Principal Component Analysis

Hee-il Hahn

Professor

Department of Information and Communications Engineering

Hankuk University of Foreign Studies

hihahn@hufs.ac.kr

PCA

- The most popular algorithm for dimensionality reduction
  - Finds the direction vectors along which the data has maximum variance.
  - Most useful in the case when the data lies on or close to a linear subspace of the data set.
  - Suppose that $\boldsymbol{X} \in \boldsymbol{R}^{n \times p}$ is a matrix whose rows are p −dimensional data points.
  - We are looking for the d −dimensional linear subspace of $\boldsymbol{R}^{p}$ along which the data has maximum variance.
  - The objective function it optimizes is

$$\max_{\boldsymbol{V}} var(\boldsymbol{XV})$$

where $\boldsymbol{V}$ is an orthogonal $\mathrm{p} \times d$ matrix.

  - If $d = 1$, then $\boldsymbol{V}$ is simply a unit-length vector which gives the direction of maximum variance.

Deriving the vector of maximum variance

$$\max_{\|\boldsymbol{v}\|=1} var\,(\boldsymbol{Xv}) = \max_{\|\boldsymbol{v}\|=1} E(\boldsymbol{Xv})^2 - \left(E(\boldsymbol{Xv})\right)^2$$

$$= \max_{\|\boldsymbol{v}\|=1} E(\boldsymbol{Xv})^2 \qquad \text{if mean-centered}$$

$$= \max_{\|\boldsymbol{v}\|=1} \textstyle\sum_{i=1}^{n}(\boldsymbol{x}_i\boldsymbol{v})^2$$

$$= \max_{\|\boldsymbol{v}\|=1} \textstyle\sum_{i=1}^{n}(\boldsymbol{x}_i\boldsymbol{v})^T\,\boldsymbol{x}_i\boldsymbol{v}$$

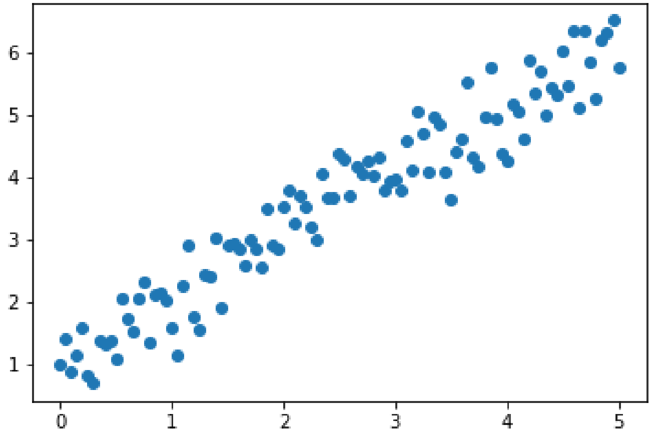
$$= \max_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^T\left(\textstyle\sum_{i=1}^{n}\boldsymbol{x}_i{}^T\boldsymbol{x}_i\right)\boldsymbol{v}$$

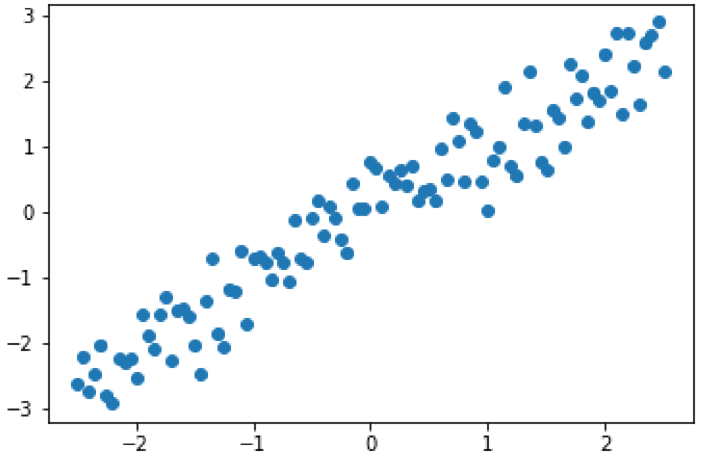
$$= \max_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^T\boldsymbol{X}^T\,\boldsymbol{Xv}$$

$$= \max_{\|\boldsymbol{v}\|=1} \boldsymbol{v}^T\lambda\,\boldsymbol{v} = \lambda_{max}$$

PCA

- Mean-centered data





Principal component analysis (another ver.)

