STAD–General statistics

Mathieu Ribatet—Full Professor of Statistics

References

[1] D. Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CR, 3rd edition, 2014.

[2] J. P. Klein and M. L. Moeschberger. *Survival analysis*. Springer–Verlag, 2nd edition, 2003.

[3] D. G. Kleinbaum and M. Klein. *Survival analysis: A self learning text*. Springer–Verlag, New–York, 3rd edition, 2012.

1. Descriptive statistics

Types of variables

- □ There are two main type of variables:

  Quantitative such as height, weights, . . .
  Qualitative such colors, lefty/righty, . . .

- □ Often qualitative variables are encoded as integers.
- □ Possible side effect is that computer may wrongly perform standard algebra on those values!
- □ Pressing need to encode them as factors
- □ Note that, if needed, one can convert a quantitative variable to a factor using discretization, e.g., $[0, 5], [5, 10], \dots$

Summary statistics

- □ Having observed a sample $x_1, \dots, x_n$, it is common practice to give a brief summary of the data using summary statistics.
- □ Measures of location refer to the central position of the data, i.e., where a future observation would typically lie.
- □ Measure of dispersion refer to the spread of the data, i.e., does observation can vary a lot or not?

Location sample mean, sample median, midhinge

Dispersion sample standard deviation, range, inter quartile range, MAD

Shape Skewness, kurtosis

Measures of location

Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median

$$\text{Median} = \begin{cases} x_{\frac{n+1}{2}:n}, & n \text{ is odd} \\ 0.5\left(x_{\frac{n}{2}:n} + x_{(\frac{n}{2}+1):n}\right), & n \text{ is even.} \end{cases}$$

Quantile of order p with $0 < p < 1$

$$Q_p = (1-\gamma)x_{j:n} + \gamma x_{j+1:n}, \qquad j = [np + 1 - p], \quad \gamma = np + 1 - p - j$$

Quartiles are special cases with $p = 1/4, 3/4$ and often denoted Q_1 and Q_3.

Measures of dispersion

Standard deviation

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Range

$$\mathsf{Range} = \max x_i - \min x_i$$

Interquartile Range

$$\mathsf{IQR} = Q_3 - Q_1$$

Statistical graphics

□ A picture worths a thousand words

Statistical graphics

- □ A picture worths a thousand words but takes place so need to worth it
- □ Widely used statistical plots are
  - histograms, barplots
  - boxplots
  - scatterplots
  - quantile–quantile plots

Histograms

- □ Histograms are used to visualized the distribution of the data.
- □ They are empirical versions of the probability density function of a quantitative variable
- □ Each class/modality is depicted by a rectangle whose area is proportional to the corresponding class frequency.
- □ Statisticians usually used normalized versions so that the total area of the histogram is 1[1].
- □ More precisely we have

$$h_j = \frac{n_j}{n\ell_j}, \qquad j = 1, \ldots, J, \quad n_j = \# \text{ obs. in class j.}$$

[1] as the probability density function integrates to 1

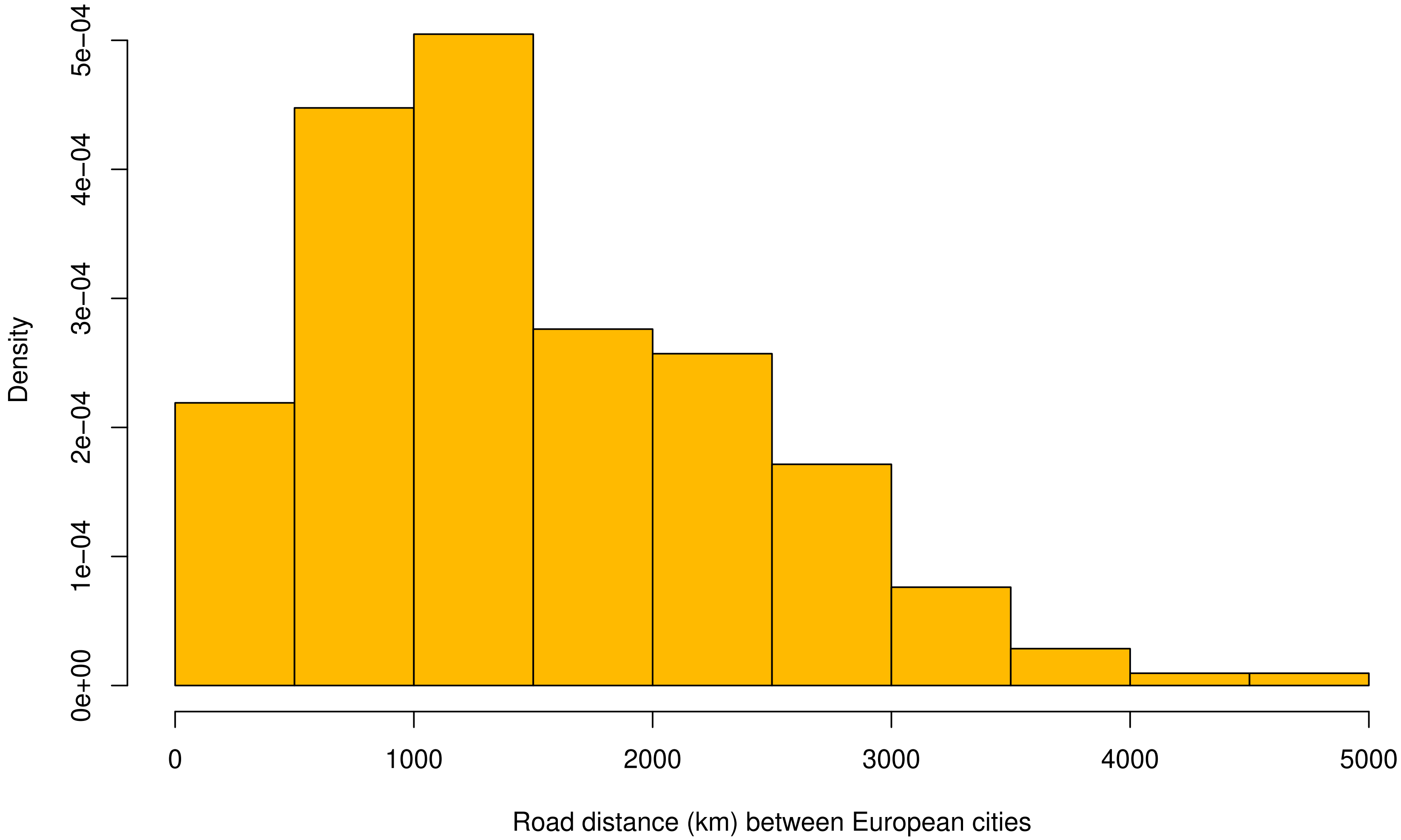


Figure 1: *Histogram of distance in km between 21 European cities.*

Barplots

- □ Barplots are somehow similar to histograms but for categorical variable or variable with finite numbers of possible outcomes.

