Tutorial Business Analytics

Tutorial 10 - Tutorial

Exercise 10.1 Compute the Principal Components

Given the following dataset D = {(-3, -1, -1), (0, -1, 0), (-2, -1, 2), (1, -1, 3)}, compute its principal components by following the PCA algorithm introduced in class and generate the transformed data.
Each tuple of the set D stands for an observation or row vector.

a) Calculate the zero-mean dataset X from the given dataset D. Note down the means.

b) Calculate the 3 x 3 covariance matrix ∑ using the following formulas. What can you infer from it?

$$var(x_j) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2$$

$$cov(x_{j_1}, x_{j_2}) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij_1} - \bar{x}_{j_1}) \cdot (x_{ij_2} - \bar{x}_{j_2})$$

Reminder: Since the matrix X is centered you can use the following formulas:

$$var(x_j) = \frac{1}{N-1}\sum_{i=1}^{N} {x_{ij}}^2$$

$$cov(x_{j_1}, x_{j_2}) = \frac{1}{N-1}\sum_{i=1}^{N} x_{ij_1}\, x_{ij_2}$$

c) Find the eigenvalues for the covariance matrix by solving the equation: $|\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_3}| = 0$.

d) Find the corresponding eigenvectors and order them by significance. How is the variance distributed among them?

Hint: Solving the equation $(\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_3})v = 0$ gives you the corresponding eigenvectors.

e) Compute a one-dimensional PCA projection of the dataset.

f) Compute a two-dimensional PCA projection of the dataset.

Hint for e) and f): The general formula for projections is: $Z = X\Phi$

Exercise 10.2 Reconstruction of the Original Data

Making use of the PCA projections computed in Exercise 10.1, restore the original dataset using the formula: $D \approx Z\Phi^T + means$

a) Reverse the one-dimensional PCA projection to restore the original data. How would the data look when plotted into the original coordinate system?

b) What result do you expect when reconstructing the original data from the two-dimensional PCA projection? What is the information loss?

Exercise 10.3 Principal Component Regression vs Linear Regression

Install/open the "AER" (Applied Econometrics with R) package and open the "HousePrices" data set, which holds information about the prices of houses sold in Canada during three months in 1987.

a) Check the structure of the dataset. Filter the numerical attributes and discard the rest.

```
HousePrices <- HousePrices[,unlist(lapply(HousePrices, is.numeric))]
```

b) Build a model to predict the price of a house given the other independent variables using principal component regression with one component. How much of the dependent variable is explained by the model?

```
pcr_auto <- pcr(price~., data=HousePrices, scale=TRUE, ncomp = 1)
```

c) Build a model to predict the price of a house using simple OLS regression for each independent variable separately. Which OLS model explains best the price? What percentage of variation is explained in this case?

d) Compare the models derived from b) and c). Which one would you choose in this scenario? Give reasons.

Note: Use R to solve this exercise (Exercise10.3_R_template.R).

Tutorial Business Analytics

Tutorial 10 – Solution

Exercise 10.1 Compute the Principal Components

a) The matrix of our dataset D:

$$D = \begin{bmatrix} -3 & -1 & -1 \\ 0 & -1 & 0 \\ -2 & -1 & 2 \\ 1 & -1 & 3 \end{bmatrix}$$

We calculate the mean values for each feature:

$$\bar{d}_1 = \frac{1}{4} \cdot \big((-3) + 0 + (-2) + 1\big) = \frac{1}{4} \cdot (-4) = -1$$

$$\bar{d}_2 = \frac{1}{4} \cdot \big((-1) + (-1) + (-1) + (-1)\big) = \frac{1}{4} \cdot (-4) = -1$$

$$\bar{d}_3 = \frac{1}{4} \cdot \big((-1) + 0 + 2 + 3\big) = \frac{1}{4} \cdot 4 = 1$$

We transform our dataset to a zero means dataset by subtracting the means:

$$x_j = d_j - \bar{d}_j$$

$$X = \begin{bmatrix} -3-(-1) & -1-(-1) & -1-1 \\ 0-(-1) & -1-(-1) & 0-1 \\ -2-(-1) & -1-(-1) & 2-1 \\ 1-(-1) & -1-(-1) & 3-1 \end{bmatrix} = \begin{bmatrix} -2 & 0 & -2 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix}$$

b) The covariance matrix of the centered dataset is computed by determining the variances $var(x_j)$ for each feature and the covariance $cov(x_{j_1}, x_{j_2})$ between features.

$$\sum_x = \begin{bmatrix} var(x_1) & cov(x_1, x_2) & cov(x_1, x_3) \\ cov(x_2, x_1) & var(x_2) & cov(x_2, x_3) \\ cov(x_3, x_1) & cov(x_3, x_2) & var(x_3) \end{bmatrix}$$

Given that the mean of each feature is now 0, we calculate:

$$var(x_j) = \frac{1}{N-1} \sum_{i=1}^{N} \big(x_{ij} - \bar{x}_j\big)^2 = \frac{1}{N-1} \sum_{i=1}^{N} {x_{ij}}^2$$

$$var(x_1) = \frac{1}{4-1} \cdot ((-2)^2 + 1^2 + (-1)^2 + 2^2) = \frac{1}{3} \cdot (4 + 1 + 1 + 4) = \frac{10}{3}$$

Similarly, $var(x_2) = 0$ and $var(x_3) = \frac{10}{3}$.

