PENNSTATE

STAT 505 - Applied Multivariate Statistical

Start Here!

- Welcome to STAT 505!
- Faculty login (PSU Access Account)

Lessons

- Lesson 0: Matrices and Vectors
- Lesson 1: Graphical Display of Multivariate Data
- Lesson 2: Measures of Central Tendency, Dispersion and Association
- Lesson 3: Linear Combinations of Random Variables
- Lesson 4: Multivariate Normal Distribution
- Lesson 5: Sample Mean Vector and Sample Correlation and Related Inference Problems
- Lesson 6: Multivariate Conditional Distribution and Partial Correlation
- Lesson 7: Principal Components Analysis (PCA)
  - 7.1 - Principal Component Analysis (PCA) Procedure
  - 7.2 - How do we find the coefficients?
  - 7.3 - Example: Places Rated
  - 7.4 - Interpretation of the Principal Components
  - 7.5 - Alternative: Standardize the Variables
  - 7.6 - Example: Places Rated after Standardization
  - 7.7 - Once the Components Have Been Calculated
  - 7.8 - Summary
- Lesson 8: Canonical Correlation Analysis
- Lesson 9: Factor Analysis
- Lesson 10: Discriminant Analysis
- Lesson 11: Inferences Regarding Multivariate Population Mean
- Lesson 12: Cluster Analysis
- Lesson 13: Multivariate Analysis of Variance (MANOVA)

Home // Lesson 7: Principal Components Analysis (PCA)

7.1 - Principal Component Analysis (PCA) Procedure

Printer-friendly version

Suppose that we have a random vector **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

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

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

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

Resources

Moreover, Y_i and Y_j will have a population covariance

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

Here the coefficients e_{ij} are collected into the vector

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

First Principal Component (*PCA*1): Y_1

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

Specifically we will define coefficients $e_{11}, e_{12}, \dots, e_{1p}$ for that 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 $e_{11}, e_{12}, \dots, e_{1p}$ that maximizes

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

subject to the constraint that

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

Second Principal Component (*PCA*2): 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 $\boldsymbol{e}_{21}$, $\boldsymbol{e}_{22}$, ... , $\boldsymbol{e}_{2p}$ that maximizes the variance of this new component...

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

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

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

along with the additional constraint that these two components will be uncorrelated with one another.

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

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* (*PCAi*): Y_{i}

We select $\boldsymbol{e}_{i1}$, $\boldsymbol{e}_{i2}$, ... , $\boldsymbol{e}_{ip}$ that maximizes

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

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

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

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

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

$$\vdots$$

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

Therefore all principal components are uncorrelated with one another.

‹ Lesson 7: Principal Components Analysis (PCA) up 7.2 - How do we find the coefficients? ›

Printer-friendly version