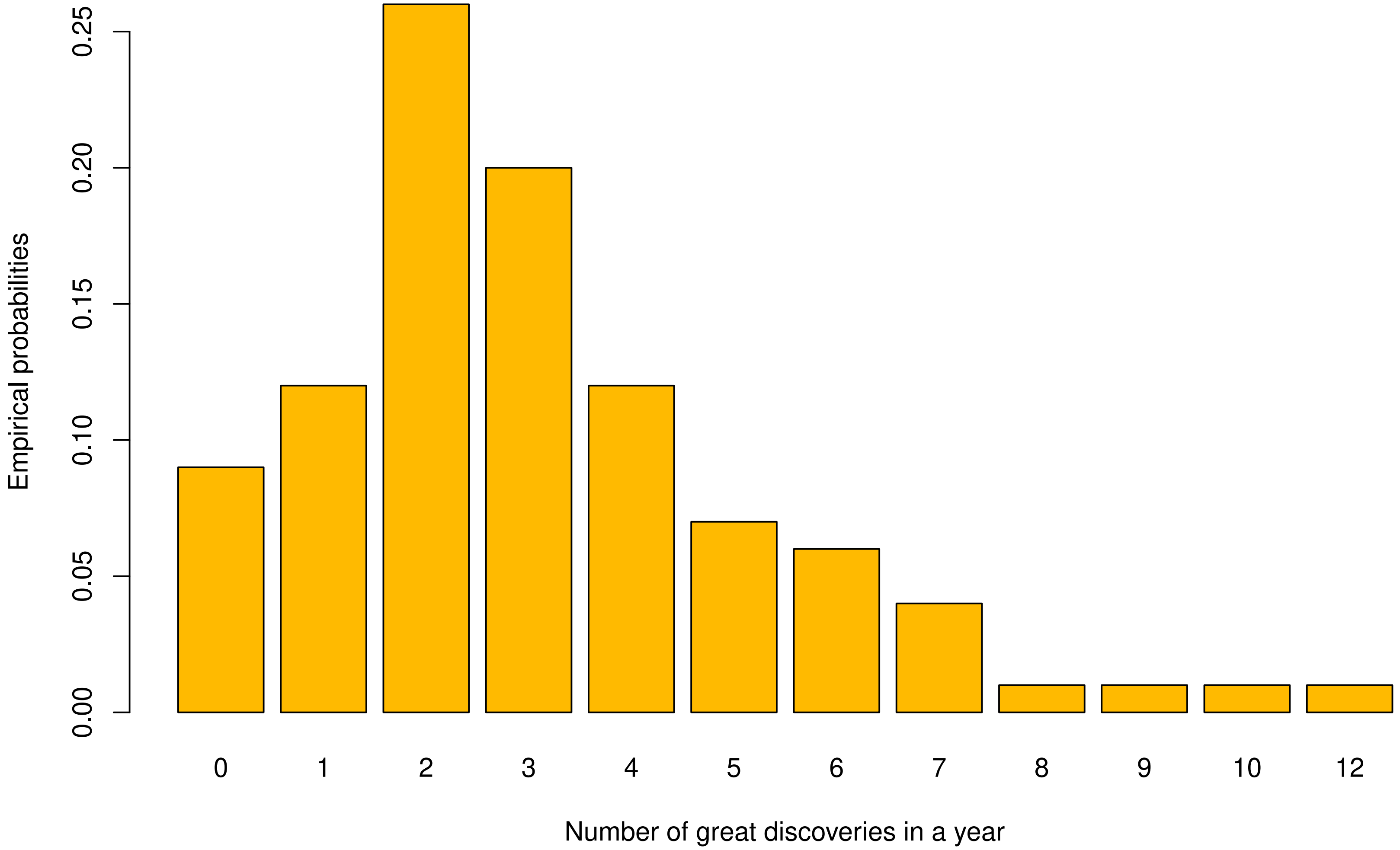


Figure 2: *Barplot of the number of yearly “great” discoveries from 1860 to 1959.*

Boxplots

□ Boxplots also helps visualizing the distribution of the data but take less space.

□ They are never used alone but rather in groups to spot any differences.

□ It consists of a box (Q_1, Q_3 and the median) and whiskers defined as the closest observation[2] to $Q_{1,3} \mp 1.5IQR$.

□ Observation outside those whiskers are usually denoted as outliers.

! Outliers are not spurious observations and should not be discarded. To do so, you need a justification such as measurement problem.