The covariance is calculated:

$$cov(x_{j_1}, x_{j_2}) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij_1} - \bar{x}_{j_1}) \cdot (x_{ij_2} - \bar{x}_{j_2}) = \frac{1}{N-1}\sum_{i=1}^{N} x_{ij_1}\, x_{ij_2}$$

$$cov(x_1, x_2) = \frac{1}{4-1} \cdot \big((-2)\cdot 0 + 1 \cdot 0 + (-1)\cdot 0 + 2 \cdot 0\big) = \frac{1}{3} \cdot 0 = 0$$

$$cov(x_1, x_3) = \frac{1}{4-1} \cdot \big((-2)\cdot(-2) + 1 \cdot (-1) + (-1)\cdot 1 + 2 \cdot 2\big) = \frac{1}{3} \cdot 6 = 2$$

$$cov(x_2, x_3) = \frac{1}{4-1} \cdot (0\cdot(-2) + 0 \cdot (-1) + 0 \cdot 1 + 0 \cdot 2) = \frac{1}{3} \cdot 0 = 0$$

We do not need to compute the other covariance values as we know that the covariance matrix is symmetric.

$$\sum_x = \begin{bmatrix} 10/3 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 10/3 \end{bmatrix}$$

From the covariance matrix, we can expect the variables x_1 and x_3 to increase together, they are positively correlated. We also notice that for x_2 the variance and the relative covariance values are 0. That is because it is a constant feature, i.e. its column has only one value, and therefore no variance. We can already assume that the characteristic polynomial of the covariance matrix will yield an eigenvalue of value 0.

c) To compute the eigenvalues of the covariance matrix $\sum_x$, we need to solve the characteristic equation $|\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_3}| = 0$.

First, we derive the characteristic polynomial of $\sum_{\boldsymbol{x}}$:

$$\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_3} = \begin{bmatrix} \frac{10}{3} & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & \frac{10}{3} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{10}{3} - \lambda & 0 & 2 \\ 0 & -\lambda & 0 \\ 2 & 0 & \frac{10}{3} - \lambda \end{bmatrix}$$

$$\begin{aligned} |\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_3}| &= \left(\frac{10}{3} - \lambda\right)(-1)^2\left(-\lambda\left(\frac{10}{3} - \lambda\right)\right) + 2 \cdot (-1)^4(0 + 2 \cdot \lambda) \\ &= -\lambda\left(\frac{10}{3} - \lambda\right)^2 + 4 \cdot \lambda = -\lambda\left(\frac{100}{9} - \frac{20}{3} \cdot \lambda + \lambda^2\right) + 4 \cdot \lambda \\ &= -\lambda^3 + \frac{20}{3}\lambda^2 - \frac{64}{9}\lambda \end{aligned}$$

Then, we solve the characteristic equation for λ:

$$-\lambda^3 + \frac{20}{3}\lambda^2 - \frac{64}{9}\lambda = 0 \Rightarrow -9\lambda^3 + 60\lambda^2 - 64\lambda = 0 \Rightarrow \lambda(-9\lambda^2 + 60\lambda - 64) = 0$$

Hence, $\lambda_1 = 0$

We solve the second degree equation for the other two eigenvalues:

$-9\lambda^2 + 60\lambda - 64 = 0$

$\lambda = \frac{-b \pm \sqrt{b^2-4ac}}{2a} = \frac{-60 \pm \sqrt{60^2-4\cdot(-9)\cdot(-64)}}{2\cdot(-9)} = \frac{-60 \pm \sqrt{3600-2304}}{-18} = \frac{-60 \pm 36}{-18}$

$\lambda_2 = \frac{-24}{-18} = \frac{4}{3}$

$\lambda_3 = \frac{-96}{-18} = \frac{16}{3}$

d) The corresponding eigenvectors are found by using these values of λ in the equation $(\sum_{\boldsymbol{x}} - \lambda \boldsymbol{I_3})v = 0$.

We can already ignore $\lambda_1 = 0$, because variance along the corresponding eigenvector would be exactly 0, i.e. all data points fall in the exact same point. Consequently, no significant principal component will be computed. Nevertheless, for correctness:

- For $\lambda_1 = 0$:

$(\sum_x - 0\, I_3)v = 0 \Rightarrow \begin{bmatrix} \frac{10}{3} & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & \frac{10}{3} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0 \Rightarrow \begin{cases} \frac{10}{3} v_1 + 2v_3 = 0 \\ 2v_1 + \frac{10}{3} v_3 = 0 \end{cases} \Rightarrow v_1 = v_3 = 0$, while v_2 can take any values. Thus the eigenvectors of $\sum_x$ corresponding to $\lambda_1 = 0$ are of the form $r \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, where r is a scalar.

