



Published on STAT 505 (<https://onlinecourses.science.psu.edu/stat505>)

[Home](#) > 11.1 - Principal Component Analysis (PCA) Procedure

11.1 - Principal Component Analysis (PCA) Procedure

Suppose that we have a random vector \mathbf{X} .

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

with population variance-covariance matrix

$$\text{var}(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}$$

Consider the linear combinations

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \dots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \dots + e_{2p}X_p \\ &\vdots \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \dots + e_{pp}X_p \end{aligned}$$

Each of these can be thought of as a linear regression, predicting Y_i from X_1, X_2, \dots, X_p . There is no intercept, but $e_{i1}, e_{i2}, \dots, e_{ip}$ can be viewed as regression coefficients.

Note that Y_i is a function of our random data, and so is also random. Therefore it has a population variance

$$\text{var}(Y_i) = \sum_{k=1}^p \sum_{l=1}^p e_{ik}e_{il}\sigma_{kl} = \mathbf{e}_i' \Sigma \mathbf{e}_i$$

Moreover, Y_i and Y_j have population covariance

$$\text{cov}(Y_i, Y_j) = \sum_{k=1}^p \sum_{l=1}^p e_{ik} e_{jl} \sigma_{kl} = \mathbf{e}_i' \Sigma \mathbf{e}_j$$

Collect the coefficients e_{ij} into the vector

$$\mathbf{e}_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{pmatrix}$$

First Principal Component (PCA1): Y_1

The *first principal component* is the linear combination of x-variables that has maximum variance (among all linear combinations). It accounts for as much variation in the data as possible.

Specifically we define coefficients $\mathbf{e}_{11}, \mathbf{e}_{12}, \dots, \mathbf{e}_{1p}$ for the first component in such a way that its variance is maximized, subject to the constraint that the sum of the squared coefficients is equal to one. This constraint is required so that a unique answer may be obtained.

More formally, select $\mathbf{e}_{11}, \mathbf{e}_{12}, \dots, \mathbf{e}_{1p}$ that maximizes

$$\text{var}(Y_1) = \sum_{k=1}^p \sum_{l=1}^p e_{1k} e_{1l} \sigma_{kl} = \mathbf{e}_1' \Sigma \mathbf{e}_1$$

subject to the constraint that

$$\mathbf{e}_1' \mathbf{e}_1 = \sum_{j=1}^p e_{1j}^2 = 1$$

Second Principal Component (PCA2): Y_2

The *second principal component* is the linear combination of x-variables that accounts for as much of the remaining variation as possible, with the constraint that the correlation between the first and second component is 0

Select $\mathbf{e}_{21}, \mathbf{e}_{22}, \dots, \mathbf{e}_{2p}$ that maximizes the variance of this new component...

$$\text{var}(Y_2) = \sum_{k=1}^p \sum_{l=1}^p e_{2k} e_{2l} \sigma_{kl} = \mathbf{e}_2' \Sigma \mathbf{e}_2$$

subject to the constraint that the sums of squared coefficients add up to one,

$$\mathbf{e}_2' \mathbf{e}_2 = \sum_{j=1}^p e_{2j}^2 = 1$$

along with the additional constraint that these two components are uncorrelated.

$$\text{cov}(Y_1, Y_2) = \sum_{k=1}^p \sum_{l=1}^p e_{1k} e_{2l} \sigma_{kl} = \mathbf{e}'_1 \Sigma \mathbf{e}_2 = 0$$

All subsequent principal components have this same property – they are linear combinations that account for as much of the remaining variation as possible and they are not correlated with the other principal components.

We will do this in the same way with each additional component. For instance:

***i*th Principal Component (PCA): Y_i**

We select $\mathbf{e}_{i1}, \mathbf{e}_{i2}, \dots, \mathbf{e}_{ip}$ to maximize

$$\text{var}(Y_i) = \sum_{k=1}^p \sum_{l=1}^p e_{ik} e_{il} \sigma_{kl} = \mathbf{e}'_i \Sigma \mathbf{e}_i$$

subject to the constraint that the sums of squared coefficients add up to one...along with the additional constraint that this new component is uncorrelated with all the previously defined components.

$$\mathbf{e}'_i \mathbf{e}_i = \sum_{j=1}^p e_{ij}^2 = 1$$

$$\text{cov}(Y_1, Y_i) = \sum_{k=1}^p \sum_{l=1}^p e_{1k} e_{il} \sigma_{kl} = \mathbf{e}'_1 \Sigma \mathbf{e}_i = 0,$$

$$\text{cov}(Y_2, Y_i) = \sum_{k=1}^p \sum_{l=1}^p e_{2k} e_{il} \sigma_{kl} = \mathbf{e}'_2 \Sigma \mathbf{e}_i = 0,$$

$$\vdots$$

$$\text{cov}(Y_{i-1}, Y_i) = \sum_{k=1}^p \sum_{l=1}^p e_{i-1,k} e_{il} \sigma_{kl} = \mathbf{e}'_{i-1} \Sigma \mathbf{e}_i = 0$$

Therefore all principal components are uncorrelated with one another.

Source URL: <https://onlinecourses.science.psu.edu/stat505/node/51>