# Principal Component Analysis (PCA)

Dimensionality Reduction — exploit [structure / correlation] → ↓noise → compact representation of data
High dimension data — compression algorithm — low Dimension Data

Statistics of datasets – compact way to describe data
Average — Mean $E(D) = \frac{1}{N}\sum_{i=1}^{N} D_i$
Spread — 1D Variance $E(X-\mu)^2$
nD Covariance $E(X-\mu_X)(Y-\mu_Y)$ → Covariance Matrix D×D

Linear transformation
$E(aD+b) = aE(D)+b$
$\text{Var}(aD+a) = a^2\text{Var}(D)$ $\text{var}(AD+b) = A\,\text{Var}(D)\,A^T$

Inner product
Dimensionality Reduction → Compression — Geometric Space — Orthogonality — Distance (vector), Angle (vector space)

→ Inner product → Dot Product $xy = x^Ty = \sum x_i y_i$ — $x\cdot x = \|x\|$ — distance; $\cos\theta = \frac{x^Ty}{\|x\|\|y\|}$ — angle
Dot Product → Generalize → Inner Product
Inner Product — mapping $\langle\cdot,\cdot\rangle : V\times V \to R$
- Symmetric $\langle x,y\rangle = \langle y,x\rangle$
- Positive Definite $\langle x,x\rangle \geq 0$; $\langle x,x\rangle = 0 \Leftrightarrow x = 0$
- Bilinear $\langle \lambda x+z, y\rangle = \lambda\langle x,y\rangle + \langle z,y\rangle$; $\langle x, \lambda y+z\rangle = \lambda\langle x,y\rangle + \langle x,z\rangle$

Inner Product — Length/norm $\|x\| = \sqrt{\langle x,x\rangle}$
Properties:
- $\|\lambda x\| = |\lambda|\,\|x\|$
- $\|x+y\| \leq \|x\| + \|y\|$ Triangle inequality
- $|\langle x,y\rangle| \leq \|x\|\,\|y\|$ Cauchy-Schwarz inequality

Angle $\cos\theta = \frac{\langle x,y\rangle}{\|x\|\,\|y\|}$ — $\|\langle y,y\rangle\|$
- similarity
- orthogonal $\theta = 90°$, nothing in common except origin

Orthonormal basis vector (ONB) — 90° — Unit length

finite number of entries

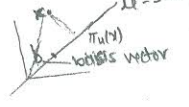
function
- Discrete
- Continuous — $\langle u,v\rangle = \int_a^b u(x)\,v(x)\,dx$ — orthogonality = 0

random variable — $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]$ if X & Y are uncorrelated random variable
$\langle X,Y\rangle = \text{Cov}(X,Y)$ $\|X\| = \sigma(X)$
$\cos\theta = \frac{\langle X,Y\rangle}{\sigma(X)\,\sigma(Y)}$

$(AB)^T = B^TA^T$

Orthogonal Projection
High dimension data ——compress——→ low dimension data
few dimension contains most information

to 1-D subspace




$U$ - subspace, orthogonal projection on subspace $U$ with basis vector $b$
1) $\pi_U(x) \in U \Rightarrow \exists\,\lambda\in\mathbb{R} : \pi_U(x) = \lambda b$ — coordinate with basis $b$ in subspace $U$.
2) $\langle b, \pi_U(x) - x\rangle = 0$ → orthogonal
①&② $\lambda = \frac{\langle b,x\rangle}{\|b\|^2}$

$\pi_U(x) = \frac{\langle b,x\rangle}{\|b\|^2} b = \underbrace{\frac{bb^T}{\|b\|^2}}_{\text{projection matrix}} x$ in terms of dot product

to n-D subspace
D dimension vector to M dim subspace
1) $\pi_U(x) = \sum_{i=1}^{M} \lambda_i b_i = B\lambda$
2) $\langle \pi_U(x) - x, b_i\rangle = 0$ for $i = 1,2,\dots M$
①&② $\lambda = (B^TB)^{-1}B^Tx$
$\pi_U(x) = \underbrace{B(B^TB)^{-1}B^T}_{\text{Projection Matrix}} x$ ONB $\pi_U(x) = BB^Tx$

$\lambda = \begin{bmatrix}\lambda_1\\ \lambda_2\\ \vdots\\ \lambda_M\end{bmatrix}$ $M\times 1$ $\quad B = [b_1\ b_2\ \dots\ b_M]$ $D\times M$ — basis vector

