Ensimag – Grenoble INP – UGA
Modélisation statistique et analyse de données
Christophe Dutang
Olivier François

Année 2023-2024

christophe.dutang@grenoble-inp.fr
olivier.francois@univ-grenoble-alpes.fr

⚠ General guidelines for TPs

Each team shall upload its report on Teide before the deadline indicated at the course website. Please **include the name of all members of the team** on top of your report.

The report should contain graphical representations. For each graph, axis names should be provided as well as a legend when it is appropriate. Figures should be explained by a few sentences in the text. Answer to the **questions in order and refer to the question number in your report**. Computations and graphics have to be performed with `R`.

The report should be written using the `Rmarkdown` format. This is a file format that allows users to format documents containing text and `R` instructions. You should include all of the `R` instructions that you have used in the `rmd` document so that it may be possible to replicate your results. From your `rmd` file, you are asked to generate an `html` file for the final report. In Teide, you are asked to submit both the `rmd` and the `html` files. In the `html` file, you should limit the displayed `R` code to the most important instructions.

TP: Principal Component Analysis on breast cancer dataset

We are interested in detecting severe breast cancers. To this end, we use a datas coming from cytopathologic measures of tissue cell of women developping breast cancers, collected in Wolberg and Mangasarian (1990). *Multisurface method of pattern separation for medical diagnosis applied to breast cytology*, Proceedings of the National Academy of Sciences. `BreastCancer` contains 699 observations 9 explanatory variables and 1 response variable (the identification column is useless), see Table 1. All explanatory variables have been converted into 11 primitive numerical attributes with values ranging from 0 through 10.

This project aims to determine if cytological measures contain enough information to diagnose women for breast cancer. We use a PCA to reduce dimensions and study relationships between cytological measures and the cancer class: `benign` cancer, `malignant` cancer.

Variable	Meaning
`Id`	Sample code number
`Cl.thickness`	Clump Thickness (ordered value)
`Cell.size`	Uniformity of Cell Size (ordered value)
`Cell.shape`	Uniformity of Cell Shape (ordered value)
`Marg.adhesion`	Marginal Adhesion (ordered value)
`Epith.c.size`	Single Epithelial Cell Size (ordered value)
`Bare.nuclei`	Bare Nuclei
`Bl.cromatin`	Bland Chromatin
`Normal.nucleoli`	Normal Nucleoli
`Mitoses`	Mitoses
`Class`	Class

Table 1: Variable list for `BreastCancer`

Data preparation

1. Load the data and remove missing values using the code below.

```
> BreastCancer <- na.omit(read.csv("BreastCancer.csv", header=TRUE, stringsAsFactors = TRUE))
> explvar <- c("Clump.Thickness", "Uniformity.of.Cell.Size", "Uniformity.of.Cell.Shape",
+              "Marginal.Adhesion", "Single.Epithelial.Cell.Size", "Bare.Nuclei",
+              "Bland.Chromatin", "Normal.Nucleoli", "Mitoses")
```

2. Explore the dataset with a (short) descriptive and graphical analysis.
3. Emphasize the fact that using the log of explanatory variables is better. For the report, please carefully select the graphics displayed and comment them.

In the following, we work on `X <- as.matrix(log(BreastCancer[, explvar]))`.

Principal component analysis using `stats` package

There are two functions in the core `stats` package to perform a PCA: `prcomp()` and `princomp()`.

4. What are the differences between these two functions? what are the default behavior?

For the next two questions we check the outputs of the following code

```
> breast.prcomp <- prcomp(X, scale=FALSE)
> breast.princomp <- princomp(X, cor=FALSE)
```

5. We compute the spectral decomposition PDP^T using `eigen()` of the empirical covariance matrix V `X.cov <- cov(X)`.
   (a) Compute the eigenvalues of the covariance V and check you retrieve the explained variances of `breast.prcomp$sdev^ 2`.
   (b) Check the first eigenvector and the first axis `breast.prcomp$loadings[,1]`.
   (c) Check the coordinates of observations (a matrix multiplication) against the output `breast.princomp$x` (up to a sign).
6. We compute the singular value decomposition $U\Sigma V^T$.
   (a) Center the observations `X.s <- scale(X, center = TRUE, scale = FALSE) /sqrt(nrow(X) - 1)`
   (b) Using `svd()`, check the explained variances against the output `breast.prcomp$sdev`.
   (c) Using `svd()`, check the right singular vectors against the output `breast.prcomp$rotation`.

We now interpret the output `breast.prcomp` (the output of `prcomp()`)

```
> breast.prcomp <- prcomp(X, scale=TRUE)
```

7. Interpret the values of explained variances. Use `summary()` and `plot()` on `breast.prcomp`
8. Interpret the components using `biplot()` and

```
> plot(., col=1+1*(BreastCancer$Class == "malignant"))
```

Principal component analysis using `FactoMineR` package

9. Redo the computation using the `PCA()` function of the `FactoMineR` package with `scale.unit=FALSE`. Use `quali.sup` argument to add the qualitative variable `Class`. Check your analysis of the previous question using `summary()`. Furthermore, `dimdesc()` provides further analysis of principal components.
10. Plot individuals in your plane(s) defined by your chosen principal components. Use `habillage` argument of `plot()` on the output of `PCA()` and comment.

```
> plot(., choix="ind", label="quali", habillage = .)
```

11. Retrieve the explanations of variable's contribution to principal components with `plot()` function.

```
> plot(., choix="var", habillage="cos2")
```