ISEP
IIF.1105 - IF.2301 - Data Science
Academic year 2024-2025

DATA SCIENCE PROJECT
Principal Component Analysis (PCA) and Linear Regression

1 Instructions to read carefully

In this project you will perform Principal Component Analysis and Linear regression with real data. You will work in groups of **3 students**. You will have to prepare a presentation and pass an oral defense. You can use any programming language that performs PCA and Linear regression, however, the use of Python is highly recommended. The instructions are the following :

About the oral defense

The defense will last about 15 minutes per group and it will consist in 10 minutes of oral presentation plus 5 minutes of questions. You should prepare a presentation with the following (minimal) content :

- a cover page with the first name, last name and the student identification number of all the members.
- a table of contents,
- a short introduction,
- the main body of the presentation (results, figures, tables, interpretations, comments or any other element that might help you answer the questions). In this part, you should answer all the questions referred to as [*graded question*] . If necessary, you can use up to three significant digits in your numerical results.
- the conclusion
- the references

It is not necessary to include your code in the presentation. However, you should have it at hand, in case, you have any related questions.

You will find in the hyperplanning the date of the your oral defense. You should submit the presentation file in pdf format one day before the oral defense. To this end, in moodle you will find a deposit box to upload the file. The file name must have the following format :

LastNameStudent1_ LastNameStudent2_ LastNameStudent3.pdf

Just one deliver per group must be done. There is no report to submit, only the presentation ! The language of the presentation can be either French or English.

About the evaluation

The oral defense is divided in 2 parts, an oral presentation and questions. The quality of the oral presentation will be appreciated and it **should not exceed 10 minutes**. It must be clear, explicit and well understandable. During the second part, in turn each member of the group will be asked some questions. The quality of the answers in terms of comments, interpretations and reasoning will be taken into account for the final mark. The evaluation is individual.

2 Data analysis

2.1 The dataset

The dataset contains information about the power consumption of three different distribution networks in Tetouan, a city located in north Morocco. The observations were taken during the summer of 2017, the period when the consumption increases considerably in comparison to the rest of the year. The power consumption is mostly influenced by the weather temperature in °C (`Temperature`), the humidity in % (`Humidity`) and the wind speed in Km/h (`WindSpeed`). In the dataset the variables representing the power consumption in KW of zones 1, 2 and 3 are denoted `PCZone1`, `PCZone2`, and `PCZone3` respectively.

2.2 Preliminary analysis : descriptive statistics

Import the datafile *Tetuan-PC.csv*. Get familiar with the data and answer the questions :

1. [*graded question*] How many observations are there ? How many variables ?
2. [*graded question*] Are there any missing values in the dataset ? If you think it is appropriate, delete the variables concerning missing values.
3. [*graded question*] Calculate descriptive statistics for all the variables. You can use graphics of your choice to help you describe the data (boxplot, scatter plot, etc.). Interpret the results.

2.3 Principal Component Analysis (PCA)

1. [*graded question*] **Theoretical question** If two variables are perfectly correlated in the dataset, would it be suitable to include both of them in the analysis when performing PCA ? Justify your answer. In contrast, what if the variables are completely uncorrelated ?

Practical application : You are going to perform PCA with the *Tetuan-Power-Consumption* dataset. In this part, you will use only the following variables : `Temperature`, `Humidity`, `WindSpeed`, `PCZone1`, `PCZone2` and `PCZone3`.

1. [*graded question*] Calculate the variance of each variable and interpret the results. Do you think it is necessary to standardize the variables before performing *PCA* for this dataset ? Why ?
2. [*graded question*] Perform PCA using the appropriate function with the appropriate arguments and options considering your answer to the previous question. Analyze the output of the function. Interpret the values of the two first principal component loading vectors.
3. [*graded question*] Calculate the percentage of variance explained *(PVE)* by each component ? Plot the *PVE* explained by each component, as well as the cumulative *PVE*. How many components would you keep ? Why ?
4. [*graded question*] Plot the correlation circle to display the loading vectors and Interpret the results.