[2] towards the center of the distribution

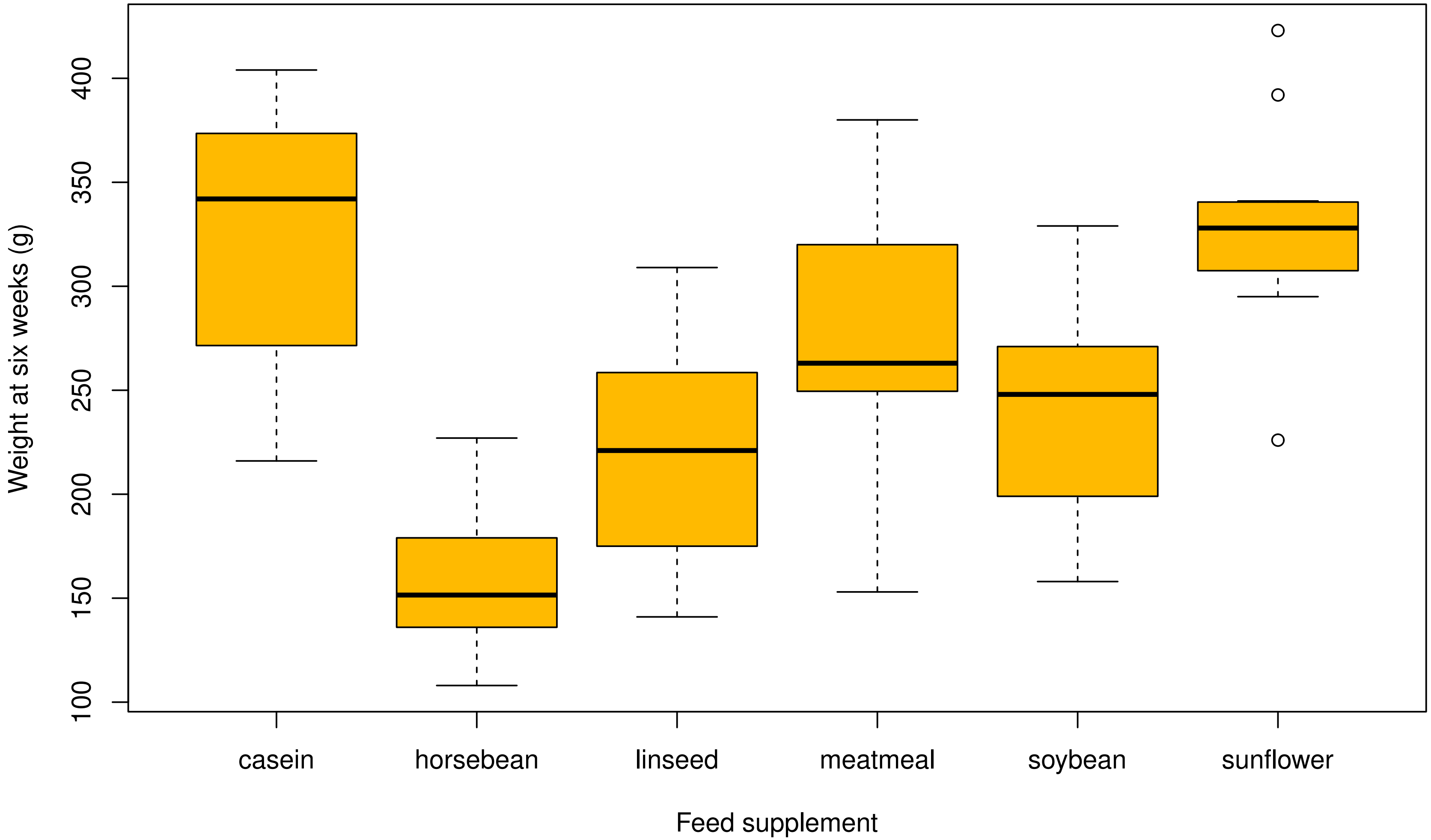


Figure 3: *Boxplots of the weights of chicks (g) with respect to their feed type supplements.*

Scatter plot

- □ Scatter plot aims at visualizing relationship between two variables
- □ Often but not necessarily, those variables are quantitative
- □ We just plot the points $\{(x_i, y_i) \colon i = 1, \ldots, n\}$.

