Ensimag – Grenoble INP – UGA Year 2023-2024
Introduction to Statistical Learning and Applications
Pedro L. C. Rodrigues pedro.rodrigues@inria.fr
Alexandre Wendling alexandre.wendling@univ-grenoble-alpes.fr

TD 2: 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

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

(c) Solve the matrix-valued optimization problem below (with $q < p$)

$$\underset{\mathbf{W}^\top \mathbf{W} = \mathbf{I}_q}{\text{minimize}}\ \mathrm{tr}(\mathbf{W}^\top \boldsymbol{\Sigma} \mathbf{W})$$

▶ 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 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 aim of this exercise is to summarize and to interpret the data `cars04` using PCA. Using `R` we run the following instruction:

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`

```
cars04.pca <- prcomp(cars04, scale=TRUE)
summary(cars04.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.6655 1.3726 0.92181 0.59751 0.52482 0.44491 0.37486
## Proportion of Variance 0.6459 0.1713 0.07725 0.03246 0.02504 0.01799 0.01277
## Cumulative Proportion  0.6459 0.8171 0.89439 0.92685 0.95189 0.96988 0.98266
##                           PC8     PC9    PC10    PC11
## Standard deviation     0.29434 0.25766 0.19229 0.02811
## Proportion of Variance 0.00788 0.00604 0.00336 0.00007
## Cumulative Proportion  0.99053 0.99657 0.99993 1.00000
```

(a) What is the effect of the argument `scale=TRUE` in the result of the PCA?

(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. The coefficients of the first 2 principal components on the 11 feature are

```
cars04.pca$rotation[,1:2]
```

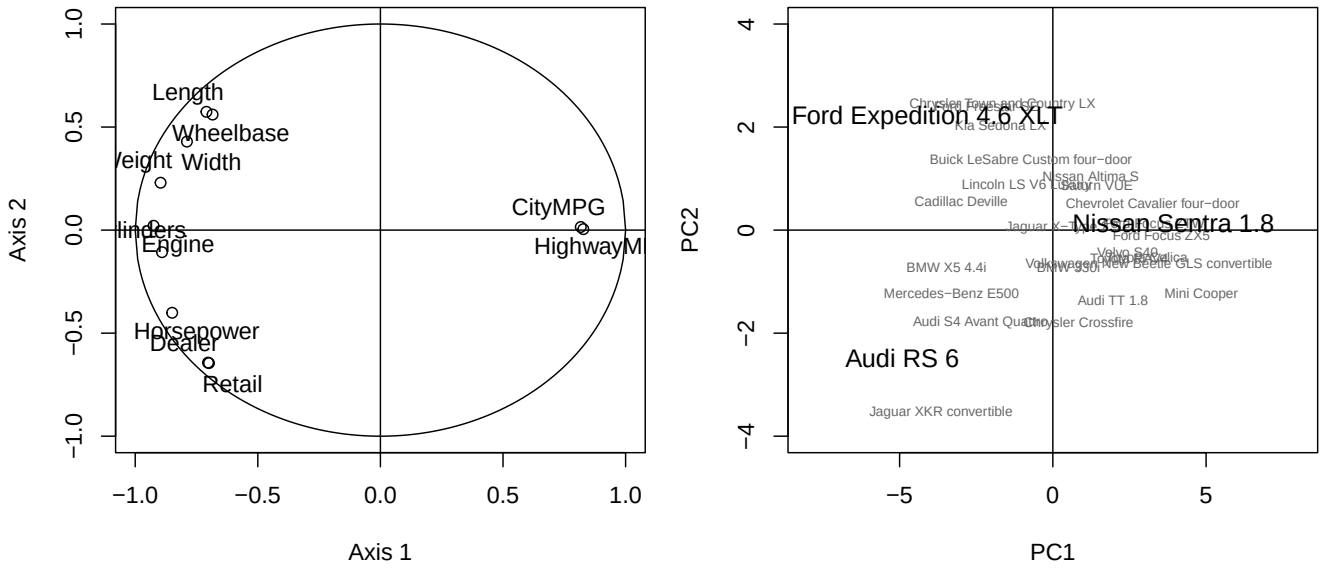
```
##                   PC1          PC2
## Retail     -0.2637504 -0.468508698
## Dealer     -0.2623186 -0.470146585
## Engine     -0.3470805  0.015347186
## Cylinders  -0.3341888 -0.078032011
## Horsepower -0.3186023 -0.292213476
## CityMPG     0.3104817  0.003365936
## HighwayMPG  0.3065886  0.010964460
## Weight     -0.3363294  0.167463572
## Wheelbase  -0.2662100  0.418177107
## Length     -0.2567902  0.408411381
## Width      -0.2960546  0.312891350
```

(c) What would be a good interpretation for these new variables in terms of the initial features of the dataset?

Figure 1 shows the projection the dataset on its first two principal components.

(d) Interpret each quadrant of the Figure.

(e) Can you describe which kind of car Audi RS 6, Ford Expedition 4.6 XLT and Nissan Sentra 1.8 are?



(a) Variable space

(b) Individual space

Figure 1: Principal component representation in the first plane of the variable and of the sample spaces.