- Suppose a collection of m points $\{\boldsymbol{x}^{(1)} \ \ldots \ \boldsymbol{x}^{(m)}\}$ in $\boldsymbol{R}^n$
  - lossy compression assumed.
  - for each point $\boldsymbol{x}^{(k)}$, corresponding code vector $\boldsymbol{c}^{(k)} \in \boldsymbol{R}^l$, $l \leq n$
  - Find some encoding function $f(\boldsymbol{x}) = \boldsymbol{c}$ and decoding function $\boldsymbol{x} \approx g\big(f(\boldsymbol{x})\big)$
  - Let $g(\boldsymbol{c}) = \boldsymbol{D}\boldsymbol{c}$ where $\boldsymbol{D} \in \boldsymbol{R}^{n \times l}$

    constraints: columns of $\boldsymbol{D}$ are orthonormal to each other.
  - For optimal code ⇨ minimize the distance between $\boldsymbol{x}$ and $g(\boldsymbol{c})$
  - i.e.

$$
\begin{aligned}
\boldsymbol{c}^* &= \underset{\boldsymbol{c}}{\operatorname{argmin}} \|\boldsymbol{x} - g(\boldsymbol{c})\|_2 \\
&= \underset{\boldsymbol{c}}{\operatorname{argmin}} \big(\boldsymbol{x} - g(\boldsymbol{c})\big)^T \big(\boldsymbol{x} - g(\boldsymbol{c})\big) \\
&= \underset{\boldsymbol{c}}{\operatorname{argmin}} \big(\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{x}^T g(\boldsymbol{c}) - g(\boldsymbol{c})^T \boldsymbol{x} + g(\boldsymbol{c})^T g(\boldsymbol{c})\big) \\
&= \underset{\boldsymbol{c}}{\operatorname{argmin}} \big(-2\boldsymbol{x}^T g(\boldsymbol{c}) + g(\boldsymbol{c})^T g(\boldsymbol{c})\big)
\end{aligned}
$$

Principal component analysis

$$\begin{aligned}
\boldsymbol{c}^* &= \underset{\boldsymbol{c}}{\operatorname{argmin}} \|\boldsymbol{x} - g(\boldsymbol{c})\|_2 \\
&= \underset{\boldsymbol{c}}{\operatorname{argmin}} \big(\boldsymbol{x} - g(\boldsymbol{c})\big)^T \big(\boldsymbol{x} - g(\boldsymbol{c})\big) \\
&= \underset{\boldsymbol{c}}{\operatorname{argmin}} \big(\boldsymbol{x}^T\boldsymbol{x} - \boldsymbol{x}^T g(\boldsymbol{c}) - g(\boldsymbol{c})^T\boldsymbol{x} + g(\boldsymbol{c})^T g(\boldsymbol{c})\big) \\
&= \underset{\boldsymbol{c}}{\operatorname{argmin}} \big(-2\boldsymbol{x}^T g(\boldsymbol{c}) + g(\boldsymbol{c})^T g(\boldsymbol{c})\big) \\
&= \underset{\boldsymbol{c}}{\operatorname{argmin}} (-2\boldsymbol{x}^T\boldsymbol{D}\boldsymbol{c} + \boldsymbol{c}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{c}) \\
&= \underset{\boldsymbol{c}}{\operatorname{argmin}} (-2\boldsymbol{x}^T\boldsymbol{D}\boldsymbol{c} + \boldsymbol{c}^T\boldsymbol{c}) \qquad (\boldsymbol{D}^T\boldsymbol{D} = \boldsymbol{I})
\end{aligned}$$