PCA
Derivation

Data set $X = \{x_1, x_2, \dots x_N\}$ $x_i \in \mathbb{R}^D$ — find low dimension data similar to X
① $x_n = \sum_{i=1}^{D} \beta_{in} b_i$ — Every vector in $\mathbb{R}^D$ can be represented as linear combination of basis vector
↳ orthonormal basis
② $\beta_{in} = x_n^T b_i$ $\beta_{in}$ - orthogonal projection of $x_n$ to one dim subspace spanned by $i^{th}$ basis vector
③ $B = [b_1\ b_2\ \dots\ b_M]$
$\tilde{x} = B\underbrace{B^Tx}_{\text{code}}$ orthogonal projection of x on subspace span by B
↳ co-ordinate of $\tilde{x}$ wrt basis vector connected in B

PCA
find lower dimension representation $\tilde{x}_n \leftarrow \tilde{x}$
fewer basis vector
Assume - Centered Data $E[x] = 0$
- ONB $b_1 \dots b_D$

Ⓐ $\tilde{x}_n = \sum_{i=1}^{M} \beta_{in} b_i + \sum_{i=M+1}^{D} \beta_{in} b_i \in \mathbb{R}^D$
coordinate — M dim subspace — D-M dim subspace
$b_1$ to $b_M$ span principal subspace — Orthogonal complimentary ignore in PCA

Ⓐ $\tilde{x}_n = \sum_{j=1}^{M} \beta_{jn} b_j \overset{①}{=} \sum_{j=1}^{M} (x_n^Tb_j) b_j = \left(\sum_{j=1}^{M} b_j b_j^T\right) x_n$ — $\tilde{x}_n$ is orthogonal projection of $x_n$ on subspace spanned by M basis vector $b_j$ where $j = 1$ to $M$
projection matrix

$x_n = \left(\sum_{j=1}^{M} b_j b_j^T\right) x_n + \left(\sum_{j=M+1}^{D} b_j b_j^T\right) x_n$ ⇔ $x_n - \tilde{x}_n = \left(\sum_{j=M+1}^{D} b_j b_j^T\right) x_n = \sum_{j=M+1}^{D} (b_j^Tx_n)\, b_j$ Ⓔ
Displacement vector — orthogonal compliment to principle subspace

Given X, find $\beta_{in}$ of ONB $b_i$, average square reconstruction error is minimized (J)
Ⓑ $J = \frac{1}{N}\sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2$ Ⓔ + ONB $= \sum_{j=M+1}^{D} b_j^T \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right) b_j$
data covariance (S)
matrix ← $E(x) = 0$

$= \sum_{j=M+1}^{D} b_j^T S b_j = \text{trace}\left(\left(\sum_{j=M+1}^{D} b_j b_j^T\right) S\right)$ Ⓕ
projection matrix → data covariance matrix → project onto orthogonal compliment to the principal subspace

$\frac{\delta J}{\delta \beta_{in}, b_i^T} = \frac{\delta J}{\delta \tilde{x}_n} \times \frac{\delta \tilde{x}_n}{\delta \beta_{in}, b_i^T}$
Ⓒ $\frac{\delta J}{\delta \tilde{x}_n} = -\frac{2}{N}(x_n - \tilde{x}_n)^T$ $\quad \frac{\delta \tilde{x}_n}{\delta \beta_{in}} = b_i$ $i = 1,2\dots,M$
$\frac{\delta J}{\delta \beta_{in}} = -\frac{2}{N}(x_n - \tilde{x}_n)^T b_i \overset{\text{ONB}}{=} -\frac{2}{N}(x_n^Tb_i - \beta_{in}) = 0$
Ⓓ ⇔ $\beta_{in} = x_n^T b_i$ optimal coordinate of $\tilde{x}_n$ Ⓐ wrt our basis
orthogonal projection of coordinate of original data onto $i^{th}$ basis vector that spans our principal subspace.

Minimize loss/ matrix - average square reconstruction error
= minimize variance of data projected onto subspace orthogonal to principal subspace
= retain as much as variance after projection as possible.

Orthonormal basis for M-dim subspace
Ⓕ J + constraint of orthonormal basis vector
Optimize → Lagrangian → Gradient wrt $\lambda$ & basis vector = 0
$Sb_j = \lambda_j b_j$ where $j = M+1,\dots,D$
$J = \sum_{j=M+1}^{D} \lambda_j$
J is minimum ↳ $\lambda$/eigen value is minimum → choose basis vectors that span ignored subspace to be eigen vectors of S that belong to smallest eigen values.
= principal subspace - spanned by - eigen vectors belonging to M largest eigen values of S.

