LAB: Principal Component Analysis

2025-03-18

```
# We will use the following packages.
# If needed, install them : pak::pkg_install().
stopifnot(
  require("corrr"),
  require("magrittr"),
  require("lobstr"),
  require("ggforce"),
  require("gt"),
  require("glue"),
  require("skimr"),
  require("patchwork"),
  require("tidyverse"),
  require("ggfortify")
  # require("autoplotly")
)
```

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

options(ggplot2.discrete.colour="viridis")
options(ggplot2.discrete.fill="viridis")
options(ggplot2.continuous.fill="viridis")
options(ggplot2.continuous.colour="viridis")
```

M1 MIDS/MFA/LOGOS
Université Paris Cité
Année 2024
Course Homepage
Moodle

Swiss fertility data

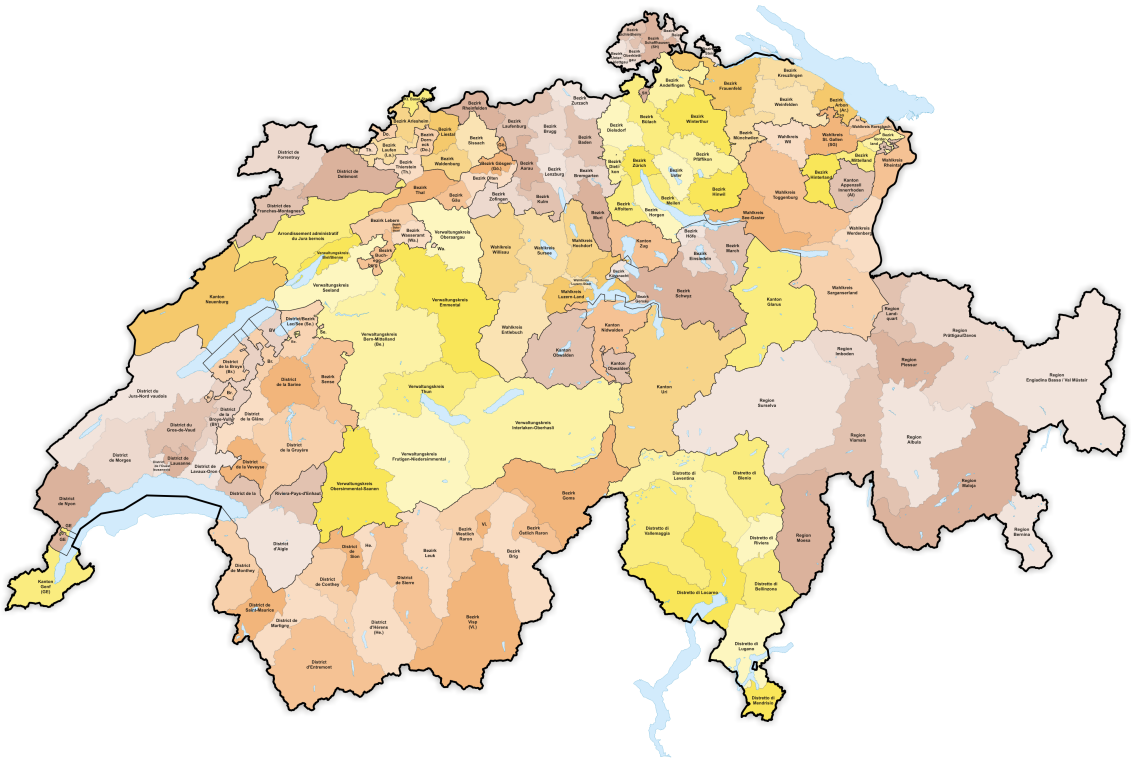
Dataset `swiss` from `datasets::swiss` connect fertility and social, economic data within 47 French-speaking districts in Switzerland.

- `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

> **Question**
>
> Compute summary for each variable

> **Question**
>
> Display graphic summary for each variable.

Investigate pairwise correlations

Question

- Compute, display and comment the sample correlation matrix
- Display jointplots for each pair of variables

Singular Value Decomposition (SVD)

Question

- Project the `swiss` dataset on the covariates (all columns but `Fertility`)
- Center the projected data using matrix manipulation
- Center the projected data using `dplyr` verbs
- Compare the results with the output of `scale()` with various optional arguments
- Call the centered matrix `Y`

Question

Check that the ouput of `svd(Y)` actually defines a Singular Value Decomposition.

Question

Relate the SVD of Y and the eigen decomposition of $Y^\top \times Y$

Perform PCA on covariates

Question

Pairwise analysis did not provide us with a clear and simple picture of the French-speaking districts.
PCA (Principal Component Analysis) aims at exploring the variations of multivariate datasets around their mean (center of inertia). In the sequel, we will perform PCA on the matrix of centered covariates, with and without standardizing the centered columns.
Base `R` offers `prcomp()`. Call `prcomp()` on the centered covariates
Note that `R` also offers `princomp`

Question

Check that `prcomp()` is indeed a wrapper for `svd()`.

Question

Check that rows and columns of component `rotation` of the result of `prcomp()` have unit norm.

Question

Check Orthogonality of V (component `rotation` of the `prcomp` object)

Question

Make a scatterplot from the first two columns of the x component of the `prcomp` object.

Question

Define a graphical pipeline for the *screeplot*.
Hint: use function `tidy()` from `broom`, to get the data in the right form from an instance of `prcomp`.

Question

Define a function that replicates `autoplot.prcomp()`
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.
Hint: use generic function `augment` from `broom`.

Question

Apply `broom::tidy()` with optional argument `matrix="v"` or `matrix="loadings"` to the `prcomp` object.
Comment.

Question

Build the third SVD plot, the so called *correlation circle*.

Question

Compute PCA after standardizing the columns, draw the correlation circle.

Compare standardized and non-standardized PCA

Question

Pay attention to the correlation circles.
1. How well are variables represented?
2. Which variables contribute to the first axis?

Question

Explain the contrast between the two correlation circles.

In the sequel we focus on standardized PCA.

Provide an interpretation of the first two principal axes

> **Question**
>
> Which variables contribute to the two first principal axes?

> **Question**
>
> Analyze the signs of correlations between variables and axes?

Add the `Fertility` variable

> **Question**
>
> 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?

Biplot

> **Question**
>
> The last svd plot (biplot) consists of overlaying the scatter plot of component `x` of the `prcomp` object and the correlation circle.
> So the biplot is a graphical object built on two dataframes derived on components `x` and `rotation` of the `prcomp` objects.
> Design a graphical pipeline.

> **Question**
>
> `autoplot.prcomp()` has optional arguments. If set to `True`, logical argument `loadings` overlays the scatterplot defined by the principal components with the correlation circle.

Generics

`autoplot()` is an example of S3 generic function. Let us examine this function using `sloop`

> Use `sloop::s3_dispatch()` to compare `autoplot(prcomp(swiss))` and `autoplot(lm(Fertility ~ ., swiss))`

> Use `sloop::s3_getmethod()` to see the body of `autoplot.prcomp`

References

S3 classes

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