$\therefore\ \boldsymbol{c} = \boldsymbol{D}^T\boldsymbol{x}$ ⇨ $f(\boldsymbol{x}) = \boldsymbol{D}^T\boldsymbol{x}$ and

$r(\boldsymbol{x}) = g(f(\boldsymbol{x})) = \boldsymbol{D}\boldsymbol{D}^T\boldsymbol{x}$

Principal component analysis

- How to choose the encoding matrix $\boldsymbol{D} \in \boldsymbol{R}^{n \times l}$

$$\boldsymbol{D}^* = \underset{\boldsymbol{D}}{\operatorname{argmin}} \sqrt{\sum_{i,j} \left(x_j^{(i)} - r(\boldsymbol{x}^{(i)})_j \right)^2} \quad \text{subject to } \boldsymbol{D}^T \boldsymbol{D} = \boldsymbol{I}$$

Consider the case of $l = 1$.

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \sum_i \left\| \boldsymbol{x}^{(i)} - \boldsymbol{d}\boldsymbol{d}^T \boldsymbol{x}^{(i)} \right\|_2^2 \quad \text{subject to } \|\boldsymbol{d}\|_2 = 1$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmin}} \sum_i \left\| \boldsymbol{x}^{(i)} - \boldsymbol{d}^T \boldsymbol{x}^{(i)} \boldsymbol{d} \right\|_2^2 \quad \text{subject to } \|\boldsymbol{d}\|_2 = 1$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmin}} \sum_i \left\| \boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i)^T} \boldsymbol{d}\boldsymbol{d} \right\|_2^2 \quad \text{subject to } \|\boldsymbol{d}\|_2 = 1$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmin}} \|X - X\boldsymbol{d}\boldsymbol{d}^T\|_F^2 \quad \text{subject to } \boldsymbol{d}^T \boldsymbol{d} = 1$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmin}}\, Tr\left((X - X\boldsymbol{d}\boldsymbol{d}^T)^T (X - X\boldsymbol{d}\boldsymbol{d}^T) \right) \quad \text{subject to } \|\boldsymbol{d}\|_2 = 1$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmin}} \left(Tr(\boldsymbol{X}^T \boldsymbol{X} - \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{d}\boldsymbol{d}^T - \boldsymbol{d}\boldsymbol{d}^T \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{d}\boldsymbol{d}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{d}\boldsymbol{d}^T) \right)$$

Principal component analysis

- How to choose the encoding matrix $\boldsymbol{D} \in \boldsymbol{R}^{n \times l}$

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \sum_i \left\| \boldsymbol{x}^{(i)} - \boldsymbol{d}\boldsymbol{d}^T \boldsymbol{x}^{(i)} \right\|_2^2 \quad \text{subject to } \|\boldsymbol{d}\|_2 = 1 \quad (= \ \boldsymbol{d}^T\boldsymbol{d} = 1\ )$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmin}} \left(-2Tr(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^T) + Tr(\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^T) \right)$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmin}} \left(-2Tr(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^T) + Tr(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{d}\boldsymbol{d}^T) \right) \quad \text{subject to } \boldsymbol{d}^T\boldsymbol{d} = 1$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmin}} \left(-2Tr(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^T) + Tr(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^T) \right) \quad \text{subject to } \boldsymbol{d}^T\boldsymbol{d} = 1$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmax}} \left(Tr(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{d}\boldsymbol{d}^T) \right) \quad \text{subject to } \boldsymbol{d}^T\boldsymbol{d} = 1$$

$$= \underset{\boldsymbol{d}}{\operatorname{argmax}} \left(Tr(\boldsymbol{d}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{d}) \right) \quad \text{subject to } \boldsymbol{d}^T\boldsymbol{d} = 1$$

⇨ optimal $\boldsymbol{d} = \boldsymbol{d}^* =$ eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$ corresponding to the largest eigenvalue