Ensimag – Grenoble INP – UGA
Introduction to Statistical Learning and Applications
Pedro L. C. Rodrigues
Alexandre Wendling

Year 2024-2025

pedro.rodrigues@inria.fr
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

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

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

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 &= \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})^\top (\mathbf{x}_i - \bar{\mathbf{x}}) \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathrm{tr}\Big((\mathbf{x}_i - \bar{\mathbf{x}})^\top (\mathbf{x}_i - \bar{\mathbf{x}})\Big) \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathrm{tr}\Big((\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top\Big) \\
&= \mathrm{tr}\left(\frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top\right) \\
&= \mathrm{tr}(\mathbf{\Sigma})
\end{aligned}$$

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

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

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

▶ 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.

In multiple linear regression we have the model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Note that if we take the expectation from both sites, we get

$$\mathbb{E}[Y] = \beta_0 + \beta_1 \mathbb{E}[X_1] + \cdots + \beta_p \mathbb{E}[X_p] + \mathbb{E}[\varepsilon]$$

Since the predictors and observations have zero-mean, then $\beta_0 = 0$.

(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

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

The SVD of the data matrix is $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ so if we plug this into the expression for $\hat{\beta}$ we get

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = (\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} = (\mathbf{V}\mathbf{D}^2\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\top\mathbf{V}\mathbf{D}\mathbf{U}^\top\mathbf{y}$$

so in the end we get $\hat{\beta} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{y}$ and $\hat{\beta}_i = \sum_{k=1}^{p} \frac{\mathbf{u}_i^\top\mathbf{y}}{d_k}\mathbf{v}_{ik}$

Note also that the predictions with the model are $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{U}\mathbf{D}\mathbf{V}^\top\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{y} = \mathbf{U}\mathbf{U}^\top\mathbf{y}$

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

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

Matrix $\mathbf{Z}$ is the projection of the data matrix on its q-top principal components. Therefore, we have:

$$\underbrace{\mathbf{Z}}_{n\times q} = \underbrace{\mathbf{X}}_{n\times p}\underbrace{\mathbf{V}_q}_{p\times q} = \underbrace{\mathbf{U}}_{n\times p}\underbrace{\mathbf{D}}_{p\times p}\underbrace{\mathbf{V}^\top}_{p\times p}\underbrace{\mathbf{V}_q}_{p\times q} = \mathbf{U}\mathbf{D}\underbrace{\begin{bmatrix}\mathbf{I}_q \\ \mathbf{0}_{(p-q)\times q}\end{bmatrix}}_{p\times q} = \underbrace{\mathbf{U}}_{n\times p}\underbrace{\mathbf{D}_q}_{p\times q} \quad \text{where} \quad \mathbf{D}_q = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_q \\ & \mathbf{0}_{(p-q)\times q} & \end{bmatrix}$$

We calculate the coefficients for the new regression model

$$\hat{\gamma} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y} = (\mathbf{D}_q^\top\mathbf{U}^\top\mathbf{U}\mathbf{D}_q)^{-1}\mathbf{D}_q\mathbf{U}^\top\mathbf{y} = \underbrace{(\mathbf{D}_q^\top\mathbf{D}_q)^{-1}}_{q\times q}\underbrace{\mathbf{D}_q^\top}_{q\times p}\underbrace{\mathbf{U}^\top}_{p\times n}\underbrace{\mathbf{y}}_{n\times 1} = \begin{bmatrix} \frac{1}{d_1}\mathbf{u}_1^\top\mathbf{y} \\ \vdots \\ \frac{1}{d_q}\mathbf{u}_q^\top\mathbf{y} \end{bmatrix}$$

where we note that the γ coefficients can be calculated as if we had q independent simple linear regressions. This is due to the diagonal shape of the matrix $\mathbf{Z}$ as per:

$$\mathbf{Z}^\top\mathbf{Z} = \mathbf{D}_q^\top\mathbf{D}_q = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_q \end{bmatrix}$$

If we take it back to the original space, we get

$$\hat{\beta}^{\text{PCR}} = \mathbf{V}_q\hat{\gamma} = \mathbf{V}_q\begin{bmatrix} \frac{1}{d_1}\mathbf{u}_1^\top\mathbf{y} \\ \vdots \\ \frac{1}{d_q}\mathbf{u}_q^\top\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_q \end{bmatrix}\begin{bmatrix} \frac{1}{d_1}\mathbf{u}_1^\top\mathbf{y} \\ \vdots \\ \frac{1}{d_q}\mathbf{u}_q^\top\mathbf{y} \end{bmatrix} \quad \text{and} \quad \hat{\beta}_i^{\text{PCR}} = \sum_{k=1}^{q} \frac{\mathbf{u}_i^\top\mathbf{y}}{d_k}\mathbf{v}_{ik}$$

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

We notice that the parameters for the linear regression obtained with the q-top principal components is a truncated version of the original least squares parameters. We observe that the terms of the sum depending of small singular values have been discarded.