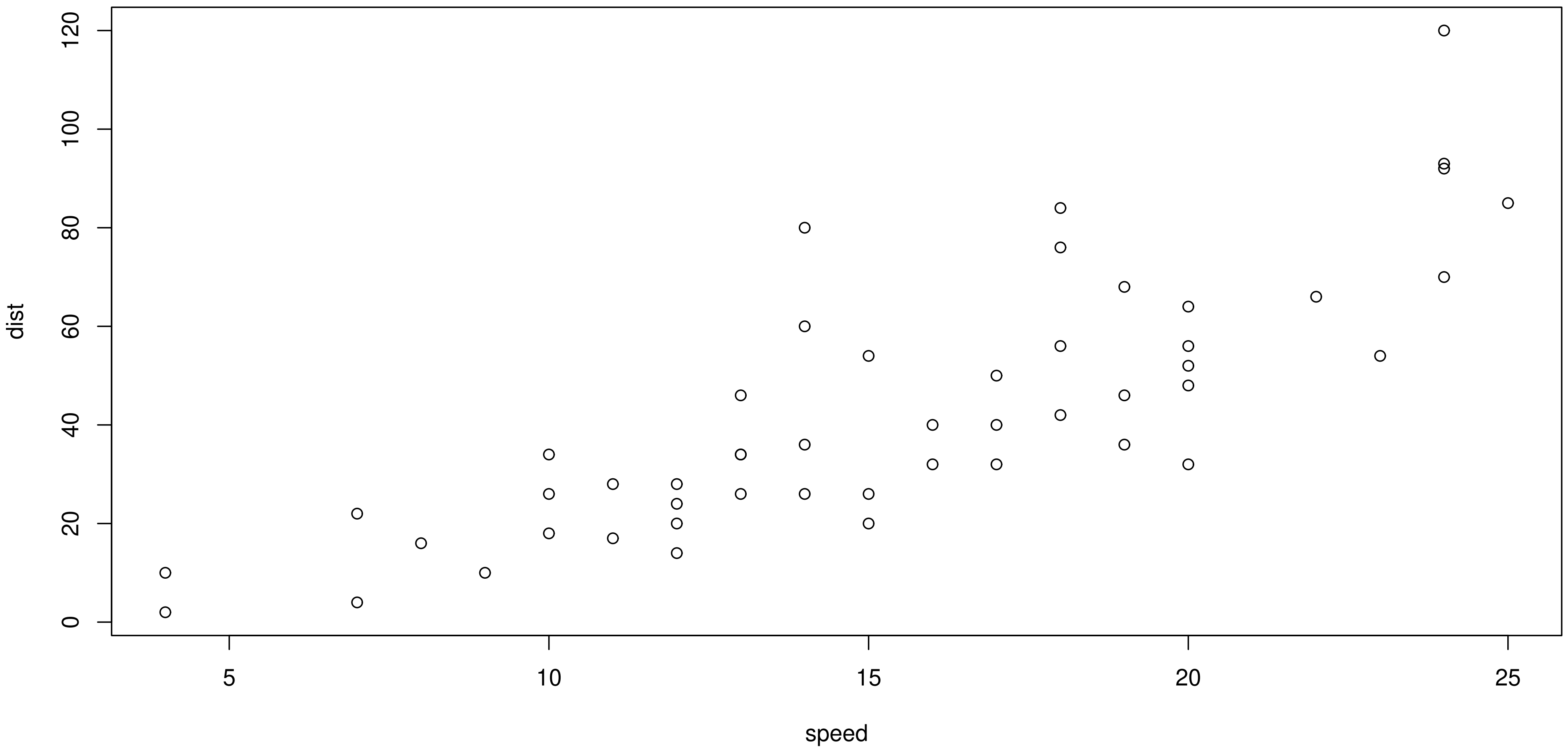
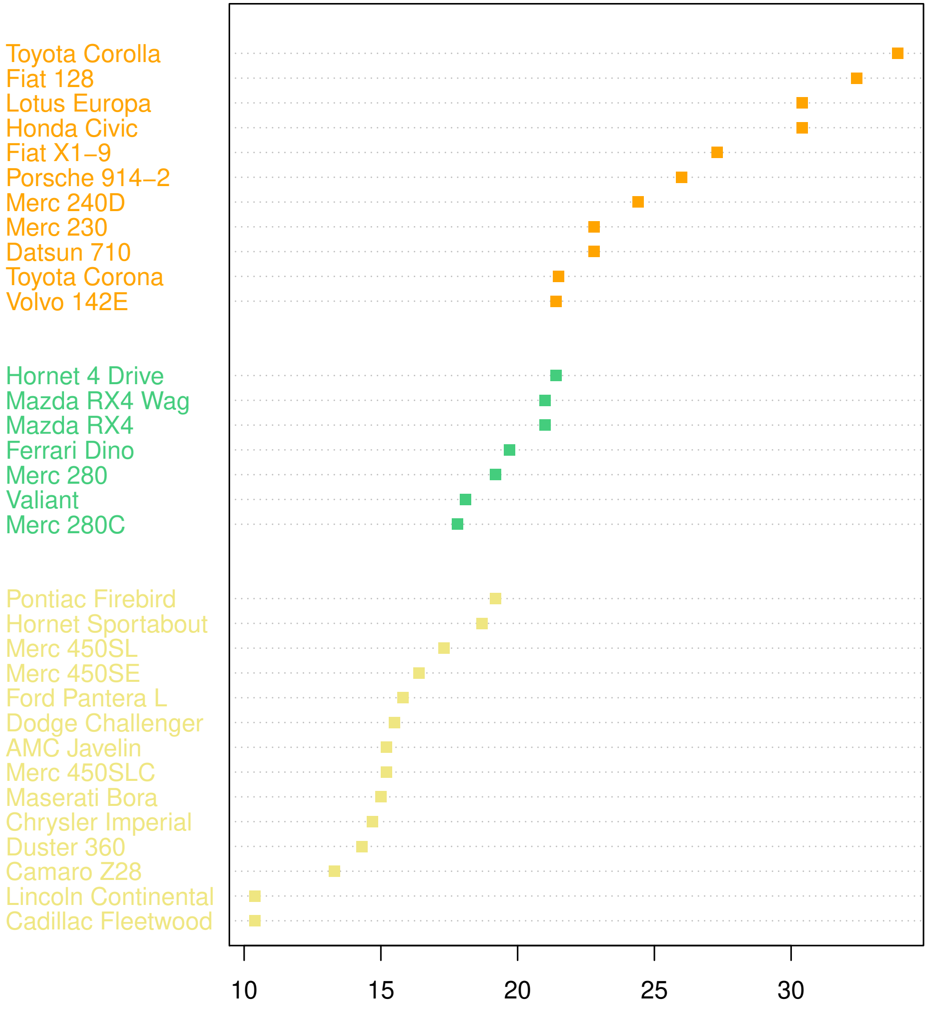


Figure 4: *Scatterplot of the distance taken to stop as the speed varies. A linear dependence seems to occur–theoretically, one would expect a quadratic one, wouldn't we?*

Dotchart



```
                   mpg  cyl
Mazda RX4         21.0    6
Mazda RX4 Wag     21.0    6
Datsun 710        22.8    4
Hornet 4 Drive    21.4    6
Hornet Sportabout 18.7    8
Valiant           18.1    6
...
Lotus Europa      30.4    4
Ford Pantera L    15.8    8
Ferrari Dino      19.7    6
Maserati Bora     15.0    8
Volvo 142E        21.4    4

[32 rows x 11 columns]
```

Figure 5: *Dotchart on the consumption of cars segmented on the number of cylinders.*

QQ-plot

- □ Quantile quantile plots are used to check whether:
  - – two samples share the same distribution
  - – a sample follows a given, e.g., fitted, distribution.
- □ The plot is based on ordered statistics

$$x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n+1}$$

- □ The first version is just a scatter plot of the ordered statistics of the two samples
- □ The second version is a scatter plot of the ordered statistics and the theoretical/fitted quantiles

