Ensimag – Grenoble INP – UGA | Year 2024-2025
Introduction to Statistical Learning and Applications
Pedro L. C. Rodrigues pedro.rodrigues@inria.fr
Alexandre Wendling alexandre.wendling@univ-grenoble-alpes.fr

TD 3: Principal component analysis

▶ Exercise 1

Consider dataset $\mathcal{D} = \left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^{i=N}$ with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$.

(a) Show that

$$\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 = \mathrm{tr}(\boldsymbol{\Sigma})$$

where $\bar{\mathbf{x}}_i$ is the average of the samples of the dataset and $\boldsymbol{\Sigma}$ is their sample covariance matrix.

(b) Show that if the samples are standardized (i.e. they have zero mean and unit standard deviation) then

$$\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^2 = p$$

▶ Exercise 2

Define $\mathbf{X}$ as a $N \times p$ data matrix with $\mathbf{x}_i$ vectors on its rows.

Define also the vector $\mathbf{y} \in \mathbb{R}^N$ containing the observations y_i.

Suppose that both the features and the observations have been re-centered so to have zero mean.

(a) Show that the intercept of a multiple linear regression using this dataset will necessarily be zero.

(b) Use the singular value decomposition (SVD) of $\mathbf{X}$ to write an expression for the parameters $\hat{\boldsymbol{\beta}}$ of the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Consider now that we project the data matrix onto a subspace spanned by the q-top principal components of the data matrix $\mathbf{X}$ with $q < p$. Call this new data matrix $\mathbf{Z}$.

(c) Use the SVD of $\mathbf{Z}$ to write an expression for the parameters $\hat{\boldsymbol{\gamma}}$ of the multiple linear regression model

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

(d) Compare and interpret the expressions obtained in exercises (b) and (c).

▶ Exercise 3

We consider the dataset `cars04`, which describes several properties of different car models in the market in 2004. Each observation (i.e. car) is described by 11 features (i.e. properties) listed in Table 1.The goal of this exercise is to summarize and to interpret the data `cars04` using PCA.

Using `python` we run the following instructions:

```
# first import the dataset
import pandas as pd
filename = './cars04.csv'
df = pd.read_csv(filename, index_col=0)
X = df.values[:, 7:]
```

Variable	Meaning
`Retail`	Builder recommended price(US$)
`Dealer`	Seller price (US$)
`Engine`	Motor capacity (liters)
`Cylinders`	Number of cylinders in the motor
`Horsepower`	Engine power
`CityMPG`	Consumption in city (Miles or gallon; proportional to km/liter)
`HighwayMPG`	Consumption on roadway (Miles or gallon)
`Weight`	Weight (pounds)
`Wheelbase`	Distance between front and rear wheels (inches)
`Length`	Length (inches)
`Width`	Width (inches)

Table 1: Variable list for `cars04`

```
# run scikit-learn methods
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
scl = StandardScaler()
pca = PCA()
est = make_pipeline(scl, pca)
est.fit(X)
```

(a) Explain what the code above does. What is the role and effect of the `StandardScaler`?

After running the following lines, we get the table below:

```
for i in range(X.shape[1]):
    variance = pca.explained_variance_[i]
    variance_ratio = pca.explained_variance_ratio_[i]
    print(f'PC{i+1:02d}:', f'{variance:.3f}', f'{variance_ratio:.3f}')
```

```
## PC01: 7.123 0.646
## PC02: 1.889 0.171
## PC03: 0.852 0.077
## PC04: 0.358 0.032
## PC05: 0.276 0.025
## PC06: 0.198 0.018
## PC07: 0.141 0.013
## PC08: 0.087 0.008
## PC09: 0.067 0.006
## PC10: 0.037 0.003
## PC11: 0.001 0.000
```

(b) Are the first two principal components enough to summarize most of the information (i.e. variance) of the dataset? Justify in terms of the proportion of the total variance that they represent.

