PRINCIPAL COMPONENT ANALYSIS [PCA] – I

Main objective is dimensionality reduction, i.e. to reduce the number of variables while preserving max info. Some loss of accuracy but leads to major simplifications in data analysis.

Step 1 : Feature Scaling

Important since other features with wider range will dominate.

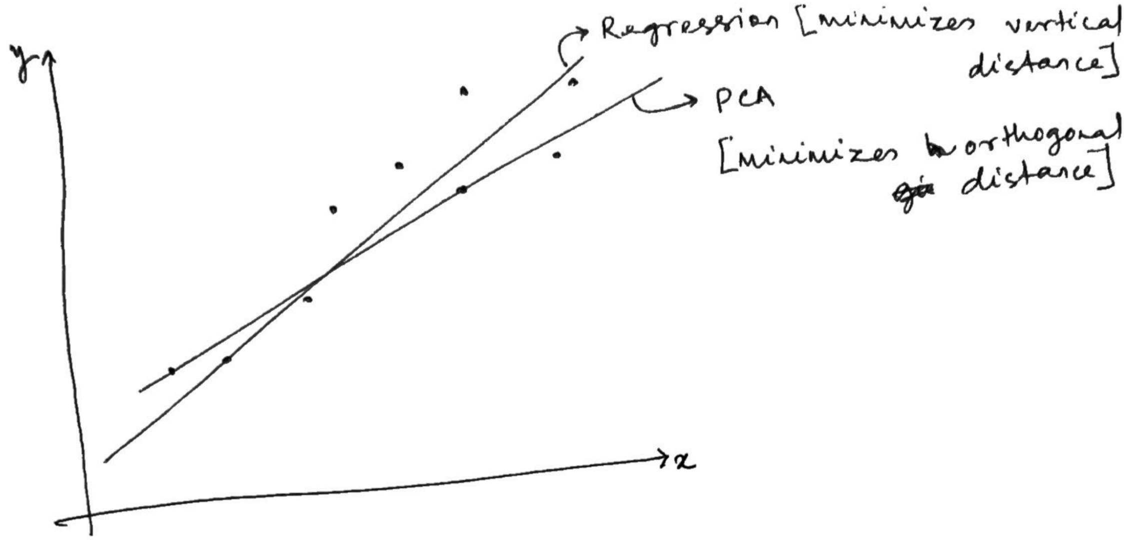
Step 2 : Covariance Matrix Computation

Variables can be correlated.
Covariance helps in finding these relationships

Step 3 : Find eigen values & eigen vectors of covariance matrix

- Eigen vectors are linear combinations of data points & chosen such that they are independent of each other.
- 10-dimensional data will give 10 eigen vectors.
- Represent directions that explain maximum amount of data.
- First principal component accounts for max variance & then the 2nd one & so on.
- Find eigen vectors $\{\vec{v}_i\}$ & eigen values $\{\lambda_i\}$
  Arrange $\{\lambda_i\}$ in decreasing order. $\lambda_i > \lambda_{i+1}$
  So $\vec{v}_1$ corresponds to first component, & so on.
  $\vec{v}_2$ to second component
- Keep a subset of the eigen vectors & discard the rest. This leads to dimensionality reduction since all the data points are now projected onto these eigen vectors through dot product.

Fraction of variance in data explained/captured by $\vec{v}_i$ $= \dfrac{\lambda_i}{\sum_i \lambda_i}$


y
x
Regression [minimizes vertical distance]
PCA [minimizes orthogonal distance]

PCA - II

$\{\vec{x}_n\}$ Data points $n = 1, 2, \ldots, N$
column vectors
of Dimensionality D.
to be projected to dimensions $M < D$.

