PCA

(= Principal Component Analysis).

Input :

① A data matrix $X = \begin{bmatrix} - x_1 - \\ \vdots \\ - x_N - \end{bmatrix}$ $(N \times n)$, always assumed centered. i.e. $x_1 + \cdots + x_N = 0$.

where each row is a vector in $\mathbb{R}^n$

② d : a positive integer w/ $d \le n$.

Output :

① A matrix $X_d = \begin{bmatrix} - \hat{x}_1 - \\ \vdots \\ - \hat{x}_N - \end{bmatrix}$ $(N \times d)$

② Principal vectors : $\{u_1, \cdots, u_d\} \subseteq \mathbb{R}^n$.

③ Principal values : $\{\lambda_1, \cdots, \lambda_d\} \subseteq \mathbb{R}_{\ge 0}$.

There are many ways to talk about PCA.

I will start w/ one and present some other equivalent interpretations.

Interpretation 1. : (Decorrelate).

Regard the columns of X as random variables.

"Decorrelate" them via rotation (and preserve max. var.)

Let $U = \begin{pmatrix} u_1 & \cdots & u_n \end{pmatrix}$ be an orthogonal matrix.

(i.e. $\{u_1, \cdots, u_n\}$ is an o.n. basis of $\mathbb{R}^n$).

Then $Y := X \underset{N \times n}{U}$ is the coordinates w.r.t. U.

$\because X$ is centered $\therefore Y$ is again centered.

$\Rightarrow \operatorname{Cov}(Y) = Y^T Y = U^T X^T X U$.

To make Y "decorrelated", since $X^T X$ is psd (positive semi-definite), we may choose $\{u_1, \cdots, u_n\}$ as the eigenvectors of $X^T X$.

In this case,

$\operatorname{cov}(Y) = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$. i.e. Y is decorrelated.

WLOG, we may assume $\lambda_1 \ge \cdots \ge \lambda_n (\ge 0)$.

Note that in the new coordinates w.r.t. U, λ_i is the variance of the ith coordinate variable.

Thus,

Principal vectors $= \{u_1, \cdots, u_d\}$ and

" values $= \{\lambda_1, \cdots, \lambda_d\}$. #

Interpretation 2 : (Greedy)

Preserve max. var. step-by-step for d times.

Formally,

(1) find $v_1 \in \mathbb{R}^n$ w/ $\|v_1\| = 1$ s.t.

$$v_1 = \underset{\|v\|=1}{\operatorname{argmax}} \sum_{i=1}^{N} (\langle x_i, v \rangle)^2$$

(2) find $v_2 \in \mathbb{R}^n$, $v_1 \perp v_2$, w/ $\|v_2\| = 1$ s.t.

$$v_2 = \underset{\substack{\|v\|=1 \\ v \perp v_1}}{\operatorname{argmax}} \sum_{i=1}^{N} (\langle x_i, v \rangle)^2 .$$

(3) Repeat the process d times and get $v_1, \cdots, v_d$. Set these as principal vectors.

(4) $X_d := X \cdot \begin{bmatrix} v_1 & \cdots & v_d \end{bmatrix}$.

Reason of equivalence :

$$\sum_{i=1}^{N} (\langle x_i, v \rangle)^2 = ((Xv)^T (Xv))$$

$$= v^T X^T X v = \langle X^T X v, v \rangle .$$

$\because \|v\| = 1 \quad \therefore v = c_1 u_1 + \cdots + c_n u_n$ w/ $\sum_{i=1}^{n} c_i^2 = 1$.

Then $\langle X^T X v, v \rangle = \sum_{i=1}^{n} \lambda_i c_i^2 \le \sum_{i=1}^{n} \lambda_1 c_i^2 = \lambda_1$.

However, choosing $v = u_1$ can achieve this max. Thus $v_1 = u_1$.

The same arguments holds subsequently.

Interpretation 3 : (Orthogonal proj. / Max var.)

Fix $d \le n$. Find

$V = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_d \\ | & & | \end{pmatrix}$ $(n \times d)$ w/ $\{v_1, \cdots, v_d\}$: o.n.

s.t. $Var(XV)$ is maximized.

i.e. We project $\{x_1, \cdots, x_N\} \subseteq \mathbb{R}^n$ to the subsp.

$sp(\{v_1, \cdots, v_d\})$ orthogonally and preserve max. variance.

Here $Var(XV) := tr((XV)^T XV)$.

Reason :

$tr((XV)^T(XV))$

$= tr(V^T X^T X V) = tr((V^T U) U^T X^T X U (U^T V))$

$= tr\left((V^T U)\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}(U^T V)\right)$. ⊛

Denote $\langle v_i, u_j \rangle = v_{ij}$ (i.e. $v_i = \sum_{j=1}^{n} v_{ij} u_j$).

Then ⊛ $= \lambda_1 (\boxed{v_{11}^2} + \cdots + \boxed{v_{d1}^2})$
$+ \lambda_2 (\boxed{v_{12}^2} + \cdots + \boxed{v_{d2}^2})$
$+ \cdots$
$+ \lambda_n (\boxed{v_{1n}^2} + \cdots + \boxed{v_{dn}^2})$.

sum=1 sum=1

We can claim by Lagrange multiplier that the max of ⊛ occurs when

$(v_{11}, v_{12}, \cdots, v_{1n}) = (1, 0, \cdots, 0)$,

$\vdots$

$(v_{d1}, v_{d2}, \cdots, v_{dn}) = (0, 0, \cdots, 1)$.

Thus, max ⊛ $= \lambda_1 + \cdots + \lambda_d$ and we may choose $v_i = u_i$, $i = 1, \cdots, d$, to achieve maximum. #

Interpretation 4 : (min. squared loss).

This can be viewed as a lossy compression problem :

We want to compress X as follows : PCA P2

(i) Encode :

Choose a d-dim. subsp. V of $\mathbb{R}^n$ spanned orthonormally by $\{v_1, \cdots, v_d\}$.

Abusing notation, denote $V = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_d \\ | & & | \end{bmatrix}$.

Compute $X_d = XV$. ← $\boxed{N \times d}$

(ii) Decode :

We can recover X w/ some loss as

$\tilde{X} = X_d V^T = X V V^T$. ← $\boxed{N \times n}$

The goal is to find o.n. $\{v_1, \cdots, v_d\}$ s.t. X and $\tilde{X}$ have least difference. (min. loss).

Formally, we use $\| X - \tilde{X} \|_2^2$. i.e.

$$V = \underset{V}{\operatorname{argmin}} \| X - \tilde{X} \|_2^2$$

$$= \underset{V}{\operatorname{argmin}}\ tr((X - XVV^T)^T (X - XVV^T)).$$

Reason for equivalence :

$tr((X - XVV^T)^T (X - XVV^T))$

$= tr((X^T - VV^T X^T)(X - XVV^T))$

$= tr(X^T X - X^T X V V^T - \cancel{V V^T X^T} X + \cancel{V V^T X^T} X V V^T)$

$= tr(X^T X) - tr(V^T X^T X V)$.

$\because tr(X^T X)$ is const.

$\therefore$ It suffices to minimize $-tr(V^T X^T X V)$.

i.e. $V = \underset{V}{\operatorname{argmax}}\ tr(V^T X^T X V)$

This is exactly the same as interpretation 3. #