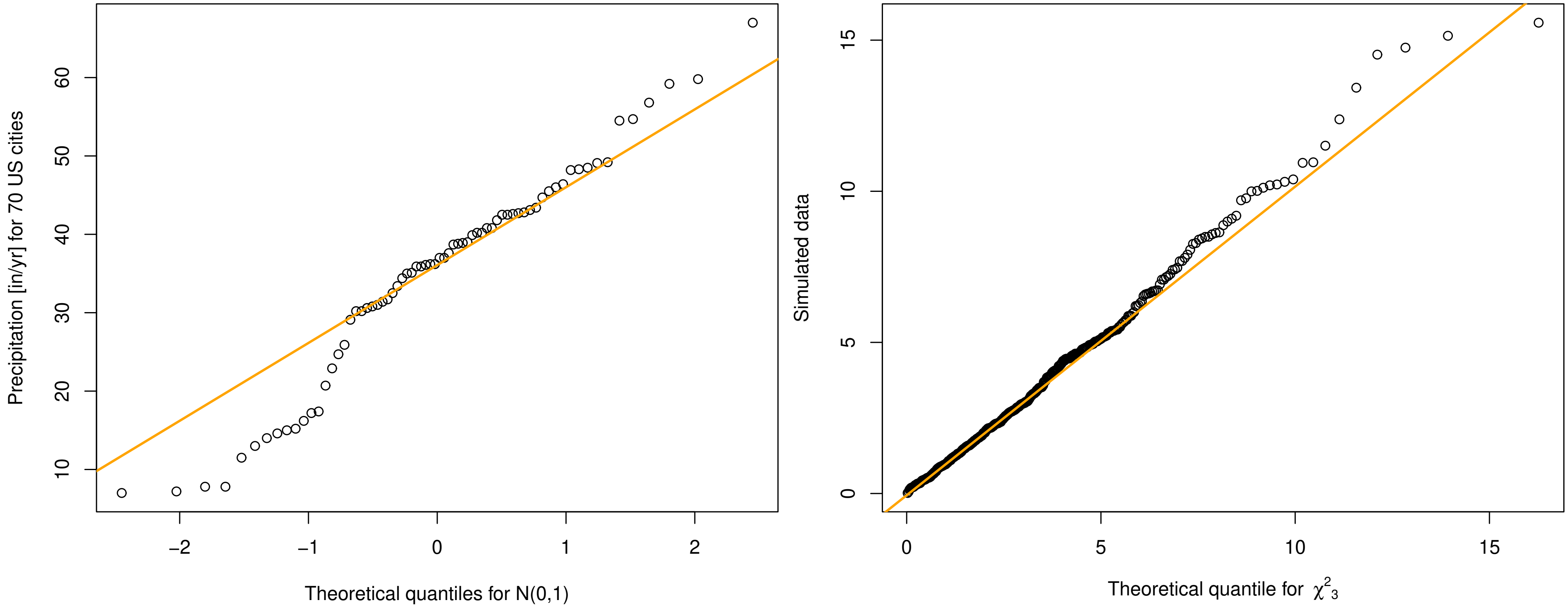


Figure 6: *Illustration on the use of qq-plots. Left: Precipitations doesn't appear to be Gaussian. In particular, the Gaussian distribution appears to overestimate the smallest precipitation amount. Right: The χ^2_3 distribution is reasonnable choice. (here confidence intervals are missing which is (very) unfortunate.)*

2. Statistics models

Probability density function

Definition 1. A probability density function, or density, is a non–negative function f defined on a (non finite) set E and such that

$$\int_E f(x)\mathrm{d}x = 1.$$

Probability density function

Definition 1. A probability density function, or density, is a non–negative function f defined on a (non finite) set E and such that

$$\int_E f(x)\mathrm{d}x = 1.$$

Definition 2. A probability mass function is just as a p.d.f. but for at most enumerable set E, i.e., a non negative function m and such that

$$\sum_{x \in E} m(x) = 1.$$

Probability density function

Definition 1. A probability density function, or density, is a non–negative function f defined on a (non finite) set E and such that

$$\int_E f(x)\mathrm{d}x = 1.$$

Definition 2. A probability mass function is just as a p.d.f. but for at most enumerable set E, i.e., a non negative function m and such that

$$\sum_{x \in E} m(x) = 1.$$

Remark. In general, a random variable can be a mixture of both discrete and continuous cases.

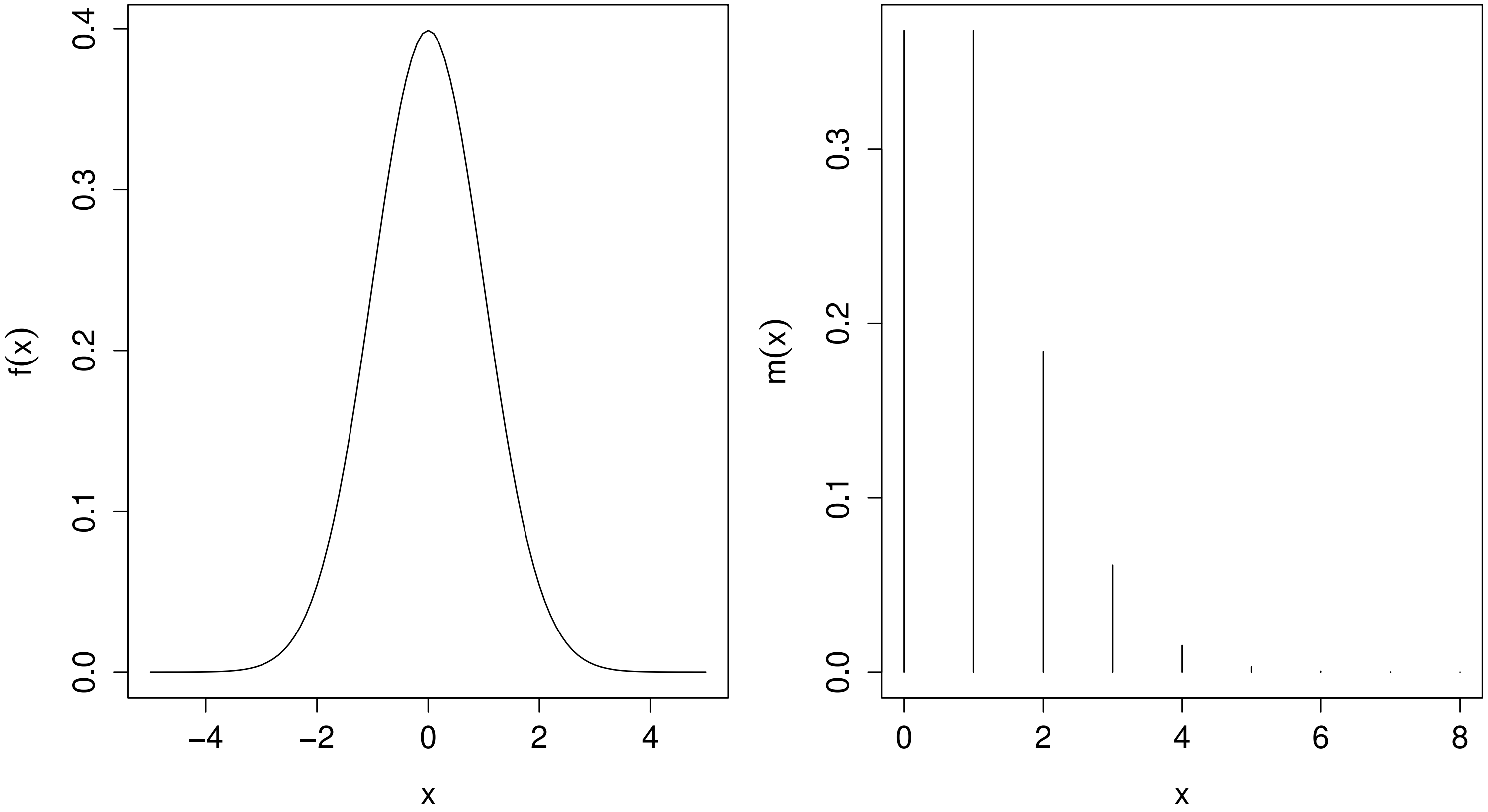


Figure 7: *Examples of probability density/mass functions. Left: Gaussian distribution. Right: Poisson distribution.*

