Principal Component Analysis

Qingyang Li

1 Introduction

PCA involves a mathematical procedure that transforms a set of correlated variables into a smaller set of uncorrelated variables called principal components (PCs). These PCs are linear combination of original variables and can be thought as "new" variables.

1.1 Characteristic of Principal Components

- Uncorrelated with each other
- 1st principal component accounts for as much variability in the data as possible
- Each successive principal component accounts for as much variability in the data as possible.

2 PCA with covariance matrix

Let $\mathbf{x} \sim (\mu, \Sigma)$ and $\mathbf{x}$ is $p \times 1$ vector, $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$. Notice that we do not use a multivariate normal distribution assumption.

2.1 First Principal Component

- $Y_1 = \mathbf{a_1^\intercal}(\mathbf{x} - \mu) = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_p - \mu_p \end{bmatrix} = a_{11}(x_1 - \mu_1) + a_{12}(x_2 - \mu_2) + \cdots + a_{1p}(x_p - \mu_p)$, where $\mathbf{a_1}$ is $p \times 1$ vector chosen so that $Var[\mathbf{a_1^\intercal}(\mathbf{x} - \mu)]$ is maximized over all vectors.
- The constraint for $\mathbf{a_1^\intercal}$ is that the length is equal to 1, which is $a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1$.
- The maximum value of the variance is the largest eigenvalue of $\mathbf{\Sigma}$, which is denoted as λ_1. The $\mathbf{a_1}$ itself is the eigenvector corresponding to the largest eigenvalue.
- Since the variance of Y_1 is being maximized, the new variable Y_1 will explain as much variability of $\mathbf{x}$ as possible.

2.2 Second Principal Component

- $Y_2 = \mathbf{a_2^\intercal}(\mathbf{x} - \mu)$, where $\mathbf{a_2}$ is the eigenvector corresponding to the second largest eigenvalue λ_2 from $\mathbf{\Sigma}$.
- Y_1 and Y_2 are uncorrelated, which implies that $\mathbf{a_1}$ and $\mathbf{a_2}$ are orthogonal ($\mathbf{a_1}'\mathbf{a_2} = 0$).

2.3 Total Variance

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

The total variance is then defined as the sum of the diagonal elements of the population covariance matrix $\mathbf{\Sigma}$.

$$tr(\Sigma) = \sum_{i=1}^{p} \sigma_{ii} = \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp}$$

Also notice the trace of the matrix is the sum of the eigenvalues.

$$tr(\Sigma) = \sum_{i=1}^{p} \lambda_i$$

Then a measure of the importance of the j^{th} principal component is $\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$. The larger the value, the more variability the j^{th} principal component accounts for.

2.4 Estimation

Since μ and $\mathbf{\Sigma}$ are usually NEVER known, the corresponding estimates $\hat{\mu}$ and $\widehat{\mathbf{\Sigma}}$ are used instead in calculations. Therefore, we will from now on use

$$\hat{Y}_j = \hat{\mathbf{a}}_\mathbf{j}(\mathbf{x} - \hat{\mu}), j = 1, 2, \cdots, p$$

where the j^{th} largest eigenvalue of $\widehat{\mathbf{\Sigma}}$ and its corresponding eigenvector are denoted as $\hat{\lambda}_j$ and $\hat{\mathbf{a}}_\mathbf{j}$ respectively.

2.5 Possible Issues with PCA

- If original set of variables are already uncorrelated, PCA will not help.
- PCA does not generally eliminate variables, because the principal components are linear combinations of the original variables.
- The original variables need to be measured in the same units and have similar variances. The solution for this issue is to use standardized data. Equivalently, use $\mathbf{R}$ in place of $\widehat{\mathbf{\Sigma}}$, because sample correlation matrix is the sample covariance matrix of standardized random variables.

3 PCA with the correlation matrix

PCA is most often performed using the correlation matrix $\mathbf{P}$ rather than the covariance matrix $\mathbf{\Sigma}$ to eliminate the problem with different numerical scales being used with variables. Again, because $\mathbf{P}$ will not be known, we will use the sample correlation matrix $\mathbf{R}$ instead. The corresponding eigenvalues and eigenvectors are denoted by $\hat{\lambda}_j^*$ and $\hat{\mathbf{a}}_\mathbf{j}^*$, respectively.

3.1 Determine the number of principal components

1. Plot $\hat{\lambda}_1^*, \hat{\lambda}_2^*, \cdots, \hat{\lambda}_p^*$ vs. $1, 2, \cdots, p$. When the point on the plot level off close to 0, the corresponding principal component are probably not contributing too much information to understand the data.

2. Find the number of eigenvalues greater than 1.

3.2 PC scores

For each observation, we calculate the j^{th} principal component value or score based on sample covariance matrix as:

$$\hat{y}_{ij} = \hat{\mathbf{a}}_\mathbf{j}^\intercal(\mathbf{x_i} - \hat{\mu})$$

For each observation, we calculate the j^{th} principal component value or score based on sample correlation matrix as:

$$\hat{y}_{ij}^* = \hat{\mathbf{a}}_\mathbf{j}^{*\intercal}\mathbf{z_i} = \begin{bmatrix} \hat{a}_{j1}^* & \hat{a}_{j2}^* & \cdots & \hat{a}_{jp}^* \end{bmatrix} \times \begin{bmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{ip} \end{bmatrix}$$

where $i = 1, 2, 3, \cdots, n$ represents the number of observations in the data; and $j = 1, 2, 3, \cdots, p$ represents the number of principal components.