Dimensionality Reduction

Yin-Yun Li

December 24, 2025

1 Principal Component Analysis

Suppose we have a dataset with n observations and k variables, represented as a matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}.$$

For convenience, we usually center the data by subtracting the sample mean across individual for each variable. Let each column vector $\mathbf{x}_i \in \mathbb{R}^k$ be i-th observation with k dimensions, we can also write X as:

$$\begin{bmatrix} — & \mathbf{x}_1^\top & — \\ — & \mathbf{x}_2^\top & — \\ & \vdots & \\ — & \mathbf{x}_n^\top & — \end{bmatrix}.$$

Intuitively, PCA aims to find a new coordinate system such that the first direction (principal component) caputures the largest variance, and the second one captures the maximum remaining variance unrelated to the previous one, and so forth.

Let the first new variable $z_{i1} = \sum_{p=1}^{k} \omega_{p1} x_{ip}$ for each $i = 1, 2, \cdots n$, or equivalently in matrix form:

$$\underbrace{\begin{bmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{bmatrix}}_{\mathbf{z}_1} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \omega_{11} \\ \omega_{21} \\ \vdots \\ \omega_{k1} \end{bmatrix}}_{\boldsymbol{\omega}_1}$$

Our goal is to choose $\boldsymbol{\omega}_1$ such that the variance of $\mathbf{z}_1$ is maximized, subject to the constraint that $\boldsymbol{\omega}_1$ is a unit vector, so we can write down the maximization problem as:

$$\begin{aligned} \max_{\boldsymbol{\omega}_1} \quad & \text{Var}(\mathbf{z}_1) = \frac{1}{n}(X\boldsymbol{\omega}_1)^\top X\boldsymbol{\omega}_1 \\ \text{subject to} \quad & \boldsymbol{\omega}_1^\top \boldsymbol{\omega}_1 = \|\boldsymbol{\omega}_1\| = 1, \end{aligned} \tag{1}$$

where

$$X\boldsymbol{\omega}_1 = \begin{bmatrix} \mathbf{x}_1^\top \boldsymbol{\omega_1} \\ \mathbf{x}_2^\top \boldsymbol{\omega_1} \\ \vdots \\ \mathbf{x}_n^\top \boldsymbol{\omega_1} \end{bmatrix}, \quad (X\boldsymbol{\omega}_1)^\top X\boldsymbol{\omega}_1 = \sum_{i=1}^{n} (\mathbf{x}_i^\top \boldsymbol{\omega}_1)^2 = \sum_{i=1}^{n} (\mathbf{x}_i^\top \boldsymbol{\omega}_1)^\top (\mathbf{x}_i^\top \boldsymbol{\omega}_1) = \boldsymbol{\omega}_1^\top \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) \boldsymbol{\omega}_1.$$

Note that we first focus on the sum of squares of $\mathbf{z}_1$. The Lagrangian function is given by:

$$\mathcal{L}(\boldsymbol{\omega}_1, \lambda_1) = \boldsymbol{\omega}_1^\top X^\top X\boldsymbol{\omega}_1 - \lambda_1(\boldsymbol{\omega}_1^\top \boldsymbol{\omega}_1 - 1) \tag{2}$$

The First-order condition yields:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\omega}_1} &= 2X^\top X\boldsymbol{\omega}_1 - 2\lambda_1 \boldsymbol{\omega}_1 = \mathbf{0} \\ X^\top X\boldsymbol{\omega}_1 &= \lambda_1 \boldsymbol{\omega}_1. \end{aligned} \tag{3}$$

The above equation shows that $\boldsymbol{\omega}_1$ is an eigenvector of $X^\top X$ w.r.t. the eigenvalue λ_1.

Next, we try to find the second principal component $\mathbf{z}_2 = X\boldsymbol{\omega}_2$, which is uncorrelated with the first one. Thus, we need to solve the following maximization problem:

$$\begin{aligned} \max_{\boldsymbol{\omega}_2} \quad & \text{Var}(\mathbf{z}_2) = \frac{1}{n}(X\boldsymbol{\omega}_2)^\top X\boldsymbol{\omega}_2 \\ \text{subject to} \quad & \boldsymbol{\omega}_2^\top \boldsymbol{\omega}_2 = \|\boldsymbol{\omega}_2\| = 1; \quad \boldsymbol{\omega}_1^\top \boldsymbol{\omega}_2 = 0. \end{aligned} \tag{4}$$

Apply the Lagrangian method again:

$$\mathcal{L}(\boldsymbol{\omega}_2, \lambda_2, \gamma) = \boldsymbol{\omega}_2^\top X^\top X\boldsymbol{\omega}_2 - \lambda_2(\boldsymbol{\omega}_2^\top \boldsymbol{\omega}_2 - 1) - \gamma(\boldsymbol{\omega}_1^\top \boldsymbol{\omega}_2). \tag{5}$$

The First-order condition yields:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\omega}_2} = 2X^\top X\boldsymbol{\omega}_2 - 2\lambda_2 \boldsymbol{\omega}_2 - \gamma \boldsymbol{\omega}_1 = \mathbf{0} \tag{6}$$

Multiplying both sides of Equation (6) by $\boldsymbol{\omega}_1^\top$ gives:

$$2\boldsymbol{\omega}_1^\top X^\top X\boldsymbol{\omega}_2 - \gamma\boldsymbol{\omega}_1^\top\boldsymbol{\omega}_1 = \mathbf{0}. \tag{7}$$

Note that $\boldsymbol{\omega}_1^\top X^\top X = \lambda_1\boldsymbol{\omega}_1^\top$ from (3), and this forces $\gamma = 0$ and hence

$$X^\top X\boldsymbol{\omega}_2 = \lambda_2\boldsymbol{\omega}_2. \tag{8}$$

Now we construct the matrix of eigenvectors as $\Omega = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \cdots \boldsymbol{\omega}_k]$. Inductively, we can regard the system of equations as $Z = X\Omega$, where $Z = [\mathbf{z}_1, \mathbf{z}_2, \cdots \mathbf{z}_k]$. Since $X^\top X$ is a symmetric matrix, there exists an invertible matrix P such that: $P^{-1}X^\top XP$ is diagonal, i.e., $X^\top X$ is diagonalizable[1].

Moreover, the diagonal matrix takes the form of:

$$Z^\top Z = \Omega^\top X^\top X\Omega = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix},$$

which shows that $P = \Omega$ is an orthonormal matrix and $tr(Z^\top Z) = tr(X^\top X) = \sum_{i=1}^k \lambda_i$. For i-th principal component, the contribution to total variance is given by the proportion $\frac{\lambda_i}{\sum_{i=1}^k \lambda_i}$. Suppose, WLOG, $n \geq k$. If $\text{rank}(X) < k$, then some eigenvalues are exactly zero, indicating that those components contribute no variance and can be discarded. If $\text{rank}(X) = k$, we ideally hope that the trailing eigenvalues are sufficiently small, such that their contribution to the total variance is negligible.

[1]For now, I have not learned the spectral theorem, I can only write these problems in terms of diagonalization theorem.