Cumulative distribution function

Definition 3. A cumulative distribution function, or distribution, is a càd–làg function F given by

$$F(x) = \Pr(X \leq x), \qquad x \in E.$$

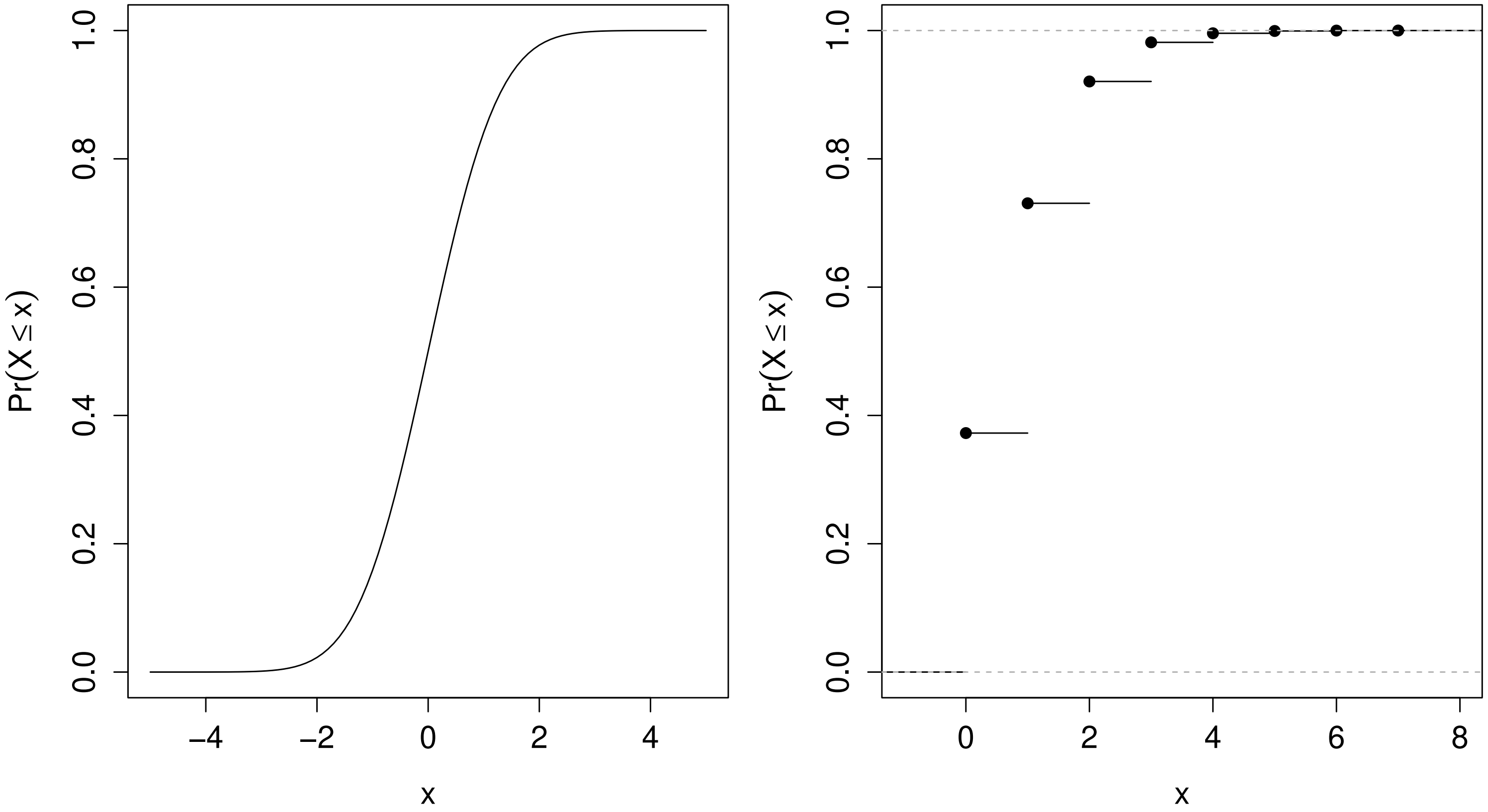


Figure 8: *Examples of cumulative distribution functions. Left: Gaussian distribution. Right: Poisson distribution.*

Statistical models

Definition 4. A parametric family of functions $\{f(x;\theta) \colon x \in E, \theta \in \Theta\}$ is a statistical model if, for any $\theta \in \Theta$, $x \mapsto f(x;\theta)$ is a probability density/mass function on E.
The sets Θ and E are respectively called parameter space and observational space.
The above model is said to be parametric if $\dim(\Theta) < \infty$.