Covariance matrix — Symmetry — eigen vector are orthogonal to each other
eigen vector with largest eigen value points in the direction of data with largest variance and variance in that direction is given by corresponding eigen value

PCA - Algorithm

- Steps 1) Data Normalization
  a) $x_i' = x_i - \bar{x}$ mean normalization [not necessary, numerical difficulties → covariance
  b) $x^* = x_1/\sigma_x$ [unit free, variance in all dimension = 1, while correlation is intact
  2) Covariance Matrix (S) → [Eigen values, Eigen vectors
  3) Project to principal subspace that is spanned by the eigenvectors that belong to largest eigenvalues

Data covariance Matrix
in D dimension = D × D → computational intensive to calculate [eigen value, eigen vector $O(n^3)$ d dim

Dataset $x_1, x_2, \dots, x_N \in \mathbb{R}^D$

$S = \frac{1}{N} X^T X$ $\quad X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} \in \mathbb{R}^{N \times D}$ $\quad N < D$

no. of datpoints < dimensionality of data

Covariance Matrix (S)
- rank(S) = N
- D − (N+1) eigen values = 0
- ∴ full rank matrix
- rows & columns are linearly dependent

Convert D×D covariance matrix into
→ full rank N×N covariance matrix without eigen values 0.

PCA → $Sb_i = \lambda_i b_i$

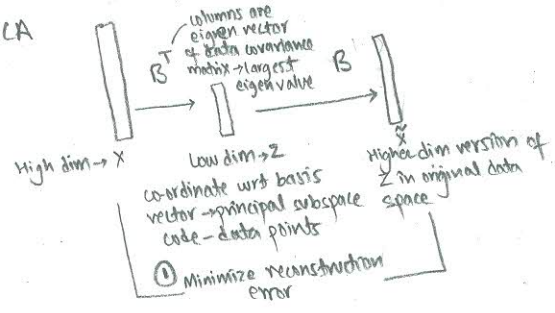
$\frac{1}{N} X^T X b_i = \lambda_i b_i \Leftrightarrow \underbrace{\frac{1}{N} X X^T}_{\in \mathbb{R}^{N \times N}} \underbrace{X b_i}_{c_i} = \lambda_i \underbrace{X b_i}_{c_i}$ (eigen value, eigen vector)

covariance matrix
- non zero eigen values

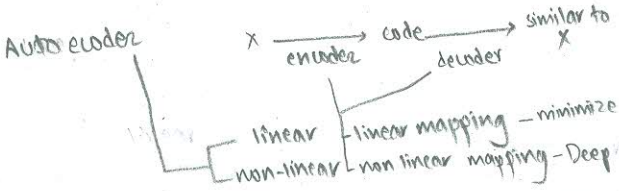
Recover original eigen vector → PCA

$\underbrace{\frac{1}{N} X^T}_{S} \underbrace{X X^T}_{\text{eigen vector}} c_i = \underbrace{\lambda_i}_{\text{eigen value}} X^T c_i$

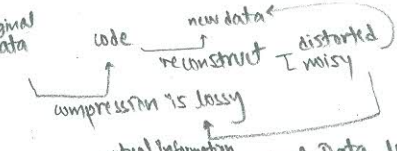
Other Interpretation of PCA




Auto encoder



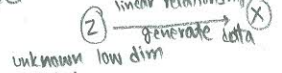

Information Theory




Maximize correlation (Original Data, lower dimension code) ③ → Mutual Information ~ PCA

④ maximizing variance of data when projected on principal subspace ↓ information contained in data

Latent Variable Model




$P(z) = N(0, I)$

$x = Bz + \mu + \epsilon$ $\quad \epsilon \sim N(0, \sigma^2 I)$ : isotropic error

Likelihood of model $P(x/z) = N(x \mid Bz + \mu, \sigma^2 I)$

Marginal likelihood $P(x) = \int p(x/z)\, p(z) \cdot dz = N(x \mid \mu, BB^T + \sigma^2 I)$

⑤ Use Maximum Likelihood Estimation (MLE) → $(\mu, B, \sigma)$ Parameter

$\mu$ = mean of data

$B$ = matrix of eigen vectors that contains largest eigen values

Find low dim code of data point,

Bayes Theorem to invert linear relationship between z & x

$P(z/x) = P(x/z) \times P(z) / P(x)$

likelihood, dist^n, marginal likelihood