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 4: Questions from previous exams

▶ Quick review of PCA

▶ Exercise 1 (credits to Berkeley CS-189)

In linear regression, we model $p(y \mid \mathbf{x}) \sim \mathcal{N}(\beta^\top \mathbf{x} + \beta_0, \sigma^2)$. The irreducible error in this model is

(A) σ^2

(B) $\mathbb{E}[y \mid \mathbf{x}]$

(C) $\mathbb{E}[(y - \mathbb{E}[y \mid \mathbf{x}])^2 \mid \mathbf{x}]$

(D) None of the above.

▶ Exercise 2 (credits to EPFL CS-433)

Which of the following transformations to a data matrix $\mathbf{X}$ will affect the principal components obtained through PCA?

(A) Adding a constant value to all elements of $\mathbf{X}$.

(B) Multiplying one of the features of $\mathbf{X}$ by a constant.

(C) Adding an extra feature to $\mathbf{X}$ (i.e. an extra column) that is constant across all data points.

(D) None of the above answers.

▶ Exercise 3

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.

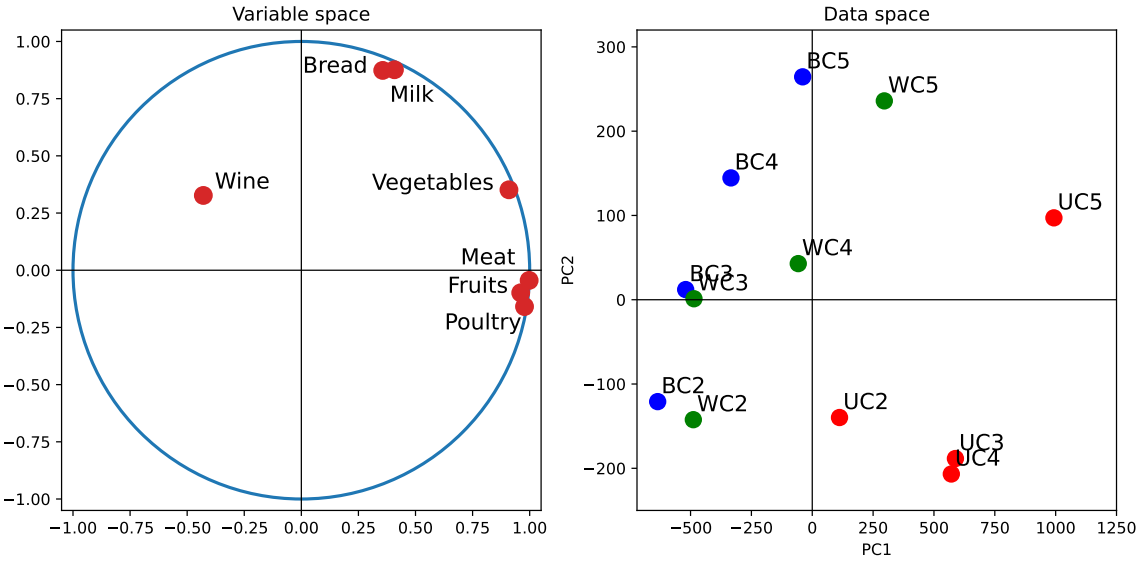
```
import pandas as pd
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import numpy as np

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".



▶ Questions from previous classes and TP1