Example 1. The Gaussian model, denoted $X \sim N(\mu, \sigma^2)$, is given by

$$f(x;\theta) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad \theta = (\mu, \sigma^2), \quad E = \mathbb{R}, \quad \Theta = \mathbb{R}\times(0,\infty).$$

Example 2. The Poisson model, denoted $X \sim \text{Poisson}(\lambda)$, corresponds to

$$m(x;\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda), \qquad E = \mathbb{N}, \quad \Theta = (0,\infty).$$

Some statistical models

Table 1: *Examples of useful statistical models*

Name	Support	Scope	p.d.f. // p.m.f.
		Continuous variable	
Gaussian	$\mathbb{R}$	General	$(2\pi\sigma^2)^{-1/2}\exp\left\{-(x-\mu)^2/2\sigma^2\right\}$
Student	$\mathbb{R}$	Heavy tailed	$(\nu\pi)^{-1/2}\Gamma(\nu/2)^{-1}\Gamma\{(\nu+1)/2)\left(1+x^2/\nu\right)^{-(\nu+1)/2}$
Log-normal	$(0,\infty)$	Positive	$(2\pi\sigma^2)^{-1/2}x^{-1}\exp\left\{-(\log x-\mu)^2/2\sigma^2\right\}$
Exponential	$(0,\infty)$	Duration	$\lambda\exp(-\lambda x)$
Weibull	$(0,\infty)$	Duration	$\kappa\lambda^{-\kappa}x^{\kappa-1}\exp\{-(x/\lambda)^\kappa\}$
Beta	$(0,1)$	Bounded	$B(\alpha,\beta)^{-1}x^{\alpha-1}(1-x)^{\beta-1}$
		Discrete variable	
Bernoulli	$\{0,1\}$	Binary	$p^x(1-p)^{1-x}$
Binomial	$\{0,\dots,n\}$	# success	$\binom{n}{x}p^x(1-p)^{n-x}$
Geometric	$\mathbb{N}_*$	# attempt	$p(1-p)^{x-1}$
Poisson	$\mathbb{N}$	Counts	$\lambda^x\exp(-\lambda)/x!$
Categorical	$\{1,\dots,k\}$	Factor	$p_j, j=1,\dots,k$

Example: FC Nantes scoring abilities

We are interesting in modelling the number of goals scored by FC Nantes—or your favourite football team.

Example: FC Nantes scoring abilities

We are interesting in modelling the number of goals scored by FC Nantes—or your favourite football team.

Since the number of goals is a count a sensible statistical model may be the Poisson distribution

$$N_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda), \qquad i = 1, \ldots, n,$$

where $\lambda > 0$ is the unknown parameter to be estimated from data.

Example: FC Nantes scoring abilities

We are interesting in modelling the number of goals scored by FC Nantes—or your favourite football team.

FC NANTES

Since the number of goals is a count a sensible statistical model may be the Poisson distribution

$$N_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda), \qquad i = 1, \ldots, n,$$

where $\lambda > 0$ is the unknown parameter to be estimated from data.

☞ If you want to show off a bit, you can even invoke the law of rare events

$$\text{Binomial}(n, p_n) \xrightarrow{\text{d.}} \text{Poisson}(\lambda), \qquad n \to \infty, \quad np_n \to \lambda.$$

Ligue 1 dataset

```
    Div        Date   Time    HomeTeam  ... MaxCAHH  MaxCAHA  AvgCAHH AvgCAHA
0    F1  06/08/2021  20:00      Monaco  ...    2.03     1.99     1.97    1.89
1    F1  07/08/2021  16:00        Lyon  ...    2.00     1.94     1.96    1.89
2    F1  07/08/2021  20:00      Troyes  ...    2.04     2.00     1.91    1.95
3    F1  08/08/2021  12:00      Rennes  ...    1.94     2.00     1.91    1.95
4    F1  08/08/2021  14:00    Bordeaux  ...    1.89     2.10     1.84    2.03
..   ..         ...    ...         ...  ...     ...      ...      ...     ...
375  F1  21/05/2022  20:00     Lorient  ...    1.99     1.99     1.93    1.93
376  F1  21/05/2022  20:00   Marseille  ...    1.91     2.15     1.87    1.99
377  F1  21/05/2022  20:00      Nantes  ...    1.86     2.25     1.81    2.07
378  F1  21/05/2022  20:00    Paris SG  ...    2.05     2.25     1.85    2.01
379  F1  21/05/2022  20:00       Reims  ...    2.01     1.96     1.95    1.92

[380 rows x 105 columns]
```

The maximum likelihood estimator (sloppy)

□ Having observed independent copies $\mathbf{Y} = (Y_1, \dots, Y_n)$ we may want to fit our statistical model using the maximum likelihood estimator

$$\hat\theta = \arg\max_{\theta\in\Theta} \ell(\theta; \mathbf{Y}), \qquad \ell(\theta; \mathbf{Y}) = \sum_{i=1}^{n} \log f(Y_i; \theta)$$

□ Widely used in practice since (under regularity conditions) it is

- consistent and asymptotically efficient
- widely applicable and versatile
- rather straightforward to implement

□ With loose notations, and provided the sample size n is large enough,

$$\hat\theta \dot\sim N\left(\theta_*, \Sigma_n\right), \qquad \Sigma_n = -\left\{\nabla^2 \ell(\hat\theta; \mathbf{Y})\right\}^{-1}.$$

The maximum likelihood estimator

Theorem 1. *Let* $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$, $n \geq 1$, *an* n*–sample of independent copies with p.d.f.* $f(\cdot;\theta_*)$*. Then, under regularity assumptions, the maximum likelihood estimator defined by*

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log f(Y_i;\theta)$$

satisfies

$$\sqrt{n}\left(\hat{\theta} - \theta_*\right) \xrightarrow{d.} N\left\{0, -H(\theta_*)^{-1}\right\}, \qquad n \to \infty,$$

where $H(\theta_*) = \mathbb{E}\{\nabla^2 \log f(X;\theta_*)\}$*.*

Proof. Taylor expansion + CLT + Slutsky □

Application: FC Nantes scoring

Exercise 1. Based on a sample $X_1, \dots, X_n$, compute the MLE for a Poisson model. What is the (approximate) distribution for this estimator? Apply your results to the Ligue 1 data set.

