PCA II: Swiss fertility data

2024-09-02

```
stopifnot(
  require(broom),
  require(DT),
  require(GGally),
  require(ggforce),
  require(ggfortify),
  require(ggvoronoi),
  require(glue),
  require(httr),
  require(magrittr),
  require(patchwork),
  require(skimr),
  require(tidymodels),
  require(tidyverse)
)

old_theme <- theme_set(theme_minimal())
```

- **M1 MIDS/MFA**
- **Université Paris Cité**
- Année 2024-2025
- Course Homepage

- Moodle

> **Objectives**

Swiss fertility data

Dataset `swiss` from `datasets::swiss` connect fertility and social, economic data within 47 French-speaking districts in Switzerland.

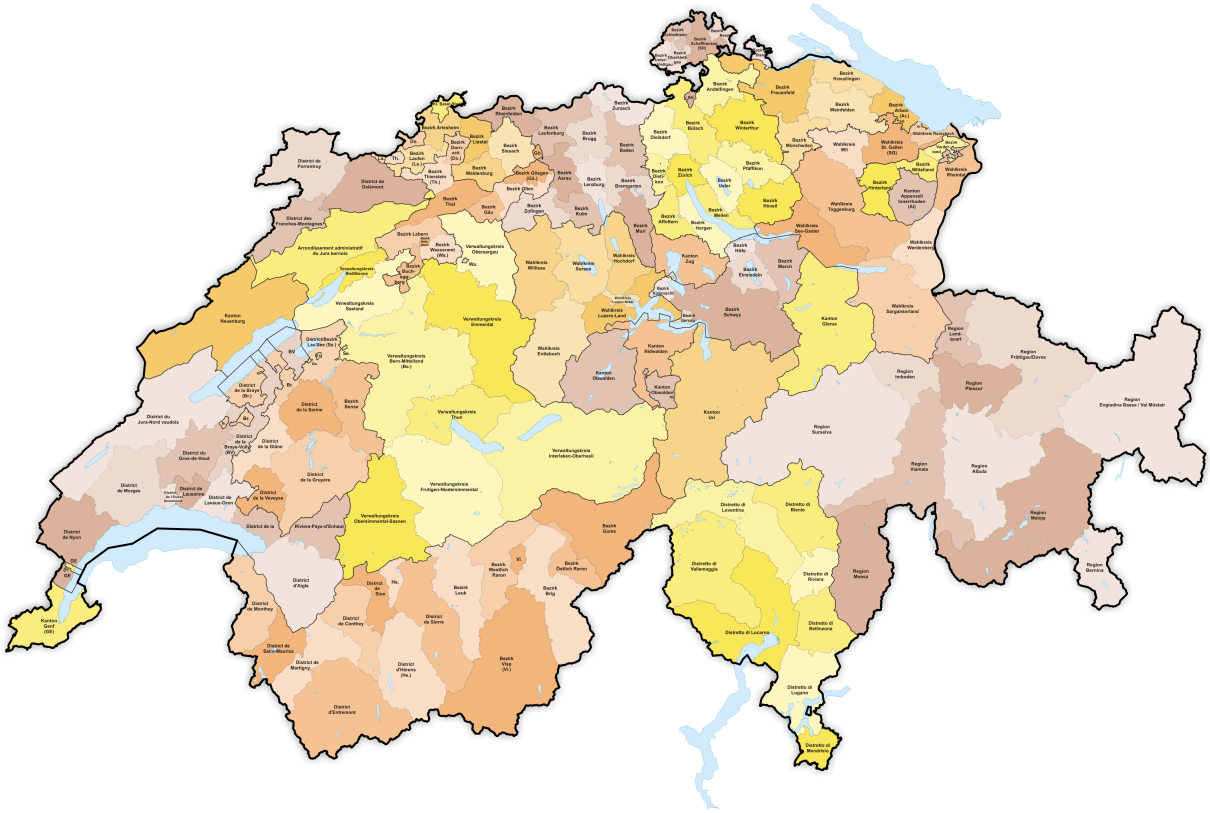
- `Fertility` : fertility index
- `Agriculture` : jobs in agricultural sector
- `Examination` : literacy index (military examination)

- `Education` : proportion of people with successful secondary education
- `Catholic` : proportion of Catholics
- `Infant.Mortality` : mortality quotient at age 0

Fertility index (`Fertility`) is considered as the *response variable*

The social and economic variables are *covariates* (*explanatory* variables).

See European Fertility Project for more on this dataset.



PCA (Principal Component Analysis) is concerned with covariates.

```
data("swiss")

swiss %>%
  glimpse(50)
```

```
Rows: 47
Columns: 6
$ Fertility         <dbl> 80.2, 83.1, 92.5, 85.8,~
$ Agriculture       <dbl> 17.0, 45.1, 39.7, 36.5,~
$ Examination       <int> 15, 6, 5, 12, 17, 9, 16~
$ Education         <int> 12, 9, 5, 7, 15, 7, 7, ~
$ Catholic          <dbl> 9.96, 84.84, 93.40, 33.~
$ Infant.Mortality <dbl> 22.2, 22.2, 20.2, 20.3,~
```

Have a look at the documentation of the dataset

Describe the dataset

- Compute summary for each variable
- Display graphic summary for each variable.

Investigate correlations

Compute, display and comment the sample correlation matrix.

Display jointplots for each pair of variables.

Perform PCA on covariates

Pairwise analysis did not provide us with a clear and simple picture of the French-speaking districts.

Play with centering and scaling

Project the dataset on the first two principal components (perform dimension reduction) and build a scatterplot. Colour the points according to the value of original covariates.

Sanity checks

- X : data matrix after column centering (use `scale(., center=T, scale-F)`)

$$X$$

Checking Orthogonality of V

Compare standardized and non-standardized PCA

Pay attention to the correlation circles.

1. How well are variables represented?
2. Which variables contribute to the first axis?

Explain the contrast between the two correlation circles.

In the sequel we focus on standardized PCA.

Investigate eigenvalues of covariance matrix

How many axes should we keep?

Provide an interpretation of the first two principal axes

1. Which variables contribute to the two first principal axes?
2. Analyze the signs of correlations between variables and axes?

Add the `Fertility` variable

Plot again the correlation circle using the same principal axes as before, but add the `Fertility` variable. How does `Fertility` relate with covariates? with principal axes?

Display individuals (districts)

Comment

Biplot

References

https://scholar.google.com/citations?user=xbCKOYMAAAAJ&hl=fr&oi=ao