $Tr(Cov(Y))$

$$Tr\ Cov(Y) = Tr(ACA^t)$$
$$= Tr(\Lambda AA^t)$$

↳ (if A satisfies $AC = \Lambda A$ eigenval eq^n)

→ ∵ C ≡ symm. matrix

∴ $AA^t = I_{\delta \times \delta}$

(We consider only δ-orthog eigenvectors)

$$\therefore \text{Tr}\ \text{Cov}(Y) = \text{Tr}\left(\Lambda\, A A^{t}\right)$$

$$= \text{Tr}\left(\Lambda\, \underset{5\times 5}{I}\right)$$

$$= \text{Tr}(\Lambda)$$

$$= \sum_{i=1}^{\delta} \lambda_i$$

$\delta \equiv$ no. of chosen eigenvectors

→ To maximize fraction of chosen variance we maximize $\text{Tr}\ \text{Cov}(Y)$

⇒ We have to choose the δ - largest eigenvalues.

→ The largest possible $\text{Tr}\ \text{Cov}(Y)$ is chosen by taking δ- largest eigenvalues.

TL;DR : ① compute covariance matrix of the data.

② find eigenvalues & eigenvectors

③ see the spectrum & choose truncation criteria upto d dimensions → either w.r.t. a gap, or w.r.t. reproducing a given fraction of the variance of the original data.

PCA allows to find explicitly the coordinates on the embedding manifold iff the manifold is on a hyperplane.

But, in many cases the embedding manifold is NOT on a hyperplane