Principal Component Analysis

Nirdosh Bhatnagar

Introduction

Principal component analysis (PCA) is an unsupervised learning technique to extract 'relevant' data from high-dimensional spaces.

Goal

Given a set $\mathcal{X}$ of n data points each with $t \in \mathbb{P} \backslash \{1\}$ attributes, PCA finds a representation of the data points in a space of dimension $k \in \mathbb{P}$, where $k \leq t$.

List of Commonly Used Symbols

E_{err} = the mean-squared error in the approximation of a data point
i = indexing variable, where $1 \leq i \leq n$
I = is an identity matrix of size t
j = indexing variable, where $1 \leq j \leq t$
k = number of largest eignevalues used in the approximation of a data point, $k \leq t$
n = number of data points, where $n \geq 2$
t = size of the data vector, where $t \in \mathbb{P}$
w_i = column vector of size t
$W = t \times n$ matrix
x_i = ith real-valued data vector of size t
$\overline{x}$ = average value of the data points
$\mathcal{X} = \{x_i \mid 1 \leq i \leq n\}$ is the set of data points

$\widetilde{\Sigma} = t \times t$ covariance matrix
θ_i = column vector of size t
$\widehat{\theta}_i$ = column vector of size t
$\Theta = t \times n$ matrix
λ_j = jth eigenvalue of matrix $\widetilde{\Sigma}$
Λ = diagonal matrix of size t
τ_{thresh} = the threshold value used in the selection of k
ψ_j = column vector of size t
$\Psi = t \times t$ matrix

Model

- Let $\mathcal{X} = \{x_i \mid x_i \in \mathbb{R}^t,\ 1 \leq i \leq n\}$ be the set of $n \in \mathbb{P} \backslash \{1\}$ data points.
- The data point x_i is represented as a column vector of size t, where $1 \leq i \leq n$.
- Let the average value of the data points be

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Define a column vector

$$\theta_i = (x_i - \overline{x}), \quad \text{for} \quad 1 \le i \le n$$

  and a $t \times n$ matrix Θ as

$$\Theta = \left[\begin{array}{cccc} \theta_1 & \theta_2 & \cdots & \theta_n \end{array} \right]$$

- The covariance matrix $\widetilde{\Sigma}$ of the data points is

$$\widetilde{\Sigma} = \frac{1}{(n-1)} \Theta\Theta^T = \frac{1}{(n-1)} \sum_{i=1}^{n} \theta_i \theta_i^T$$

- The $t \times t$ covariance matrix $\widetilde{\Sigma}$ is symmetric and positive semidefinite.
- Therefore $\widetilde{\Sigma} = \Psi\Lambda\Psi^T$, where Λ is a diagonal matrix with eigenvalues of the matrix $\widetilde{\Sigma}$ on its main diagonal.
- As the matrix $\widetilde{\Sigma}$ is symmetric and positive semidefinite, its eigenvalues are nonnegative.
- Let the eigenvalues of the matrix $\widetilde{\Sigma}$ be $\lambda_j \in \mathbb{R}_0^+, 1 \le j \le t$.
- The columns of the matrix Ψ are the mutually orthogonal eigenvectors of the covariance matrix $\widetilde{\Sigma}$.
- Assume that these eigenvectors are orthonormal. Therefore $\Psi\Psi^T = I$, where I is an identity matrix of size t. Thus

$$\begin{aligned} \Lambda &= \Psi^T \widetilde{\Sigma} \Psi \\ &= \frac{1}{(n-1)} \Psi^T \Theta\Theta^T \Psi \\ &= \frac{1}{(n-1)} W W^T \end{aligned}$$

  where $W = \Psi^T \Theta$ is a $t \times n$ matrix. Therefore

$$\Theta = \Psi W$$

- Let

$$\begin{aligned} W &= \left[\begin{array}{cccc} w_1 & w_2 & \cdots & w_n \end{array} \right]; \quad w_i \in \mathbb{R}^t \ \ 1 \le i \le n \\ w_i &= \left[\begin{array}{cccc} w_{i1} & w_{i2} & \cdots & w_{it} \end{array} \right]^T; \quad w_{ij} \in \mathbb{R} \ 1 \le j \le t \\ \Psi &= \left[\begin{array}{cccc} \psi_1 & \psi_2 & \cdots & \psi_t \end{array} \right]; \quad \psi_j \in \mathbb{R}^t \ 1 \le j \le t \end{aligned}$$

  Thus

$$\theta_i = \Psi w_i = \sum_{j=1}^{t} \psi_j w_{ij}, \quad 1 \le i \le n$$

  The vector θ_i is approximated in PCA as

$$\widehat{\theta}_i = \sum_{j=1}^{k} \psi_{l_j} w_{il_j}, \quad \text{for } 1 \le i \le n$$

  where $k \le t$.

– Therefore, the mean-squared error E_{err} in the approximation of a data point is

$$E_{err} = \frac{1}{(n-1)} \sum_{i=1}^{n} \left\| \theta_i - \widehat{\theta}_i \right\|^2$$

where $\|\cdot\|$ is the Euclidean norm.

– It can be shown that the mean-squared error E_{err} is equal to $\sum_{j=k+1}^{t} \lambda_{l_j}$.

– As the eigenvalues are nonnegative, E_{err} is minimized by selecting the eigenvectors ψ_{l_j}, for $1 \leq j \leq k$, in the approximation $\widehat{\theta}_i$ which correspond to the largest k eigenvalues of the covariance matrix $\widetilde{\Sigma}$.

These eigenvectors which correspond to the largest eigenvalues are called the *principal components* of the matrix Θ.

Further, the PCA projects data along directions in which the data varies most.

The magnitude of the eigenvalues quantify the variation of the data points along the directions of the eigenvectors.

– The value k is typically selected so that $\sum_{j=1}^{k} \lambda_{lj} / \sum_{j=1}^{t} \lambda_j$ is greater than a threshold τ_{thresh}, where $\tau_{thresh} \in (0, 1]$.

– Finally, note that PCA is closely related to the singular value decomposition of the matrix $\widetilde{\Sigma}$.