Date: 28/10/2023

Principal Component Analysis (PCA).

What is PCA?

— Technique of dimensionality reduction of observed data.

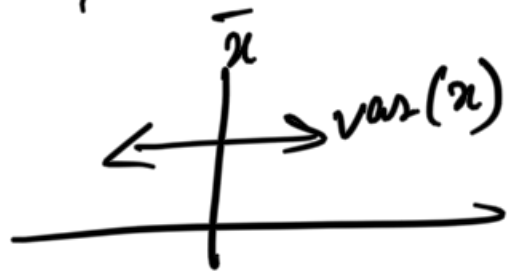
— dimensionality reduction will degrade accuracy. (Cons.)

— But DR ⟶ siplify the computation (Pros.)

— Maintain a trade-of between

Accuracy ∝

Computation Complexity ∝



Revision

— EV, EV, SV, SVector.

Let x & y are two r.v.

$$Var(x) = E\left[(x-\bar{x})^2\right], \quad \bar{x} = E(x)$$

$$Cov(x,y) = E\left[(x-\bar{x})(y-\bar{y})\right], \quad \bar{y} = E(y)$$

Covariance matrix

$$= \begin{pmatrix} \underbrace{cov(x,x)}_{var(x)} & cov(x,y) \\ cov(y,x) & \underbrace{cov(y,y)}_{var(y)} \end{pmatrix}$$

Let, an observation data/vector of 'm'- dimension (features / parameters)

$$\underline{y}_{m\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \qquad y = \underline{\mu} + W x$$

data (observation) matrix with 'n' samples.

Original Data Matrix

$$Y_{m\times n} = \begin{bmatrix} y_{1_1} & y_{1_2} & \cdots & y_{1_n} \\ y_{2_1} & y_{2_2} & \cdots & \\ y_{3_1} & y_{3_2} & \cdots & \\ \vdots & & & \\ y_{m_1} & \cdots & & y_{m_n} \end{bmatrix}$$

samples → n

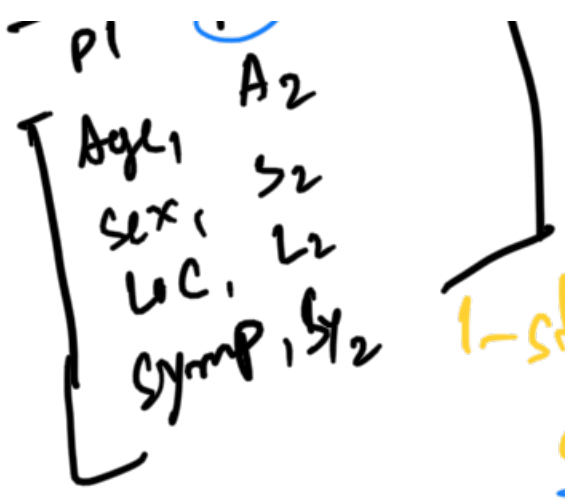
dimension.

m = 3

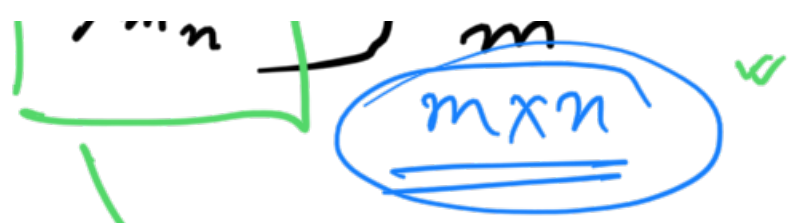
Data

Ex: Patient data

P2 P3



1-st m-dimensional data sample



n-th m-dimensional data sample.

STEPS OF PCA

⓪ After getting raw data, prepare the observation matrix.

① After getting the data (observation) matrix, find the mean in all dimensions (for each

$$\overline{y_1} = \frac{\sum_{i=1}^{n} y_{1_i}}{n} = \frac{y_{1_1} + y_{1_2} + \cdots + }{n}$$

$$\vdots$$

$$\overline{y_m} = \frac{\sum_{i=1}^{n} y_{m_i}}{n}$$

② STANDARDIZATION.

Subtracting mean from each data across each dimension.