2.4 Linear Regression

In this part, you will perform linear regression to predict the *"power consumption of zone 1"* (`PCZone1`), that is, the most consuming one. In this section you will consider only the following predictors : `Temperature`, `Humidity`, `WindSpeed`, `PCZone1`, `PCZone2` and `PCZone3`.

[*graded question*] **theoretical question :** Let us suppose that we fit a linear regression model to explain Y as a linear function of two variables X_1 and X_2. Let us denote R^2 the associated coefficient of determination. Interpret R^2 What is the range of values that can be taken by R^2 ? If we denote r_1 and r_2 the coefficient of correlation between X_1 and Y and the coefficient of correlation between X_2 and Y respectively. What is the relationship between R^2 and r_1 and r_2 ?

2.5 Simple linear regression

[*graded question*] Calculate the correlation between the target and all the 5 predictors. Which variable is the most correlated with the target `PCZone1` ? Comment on the results.

[*graded question*] Fit a simple linear regression model using as target variable `PCZone1`, denoted Y, and as feature variable the most correlated variable to it that you identified in the previous question, denoted X :

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

Then, answer the following questions :

1. What are the coefficient estimates ? Interpret coefficient estimate $\hat{\beta}_1$.
2. Give the general expression of a $1 - \alpha$ confidence interval for the parameter β_1. Calculate the 95% confidence interval for this coefficient. Interpret the results.
3. Elaborate the zero slope hypothesis test for coefficient β_1 and conclude if there is an impact of the predictor on the power consumption in zone 1 (`PCZone1`). Is β_1 significantly non zero ?
4. What is the value of the coefficient of determination R^2 ? Interpret this result. Is this model suitable to predict the power consumption in zone 1 (`PCZone1`) ?

2.5.1 Feature selection for multiple linear regression

Now you are going to fit multiple linear regression models in order to predict the target variable `PCZone1` as a function of two or more other predictors.

In some practical situations it is suitable to select only a subset of the predictors instead of considering all the available variables, since some variables can have no or just little statistical significance to predict the target. The *best subset selection* method consists in fitting a separate least squares regression for each possible combination of the available features. In Python you can use the function `combinations()` of the module `itertools` to get all the possible combinations of k predictors for $k \in \{1, ...5\}$.

Perform the following tasks and answer the questions :

1. [*graded question*] Use Best Subset Selection method to select the best model for any possible number of features ranging from 1 to 5. Select the best model. That is, the model for which the adjusted coefficient of determination $\bar{R}^2$ is the highest.
2. [*graded question*] How many features did you keep ? Which ones ?
3. [*graded question*] Why is it more appropriate to use the adjusted coefficient of determination $\bar{R}^2$ instead of the coefficient of determination R^2 when comparing two models with different numbers of predictors ?
4. [*graded question*] For the selected model, what are the values of the coefficient estimates ? Interpret them. What is the value of the coefficient of determination R^2 ? Interpret this value.

5. [*graded question*] For the selected model, perform the zero slope hypothesis test for all the coefficients except β_0 and conclude.
6. [*graded question*] For the selected model, make a prediction of the power consumption in zone 1 given the following conditions : temperature of 26°C, humidity 65%, wind speed 4.2 Km/h. In addition, the power consumption of zones 2 and 3 are 18840 KW and 25700 KW respectively.

3 References

— Salam, A., & El Hibaoui, A. (2018) "Power Consumption of Tetouan City". UCI Machine Learning Repository. `https://archive.ics.uci.edu/dataset/849/power+consumption+of+tetouan+city`.

— A. Salam and A. E. Hibaoui, "Comparison of Machine Learning Algorithms for the Power Consumption Prediction : - Case Study of Tetouan city âĂŞ," 2018 6th International Renewable and Sustainable Energy Conference (IRSEC), Rabat, Morocco, 2018, pp. 1-5.