Contents

1 PCA Via Eigen-Decomposition

Given a data matrix of n points in $\mathbb{R}^p$ (i.e., a row in the matrix below)

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \dots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

Principal Components Analysis (PCA) finds **linear combinations of original column vectors X that best explain the covariation structure of those columns** $X = (X_1, \dots, X_p)$.

1.1 Step 1: Center Data Matrix

Find the column mean for each k^{th} column in $\mathbf{X}$

$$\mu_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik}$$

Centering the data per column:

$$X_c = X - \begin{pmatrix} \mu_1 & \cdots & \mu_p \\ \vdots & \dots & \vdots \\ \mu_1 & \cdots & \mu_p \end{pmatrix}$$

In some applications, we may also scale the data per column by dividing each element of the column vectors by the standard deviations of those column vectors. In other words, we essentially obtain the standardized z-scores for each k^{th} column of X.

1.2 Step 2: Sample Covariance of Data Matrix

Compute the sample covariance matrix of p column vectors $X = (X_1, \dots, X_p)$:

$$\text{Cov}(X) = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \ddots & & \vdots \\ \vdots & & \ddots & \text{Cov}(X_{p-1}, X_d) \\ \text{Cov}(X_p, X_1) & \cdots & \text{Cov}(X_p, X_{p-1}) & \text{Var}(X_p) \end{pmatrix} \in \mathbb{R}^{p \times p}$$

Computationally, each entry $\text{Cov}(X)_{k\ell} = \text{Cov}(X_k, X_\ell) = \frac{1}{n-1}(\vec{x}_k \cdot \vec{x}_\ell)$. Therefore,

$$\text{Cov}(X) = \frac{1}{n-1} \begin{pmatrix} - & \vec{x}_1 & - \\ & \vdots & \\ - & \vec{x}_p & - \end{pmatrix} \begin{pmatrix} | & & | \\ \vec{x}_1 & \dots & \vec{x}_p \\ | & & | \end{pmatrix} = \frac{1}{n-1} X_c^T X_c$$

Sometimes the correlation matrix is used instead of the covariance matrix. For here one, we will use $\mathbf{\Sigma}$ to denote the covariance matrix $\text{Cov}(X)$.

1.3 Step 3: Principal Components

For each column vectors $X = (X_1, \ldots, X_p)$, we can construct new column vectors $Y_1, ..., Y_p$ by taking p different combinations of the $X = (X_1, \ldots, X_p)$ vectors:

$$\begin{aligned}
\underbrace{Y_1}_{n\times 1} &= \underbrace{X}_{n\times p}\underbrace{\vec{b}_1}_{p\times 1} = b_{11}\underbrace{X_1}_{n\times 1} + b_{21}\underbrace{X_2}_{n\times 1} + \cdots + b_{p1}\underbrace{X_p}_{n\times 1} \\
\underbrace{Y_2}_{n\times 1} &= \underbrace{X}_{n\times p}\underbrace{\vec{b}_2}_{p\times 1} = b_{12}\underbrace{X_1}_{n\times 1} + b_{22}\underbrace{X_2}_{n\times 1} + \cdots + b_{p2}\underbrace{X_p}_{n\times 1} \\
&\vdots \\
\underbrace{Y_p}_{n\times 1} &= \underbrace{X}_{n\times p}\underbrace{\vec{b}_p}_{p\times 1} = b_{1p}\underbrace{X_1}_{n\times 1} + b_{2p}\underbrace{X_2}_{n\times 1} + \cdots + b_{pp}\underbrace{X_p}_{n\times 1}
\end{aligned}$$

Definition 1.1. The derived vectors $Y_k \in \mathbb{R}^{n\times 1}$ for $k = 1, ..., p$ are called the **principal components** or **factors**. Together, these p vectors form a matrix $Y \in \mathbb{R}^{n\times p}$ that has the same dimensions as the original data matrix $X \in \mathbb{R}^{n\times p}$.

Definition 1.2. The vectors $\vec{b}_k \in \mathbb{R}^{p\times 1}$ for $k = 1, ..., p$ are called the **loadings** or **factor loadings** of the k^{th} principal component $Y_k \in \mathbb{R}^{n\times 1}$ for $k = 1, ..., p$.

Theorem 1. The **sample means** of the principal components are:

$$\bar{Y_k} = \frac{1}{n}\sum_{i=1}^{n} Y_{ik} = \underbrace{\vec{b_k}^T}_{1\times p}\underbrace{\bar{X}}_{p\times 1}$$

where $\bar{X} \in \mathbb{R}^{p\times 1} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \vdots \\ \bar{X}_p \end{bmatrix}$.