▶ 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:

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`

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

# 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`?

The code creates a scikit-learn pipeline composed of two steps: a standard scaler and a PCA estimator. The `StandardScaler` has the role of transforming all predictors so that they have zero-mean and unit standard deviation along the training dataset.

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.

The first two components already explain 80% of the variance. This can be considered as enough for many applications, such as visualization and initial data exploratory analysis.

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?

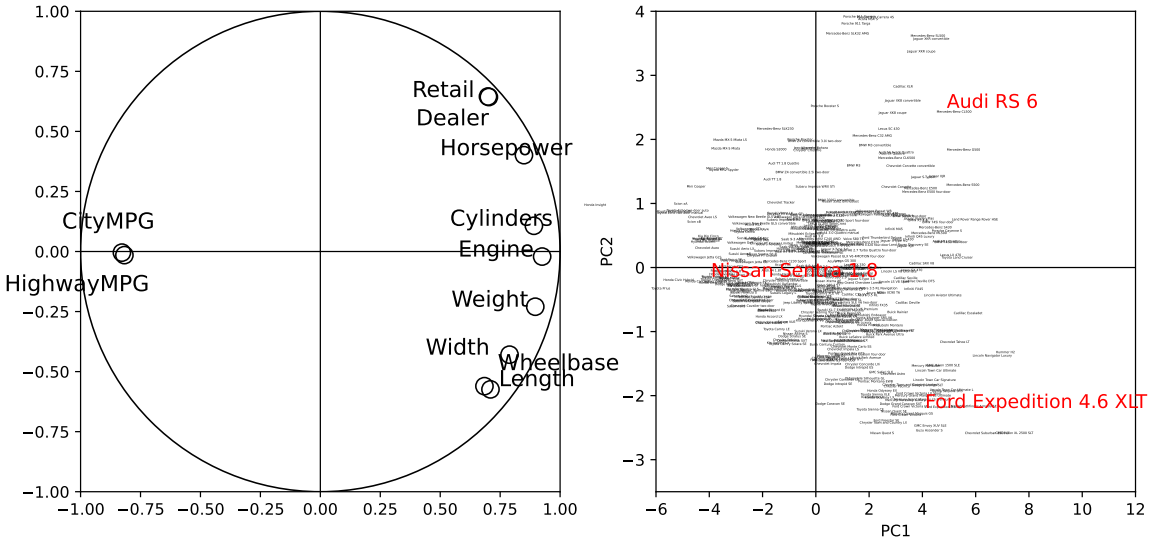
`PC1` seems to give a rather uniform weight to all features, but with opposite effects for `CityMPG` and `HighwayMPG` with respect to the other features. This seems to indicate that PC1 is a variable with strong correlation with the car's consumption. In `PC2` we see that the features with highest values are those related to the price and horsepower in one direction, and car geometry (width, weight, etc) in another direction.

(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.

We could argue that the two left quadrants describe cars which are efficient in terms of gas consumption. The upper-right quadrant seems more related to cars with powerful engines and higher price. The lower-right quadrant seems to be related to larger cars.

(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.

- Audi RS 6 : This is an expensive car with very powerful engine
- Ford Expedition 4.6 XLT : This is a rather big car
- Nissan Sentra 1.8 : Smaller and less expensive car, but very efficient



▶ 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.

No, scaling does not seem a good idea because all predictors are given in the same unity of measure. Therefore, any difference in their variances can be very informative for downstream interpretations.

(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".

The right plot shows that PC1 seems to be related to the social class of the data families, whereas PC2 describes their sizes (i.e. how many children). We also observe that the consumption of meat, poultry, and fruits seems to be orthogonal to that of more "basic" food, such as milk and bread. The

consumption of wine follows an opposite direction as compared to that of "fancy food", which could be interpreted as being an exclusive choice for most families. In the original paper where this data was presented, we have:

"The values of the contributions of the observations to the components indicate that Component 1 contrasts blue collar families with 3 children to upper class families with 3 or more children whereas Component 2 contrasts blue and white collar families with 5 children to upper class families with 3 and 4 children. In addition, the cosines between the components and the variables show that Component 1 contributes to the pattern of food spending seen by the blue collar and white collar families with 2 and 3 children and to the upper class families with 3 or more children while Component 2 contributes to the pattern of food spending by blue collar families with 5 children."

and

"We see that the first component contrasts the amount spent on wine with all other food purchases, while the second component contrasts the purchase of milk and bread with meat, fruit, and poultry. This indicates that wealthier families spend more money on meat, poultry and fruit when they have more children, while white and blue collar families spend more money on bread and milk when they have more children. In addition, the number of children in upper class families seems inversely correlated with the consumption of wine (i.e., wealthy families with 4 or 5 children consume less wine than all other types of families). This curious effect is understandable when placed in the context of the French culture of the 1950s, in which wealthier families with many children tended to be rather religious and therefore less inclined to indulge in the consumption of wine."