Let $M = 1$ for simplicity
Need to find direction $\vec{u}_1$ of projected data.
Projected data points $\{\vec{y}_n\} = \{u_1^T x_n\}$

<u>Mean</u>: $\bar{y} = u_1^T \bar{x}$, where $\bar{x} = \frac{1}{N}\sum_n x_n$

<u>Variance</u>: $S_y = \frac{1}{N}\sum_n \{u_1^T x_n - u_1^T \bar{x}\}^2$

$$= \frac{1}{N}\sum_n (u_1^T x_n - u_1^T \bar{x})(u_1^T x_n - u_1^T \bar{x})^T$$

$$= \frac{1}{N}\sum_n (u_1^T x_n - u_1^T \bar{x})(x_n^T u_1 - \bar{x}^T u_1)$$

$$= \frac{1}{N}\sum_n (u_1^T x_n - u_1^T \bar{x})(x_n^T - \bar{x}^T) u_1$$

$$= \frac{1}{N}\sum_n u_1^T (x_n - \bar{x})(x_n^T - \bar{x}^T) u_1$$

$$= u_1^T S_x u_1, \text{ where } S_x = \frac{1}{N}\sum_n (x_n - \bar{x})(x_n - \bar{x})^T$$

<u>Objective</u>: Maximize S_y keeping $\|u_1\| = 1$ fixed.
$\therefore$ find u_1 which maximizes

$$L = u_1^T S_x u_1 + \lambda_1 (1 - u_1^T u_1)$$

λ_1 → Lagrange multiplier.

$\nabla_{u_1} L = 0 \Rightarrow S_x u_1 = \lambda_1 u_1 \Rightarrow u_1$ is the eigen vector of S_x & λ_1 is the eigen-value.

$\Rightarrow S_y = u_1^T S_x u_1 = \lambda_1$

$\Rightarrow$ choose u_1 to be the eigen-vector with max eigen-value so that S_y is maximised.

Same way $u_2, u_3, \ldots$ can be chosen.

If $M = D$, PCA still works & is a rotation of the coordinate axis to maximise variance along the new directions or components.

What if $D >> N$? Very high dimensional data (small dataset of images)

Finding eigen vectors of $D \times D$ matrix has computational cost $\Theta(D^3)$.

Define X : $(N \times D)$ dimensional centred matrix
n^{th} row given by $(x_n - \bar{x})^T$

Covariance Matrix : $S = \frac{1}{N} \underbrace{X^T X}_{\hookrightarrow D \times D \text{ dimensional}}$

$$\frac{1}{N} X^T X u_i = \lambda_i u_i$$

$$\Rightarrow \frac{1}{N} X X^T X u_i = \lambda_i \underbrace{X_i u_i}_{v_i}$$: Multiplying both sides by X

$$\Rightarrow \frac{1}{N} \underbrace{X X^T}_{\hookrightarrow N \times N \text{ Dimensional}} v_i = \lambda_i v_i$$

$\Theta(N^3)$ instead of $\Theta(D^3)$
has the same eigen values as S
[$D - N + 1$ eigen values of S are zero]

One limitation of PCA is that it is limited to linear transformations of data.
Can use kernel PCA to do non-linear transformations.

$$x_n = \begin{pmatrix} x_n^{(1)} \\ x_n^{(2)} \end{pmatrix} \qquad n = 1, 2, \ldots, N$$

$$D = 2$$

To be projected to $M = 1$

$$y_n = u^T x_n \qquad u = \begin{pmatrix} u^{(1)} \\ u^{(2)} \end{pmatrix}$$

<u>y-Variance</u>:
$$S_y = \frac{1}{N} \sum_n (y_n - \bar{y})^2$$
$$= \frac{1}{N} \sum_n \left(u^T x_n - u^T \bar{x} \right)^2$$
$$= \frac{1}{N} \sum_n \left(u^T x_n - u^T \bar{x} \right) \left(u^T x_n - u^T \bar{x} \right)^T$$
$$= \frac{1}{N} \sum_n \left(u^T x_n - u^T \bar{x} \right) \left(x_n^T u - \bar{x}^T u \right)$$
$$= \frac{1}{N} \sum_n \left(u^T x_n x_n^T u - u^T x_n \bar{x}^T u - u^T \bar{x} x_n^T u - u^T \bar{x} \bar{x}^T u \right)$$
$$= \frac{1}{N} \sum_n u^T (x_n - \bar{x})(x_n - \bar{x})^T u$$
$$= u^T \underbrace{S_x}_{2 \times 2} u$$

Maximize S_y while keeping $\|u\| = 1$, constant.

$$\Rightarrow S_x u = \lambda u$$
$$\Rightarrow S_y = u^T \lambda u = \underline{\underline{\lambda}}.$$

for $D > 2$, $M > 2$, choose eigen vectors corresponding to max eigen-value of the co-variance matrix.