- For $\lambda_2 = \frac{4}{3}$:

$\left(\sum_x - \frac{4}{3} I_3\right) v = 0 \Rightarrow \begin{bmatrix} 2 & 0 & 2 \\ 0 & -\frac{4}{3} & 0 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0 \Rightarrow \begin{cases} 2v_1 + 2v_3 = 0 \\ -\frac{4}{3} v_2 = 0 \\ 2v_1 + 2v_3 = 0 \end{cases} \Rightarrow v_1 = -v_3 \; and \; v_2 = 0$

Thus, the eigenvectors of $\sum_x$ corresponding to $\lambda_2 = \frac{4}{3}$ are of the form $r \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$, where r is a scalar. After normalizing, i.e. the sum of squares of the vector elements is 1, we get: $\begin{bmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$.

- For $\lambda_3 = \frac{16}{3}$:

$$\left(\sum_x - \frac{16}{3} I_3\right) v = 0 \Rightarrow \begin{bmatrix} -2 & 0 & 2 \\ 0 & -\frac{16}{3} & 0 \\ 2 & 0 & -2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0 \Rightarrow \begin{cases} -2v_1 + 2v_3 = 0 \\ -\frac{16}{3} v_2 = 0 \\ 2v_1 - 2v_3 = 0 \end{cases} \Rightarrow \begin{cases} v_1 = v_3 \\ v_2 = 0 \end{cases}$$

Thus, the eigenvectors of $\sum_x$ corresponding to $\lambda_3 = \frac{16}{3}$ are of the form $r \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$. After normalizing, we get: $\begin{bmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ \frac{\sqrt{2}}{2} \end{bmatrix}$, clearly orthogonal to the other vectors.

The eigenvectors are ordered in decreasing order by the corresponding eigenvalues. We have $\lambda_3 > \lambda_2 > \lambda_1$. Therefore the rotation matrix, or the principal components loadings are:

$$\Phi = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \end{bmatrix}.$$

Variance distribution is calculated by the eigenvalues ratio to their sum:

$\frac{\lambda_3}{(\lambda_3+\lambda_2+\lambda_1)} = \frac{\frac{16}{3}}{\frac{20}{3}} = 80\%$ variance along the first principal component

$\frac{\lambda_3+\lambda_2}{(\lambda_3+\lambda_2+\lambda_1)} = \frac{\frac{20}{3}}{\frac{20}{3}} = 100\%$ variance explained by the first two components

As expected, the last one plays no role in describing the data.

e) For the 1D projection, we multiply the centered dataset with the first principal component:

$$Z = \begin{bmatrix} -2 & 0 & -2 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} -2\sqrt{2} \\ 0 \\ 0 \\ 2\sqrt{2} \end{bmatrix}$$

f) For the 2D projection, we multiply the centered dataset with the first two principal components:

$$Z = \begin{bmatrix} -2 & 0 & -2 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} -2\sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 0 & -\sqrt{2} \\ 2\sqrt{2} & 0 \end{bmatrix}$$

Exercise 10.2 Reconstruction of the Original Data

a) To restore the original dataset from the principal component scores, we multiply the new dataset with the transposed eigenvectors and add the original dimension means:

$$D \approx Z\Phi^T + means$$

$$D \approx \begin{bmatrix} -2\sqrt{2} \\ 0 \\ 0 \\ 2\sqrt{2} \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{bmatrix} + \begin{bmatrix} -1 & -1 & 1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} -3 & -1 & -1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \\ 1 & -1 & 3 \end{bmatrix}$$

When plotting this dataset, the data points would be aligned, as we considered the projection of the data on only one dimension, omitting the variance on the second one. We lose some information in this reconstruction, but we have conserved 80% of the original variance.

b) We do the same for the 2D PCA projection:

$$D \approx \begin{bmatrix} -2\sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 0 & -\sqrt{2} \\ 2\sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} \end{bmatrix} + \begin{bmatrix} -1 & -1 & 1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} -3 & -1 & -1 \\ 0 & -1 & 0 \\ -2 & -1 & 2 \\ 1 & -1 & 3 \end{bmatrix} = D$$

The information loss in this reconstruction is 0. That is explained by the fact that the first two principal components explained 100% of the variance of the data.

Exercise 10.3 Principal Component Regression vs Linear Regression

Solution in R: Exercise10.3_R_Solution.R

a) The dataset has 6 numerical attributes namely: price, lotsize, bedrooms, bathrooms, stories and garage.

b) One dimensional principal component regression explains 49.7% of the variance w.r.t. the dependent variable (price).

c) The OLS model with the most explanatory single covariate is the one that uses lotsize as an independent variable and explains 28.7% of the variance w.r.t. the price.

d) In this scenario, the price would be better predicted by using principle component regression. The single principle component captures much of the variance in the independent variables. In this case, this helps to explain more variance in the dependent variable.