Theorem 2. The **sample variances** of the principal components are:

$$\begin{aligned}
\text{var}(Y_k) &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ik} - \bar{Y_k})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\underbrace{\vec{b_k}^T}_{1\times p}\underbrace{X_i}_{p\times 1} - \underbrace{\vec{b_k}^T}_{1\times p}\underbrace{\bar{X}}_{p\times 1})(\underbrace{\vec{b_k}^T}_{1\times p}\underbrace{X_i}_{p\times 1} - \underbrace{\vec{b_k}^T}_{1\times p}\underbrace{\bar{X}}_{p\times 1})^T \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\underbrace{\vec{b_k}^T}_{1\times p}\underbrace{(X_i - \bar{X})(X_i - \bar{X})^T}_{\mathbf{\Sigma}}\underbrace{\vec{b}_k}_{p\times 1} \\
&= \vec{b_k}^T\mathbf{\Sigma}\vec{b}_k
\end{aligned}$$

where

- Y_{ik} is the i^{th} element of the k^{th} component $Y_k \in \mathbb{R}^{n\times 1}$ (a scalar element)
- $X_i \in \mathbb{R}^{p\times 1}$ is the column vector $X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ X_{i3} \\ \vdots \\ X_{ip} \end{bmatrix}$ for $i \in 1, ..., n$

Theorem 3. The **sample covariances** between Y_k and Y_ℓ are given by:

$$\begin{aligned} \text{Cov}(Y_k, Y_\ell) &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ik} - \bar{Y_k})(Y_{i\ell} - \bar{Y_\ell}) \\ &= \vec{b}_k^T \mathbf{\Sigma} \vec{b}_\ell \end{aligned}$$

1.4 Step 4: Solution For PCA

The principal components are the **uncorrelated** linear combinations $Y_1, ..., Y_p$ whose sample variances are as large as possible. That is, we choose up to p loadings $\vec{b}_k$ such that:

$$\begin{aligned} \text{First principal component} \quad & \vec{b}_1 = \underset{\|\vec{b}_1\|=1}{\text{argmax}}\{\vec{b}_1^T \mathbf{\Sigma} \vec{b}_1\} \\ \text{Second principal component} \quad & \vec{b}_2 = \underset{\|\vec{b}_2\|=1}{\text{argmax}}\left\{\vec{b}_2^T \mathbf{\Sigma} \vec{b}_2\right\} \quad \text{subject to} \quad \vec{b}_1^T \mathbf{\Sigma} \vec{b}_2 = 0 \\ & \vdots \\ \text{Subsequent principal component} \quad & \vec{b}_\ell = \underset{\|\vec{b}_\ell\|=1}{\text{argmax}}\left\{\vec{b}_\ell^T \mathbf{\Sigma} \vec{b}_\ell\right\} \quad \text{subject to} \quad \vec{b}_k^T \mathbf{\Sigma} \vec{b}_\ell = 0 \quad \forall \quad k < \ell \end{aligned}$$

We constrain the **loadings** so that their sums of squares are equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance. The standard technique for maximizing the variances is to use the Lagrange multiplier.

1.4.1 First Principal Component

For the **first principal component**, we maximize:

$$\vec{b}_1^T \mathbf{\Sigma} \vec{b}_1 - \lambda\left(\vec{b}_1^T \vec{b}_1 - 1\right)$$

where λ is a Lagrange multiplier. Differentiation with respect to $\vec{b}_1$ gives

$$\mathbf{\Sigma}\vec{b}_1 - \lambda\vec{b}_1 = \mathbf{0},$$

or

$$\left(\mathbf{\Sigma} - \lambda \mathbf{I}_p\right)\vec{b}_1 = \mathbf{0},$$

where $\mathbf{I}_p$ is the $(p \times p)$ identity matrix since $\mathbf{\Sigma}$ is $(p \times p)$. Thus, λ **is an eigenvalue of** $\mathbf{\Sigma}$ **and** $\vec{b}_1$ **is the corresponding eigenvector**. To decide which of the p eigenvectors gives $\underbrace{Y_1}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\vec{b}_1}_{p\times 1}$ with maximum variance, note that the quantity to be maximized is

$$\vec{b}_1^T \mathbf{\Sigma} \vec{b}_1 = \vec{b}_1^T \lambda \vec{b}_1 = \lambda \vec{b}_1^T \vec{b}_1 = \lambda 1 = \lambda,$$

so λ must be as large as possible. Thus, $\vec{b}_1$ **is the eigenvector corresponding to the largest eigenvalue of the covariance matrix of** X**, i.e.,** $\mathbf{\Sigma}$. It then follows that $\text{var}\,(Y_1) = \vec{b}_1^T \mathbf{\Sigma} \vec{b}_1 = \lambda_1$, which is the largest eigenvalue.

1.4.2 Second Principal Component

For the **second principal component**, the maximization problem is subject to $\vec{b}_1^T \mathbf{\Sigma} \vec{b}_2 = 0$. The following matrix equations are equivalent:

$$\vec{b}_1^T \mathbf{\Sigma} \vec{b}_2 = \vec{b}_2^T \mathbf{\Sigma} \vec{b}_1 = \vec{b}_2^T \lambda_1 \vec{b}_1^T = \lambda_1 \vec{b}_2^T \vec{b}_1 = \lambda_1 \vec{b}_1^T \vec{b}_2$$