Principal components are linear combinations of the 11 variables from the dataset, which are printed using the lines below:

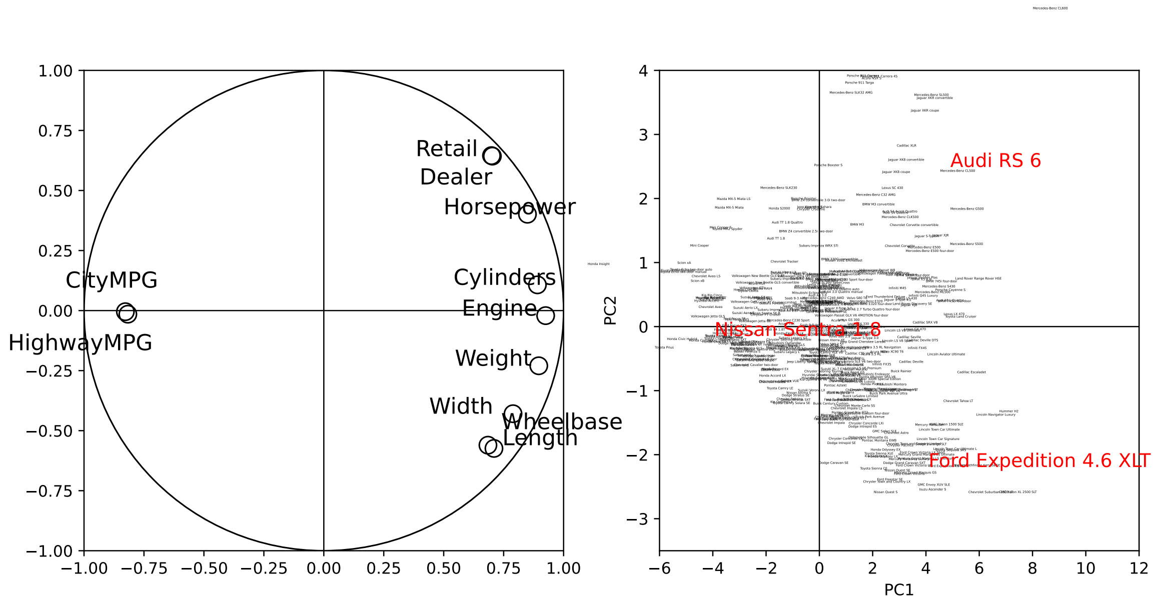
```
df_pc = pd.DataFrame()
df_pc['PC1'] = pca.components_[0, :]
df_pc['PC2'] = pca.components_[1, :]
df_pc.index = df.columns[7:]
print(df_pc)
```

```
##                 PC1       PC2
## Retail       0.263750  0.468509
## Dealer       0.262319  0.470147
## Engine       0.347080 -0.015347
## Cylinders    0.334189  0.078032
## Horsepower   0.318602  0.292213
## CityMPG     -0.310482 -0.003366
## HighwayMPG -0.306589 -0.010964
## Weight       0.336329 -0.167464
## Wheelbase    0.266210 -0.418177
## Length       0.256790 -0.408411
## Width        0.296055 -0.312891
```

(c) How would you interpret these new variables in terms of the initial features of the dataset?

(d) The left panel of the figure below portrays the correlation plot of the PCA as described in class. Recall how it is constructed and then interpret each of the quadrants for the current dataset.

(e) Based on the projections of the data points on the first two principal components shown on the right panel of the figure below, describe which kind of car Audi RS 6, Ford Expedition 4.6 XLT and Nissan Sentra 1.8 are.



▶ Exercise 4

In this exercise, we will use the results from a survey performed in the 1950s in France. The dataset contains the average number of Francs spent on several categories of food products according to social class and the number of children per family. We display below some of the rows and columns of this dataset.

```
df = pd.read_csv('foodFrance.csv', index_col=0)
print(df)
```

```
##            Class  Children  Bread  Vegetables  ...  Meat  Poultry  Milk  Wine
## 0    Blue collar         2    332         428  ...  1437      526   247   427
## 1   White collar         2    293         559  ...  1527      567   239   258
## 2    Upper class         2    372         767  ...  1948      927   235   433
## 3    Blue collar         3    406         563  ...  1507      544   324   407
## 4   White collar         3    386         608  ...  1501      558   319   363
## 5    Upper class         3    438         843  ...  2345     1148   243   341
```

```
## 6     Blue collar          4     534           660  ...  1620       638   414   407
## 7    White collar          4     460           699  ...  1856       762   400   416
## 8     Upper class          4     385           789  ...  2366      1149   304   282
## 9     Blue collar          5     655           776  ...  1848       759   495   486
## 10   White collar          5     584           995  ...  2056       893   518   319
## 11    Upper class          5     515          1097  ...  2630      1167   561   284
##
## [12 rows x 9 columns]
```

(a) Given how the dataset is defined, if we were to do a PCA, would it be preferable to scale or not the variables? Explain your reasoning.

(b) The plots below illustrate the results of the PCA carried out on the dataset. Interpret what information each principal axis convey and how it is related to the different social classes for each data point. Note that the acronyms on the right panel indicate the social class and the number of children, for instance: `WC4` means "White collar with 4 children".