Stad. data matrix

data points →

$$\begin{bmatrix} (y_{1_1} - \bar{y}_1) & (y_{1_2} - \bar{y}_1) & \cdots & (y_{1_n} - \bar{y}_1) \\ \vdots & \ddots & & \vdots \\ (y_{m_1} - \bar{y}_m) & \cdots & & (y_{m_n} - \bar{y}_m) \end{bmatrix}_{m \times n}$$

② Calculate all possible cov(,)'s

(3) Calculate all possible cov(,) across the dimensions, and form the covariance matrix.

Most computation intensive.

x, y, z

$$C = COV = \begin{pmatrix} cov(Y_1, Y_1) & cov(Y_1, Y_2) & \cdots & cov(Y_1, Y_m) \\ cov(Y_2, Y_1) & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ cov(Y_m, Y_1) & \cdots & \cdots & cov(Y_m, Y_m) \end{pmatrix}_{m \times m}$$

i.e. finding the dependencies among different dimensions

Covariance matrix is always SPD.

(4) Find the Eigen vectors (v_i) and Eigen values (λ_i) of cov matrix

$C = V \Lambda V^T$ (EVD)

Choose components in non-decreasing order of eigen-values

s.t.

$$C = v_1 \lambda_1 v_1^T + v_2 \lambda_2 v_2^T + \dots + v_m \lambda_m v_m^T$$

where $\lambda_1 > \lambda_2 > \dots > \lambda_m$.

(5) Find Featured vectors:

After arranging eigen-vectors according to non-increasing order of eigen-values, one can keep all the eigen-vectors or can discard less significant eigen-vectors (associated with smaller eigen-values) and form the remaining ones a matrix. This matrix (collection of eigen-vectors of the components that are

decided to be kept)

ex! we have

λ_1 λ_2 - - - λ_k - - - λ_m : eigen-values

associated eigen vectors

$$\begin{pmatrix} v_{11} \\ v_{12} \\ \vdots \\ v_{1m} \end{pmatrix} \begin{pmatrix} v_{21} \\ v_{22} \\ \vdots \\ v_{2m} \end{pmatrix} \cdots \begin{pmatrix} v_{k1} \\ v_{k2} \\ \vdots \\ v_{km} \end{pmatrix} \cdots \begin{pmatrix} v_{m1} \\ v_{m2} \\ \vdots \\ v_{mm} \end{pmatrix}$$

i.e. v_1 v_2 v_k v_m

as $\lambda_1 > \lambda_2 > \cdots > \lambda_k > \cdots > \lambda_m$, we can keep upto certain eigen-vector, say: 'k'.

Then, the Featured Vector/matrix

$$= \begin{pmatrix} v_{11} & v_{21} & \cdots & v_{k1} \\ v_{12} & v_{22} & & v_{k2} \\ \vdots & \vdots & & \vdots \\ v_{1m} & v_{2m} & & v_{km} \end{pmatrix}_{m \times k}$$

as $k < m$

this is the steps of dimensionality reduction.

(6) Resulting data:

$$= \text{Final/RecastedDataMatrix(Set)}_{k \times n} = \text{FeaturedVector}^{T}_{k \times m} \times \text{OriginalDataMatrix}_{m \times n}$$

k → data with PCs. (new-dimension)

n → data points.

m → $Y'_{11} \ldots Y'_{1n}$, Y'_2, Y'_k → k

Therefore dimension of data reduced from 'm' to 'k'

NOTE:

* PCA transforms data linearly into new properties that are not correlated with each other. (as eigen-vectors are orthonormals to eachother)
i.e. PCs are uncorrelated as each of them associated with eigen vectors.
* Principal components are constructed in such a manner that the first (few) component(s) accounts (account) for the largest possible variance(s) in the data set.

SPD: SYMMETRIC POSITIVE DEFINITE matrix.

An $n \times n$ symmetric real (entries are real) matrix M is called positive definite.

If $x^T M x > 0$ $\forall$ non-zero x in $\mathbb{R}^n$

$x^T M x > 0 \ \forall \ x \in \mathbb{R}^n \setminus \{0\}$.

Ex:- 3-dimensional (variable/features) data set with 4 observations (samples).

Dimension

$$\begin{matrix} l \\ h \\ b \end{matrix} \begin{bmatrix} 7 & 4 & 6 & 8 \\ 4 & 1 & 3 & 2 \\ 3 & 8 & 7 & 9 \end{bmatrix}$$

observation.

Consider the PCA and reduce upto 2-dimension (l & h).