Note that from the first principal component, we found that $\lambda_1 = \mathbf{\Sigma}$, which is used in the substitution above. Therefore, any of the equations (after dividing through by λ_1 on all sides)

$$\begin{aligned} \vec{b}_1^T \mathbf{\Sigma} \vec{b}_2 = 0, \quad \vec{b}_2^T \mathbf{\Sigma} \vec{b}_1 = 0 \\ \vec{b}_1^T \vec{b}_2 = 0, \qquad \vec{b}_2^T \vec{b}_1 = 0 \end{aligned}$$

could be used to specify zero correlation between Y_1 and Y_2. Choosing the last of these (an arbitrary choice), and noting that a normalization constraint ($\left\|\vec{b}_2\right\| = 1$) is again necessary, the quantity to be maximized is

$$\vec{b}_2^T \mathbf{\Sigma} \vec{b}_2 - \lambda \left(\vec{b}_2^T \vec{b}_2 - 1 \right) - \phi \vec{b}_2^T \vec{b}_1,$$

where λ, ϕ are Lagrange multipliers. Differentiation with respect to $\vec{b}_2$ gives

$$\mathbf{\Sigma} \vec{b}_2 - \lambda \vec{b}_2 - \phi \vec{b}_1 = \mathbf{0} \tag{1}$$

and multiplication of this equation on the left by $\vec{b}_1^T$ gives

$$\vec{b}_1^T \mathbf{\Sigma} \vec{b}_2 - \lambda \vec{b}_1^T \vec{b}_2 - \phi \vec{b}_1^T \vec{b}_1 = 0,$$

Because the first two terms (both are the covariance between Y_1 and Y_2) are zero and $\vec{b}_1^T \vec{b}_1 = 1$, the equation above reduces to $-\phi = 0$ (so $\phi = \frac{0}{-1} = 0$). Therefore:

$$\begin{aligned} \mathbf{\Sigma} \vec{b}_2 - \lambda \vec{b}_2 - \phi \vec{b}_1 &= \mathbf{0} \\ \mathbf{\Sigma} \vec{b}_2 - \lambda \vec{b}_2 - (0) \vec{b}_1 &= \mathbf{0} \\ \mathbf{\Sigma} \vec{b}_2 - \lambda \vec{b}_2 &= \mathbf{0} \\ \left(\mathbf{\Sigma} - \lambda \mathbf{I}_p \right) \vec{b}_2 &= \mathbf{0} \end{aligned}$$

And λ is once again an eigenvalue of $\mathbf{\Sigma}$, and $\vec{b}_2$ the corresponding eigenvector.

$$\vec{b}_2^T \mathbf{\Sigma} \vec{b}_2 = \vec{b}_2^T \lambda \vec{b}_2 = \lambda \vec{b}_2^T \vec{b}_2 = \lambda 1 = \lambda,$$

so λ is the quantity to be as large as possible. Assuming that $\mathbf{\Sigma}$ does not have repeated eigenvalues, λ cannot equal λ_1. If it did, it follows that $\vec{b}_2 = \vec{b}_1$, violating the constraint $\vec{b}_1^T \vec{b}_2 = 0$ (may need to apply PCA via SVD rather than Eigen-decomposition in that case). **Hence λ should be the second largest eigenvalue of $\mathbf{\Sigma}$, and $\vec{b}_2$ is the corresponding eigenvector**.

1.4.3 The k^{th} Principal Component

In general, the third, fourth, $\ldots, k^{th}, \ldots, p^{th}$ principal components $Y_3, Y_4, \ldots, Y_k, \ldots, Y_p$ are the vectors of coefficients $\vec{b}_3, \vec{b}_4, \ldots, \vec{b}_k, \ldots, \vec{b}_p$, which are the eigenvectors of $\mathbf{\Sigma}$ corresponding to $\lambda_3, \lambda_4, \ldots, \lambda_k, \ldots, \lambda_p$. These λ_k for $k \in 1, \ldots, p$ are the third and fourth largest, ..., down to the smallest eigenvalue, respectively. Furthermore,

$$\text{var}\left(\vec{b}_k^T \mathbf{x}\right) = \lambda_k \quad \text{for } k = 1, 2, \ldots, p$$

1.5 Step 5: Total Variance and Variance Explained

Note that the eigen-decomposition on the the covariance matrix of X is given by:

$$\begin{aligned} \mathbf{\Sigma} &= QDQ^T \\ &= Q \begin{pmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_p \end{pmatrix} Q^T \end{aligned}$$

where

- D is the matrix whose diagonal elements are the eigenvalues associated with the eigenvectors in Q, sorted in descending order
- Q are the eigenvectors of $\mathbf{\Sigma}$

Therefore, total variance in the data X is

$$\text{Var}(X_1) + \cdots + \text{Var}(X_p) = \text{tr}(\text{Cov}(X)) = \lambda_1 + \cdots + \lambda_p$$

And

$$\lambda_k = \left(\text{variance along } k^{th} \text{ principal component direction } \vec{b}_k\right) \propto \ \text{k}^{th} \text{ axis length of the data ellipsoid}$$

$A \propto B$ means that A is directly proportional to B or that $A = kB$ for some constant k. The percent of variance explained by the first k principal components (out of all p) is

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p} \in (0, 1)$$