Model checking

- □ Fitting a model is not enough, we have to check if our fitted model is actually good. It is model checking.
- □ One can use numerical quantities such as overall error, but if possible, graphical model checking has to be preferred
- □ Briefly the idea is to compare observations to predictions from the fitted model.
- □ Two cases arise:

Discrete Compare the empirical p.m.f. to the fitted one;
Continuous Produce a quantile-quantile plot

Application: FC Nantes

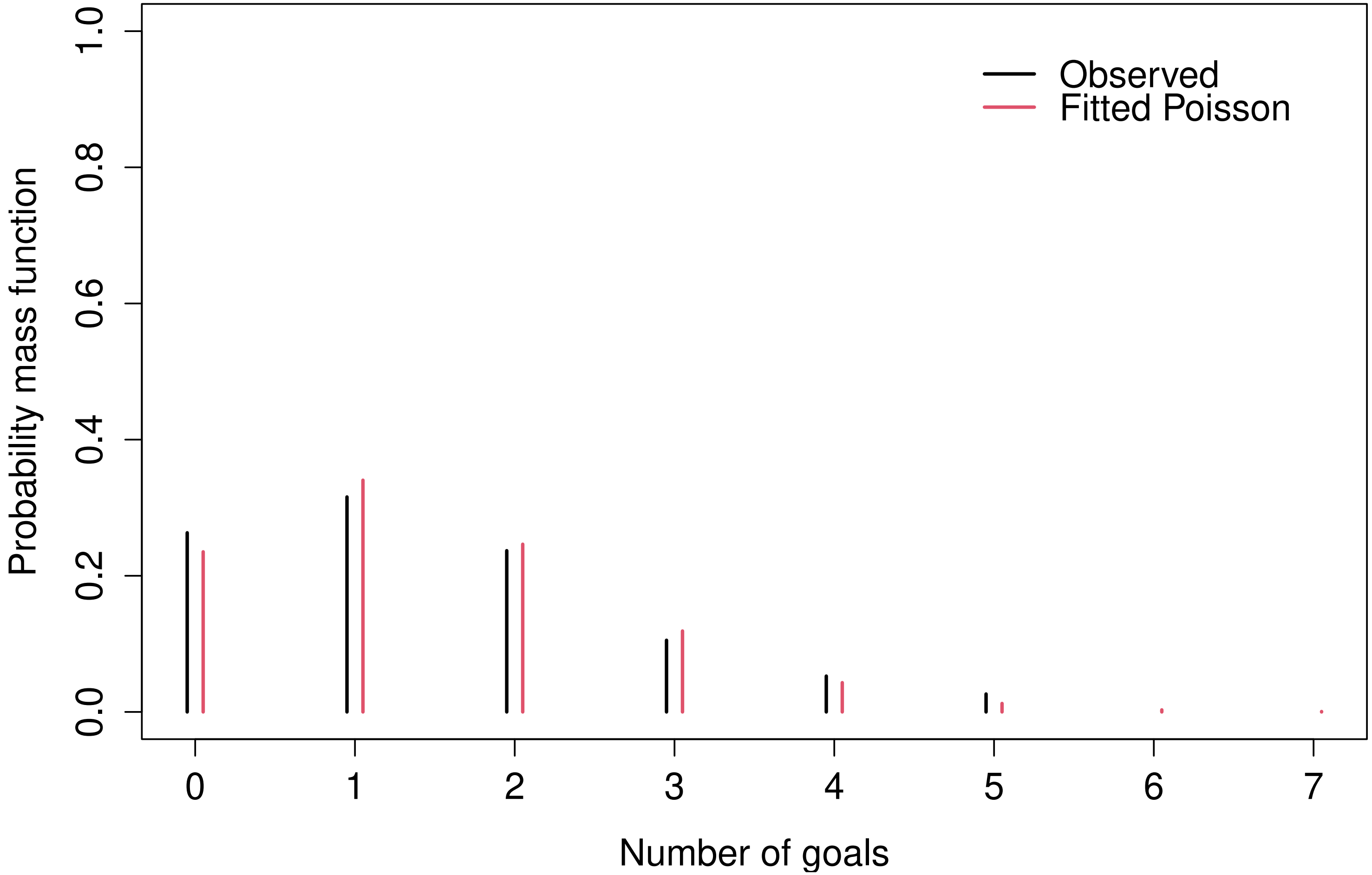


Figure 9: *Comparison of the empirical probability mass function and that from our fitted Poisson model.*

Standard errors

- □ Having an estimate of the parameter θ is not enough.
- □ It is (very!) good practice to show its respective standard errors

$$\text{std err}(\theta) = \sqrt{\text{Var}(\hat{\theta})}, \qquad \text{for some univariate parameter } \theta.$$

- □ Standard errors measure how precise is your estimate, i.e., the lower, the better.

Standard errors

- □ Having an estimate of the parameter θ is not enough.
- □ It is (very!) good practice to show its respective standard errors

$$\text{std err}(\theta) = \sqrt{\text{Var}(\hat{\theta})}, \qquad \text{for some univariate parameter } \theta.$$

- □ Standard errors measure how precise is your estimate, i.e., the lower, the better.
- □ Going back to our MLE properties, i.e., $\hat{\theta} \dot{\sim} N\left(\theta_*, \Sigma_n\right)$ where $\Sigma_n = -\left\{\nabla^2 \ell(\hat{\theta}; \mathbf{Y})\right\}^{-1}$, we conclude that standard errors are thus the square root of the diagonal elements of Σ_n.

Standard errors

- □ Having an estimate of the parameter θ is not enough.
- □ It is (very!) good practice to show its respective standard errors

$$\text{std err}(\theta) = \sqrt{\text{Var}(\hat{\theta})}, \qquad \text{for some univariate parameter } \theta.$$

- □ Standard errors measure how precise is your estimate, i.e., the lower, the better.
- □ Going back to our MLE properties, i.e., $\hat{\theta} \dot{\sim} N(\theta_*, \Sigma_n)$ where $\Sigma_n = -\left\{\nabla^2 \ell(\hat{\theta}; \mathbf{Y})\right\}^{-1}$, we conclude that standard errors are thus the square root of the diagonal elements of Σ_n.

☞ Most numerical optimizers can output Σ_n (or Σ_n^{-1}) so standard errors can easily be computed (and you have no excuse!).

Confidence intervals

Definition 5. A confidence interval of level $\alpha \in (0,1)$ for some unknown quantity $\theta_* \in \Theta$ is an interval $I_\alpha \subset \Theta$ such that $\Pr(\theta_* \in I_\alpha) = \alpha$.
Note that I_α is computed only from the sample $X_1, \dots, X_n$ and is therefore a random interval.

□ Confidence intervals can be:

approximate in which case $\Pr(\theta_* \in I_\alpha) \geq \alpha$;
asymptotic in which case $\Pr(\theta_* \in I_\alpha) \to \alpha$ as $n \to \infty$.

☞ Using the asymptotic properties of the MLE, i.e., $\hat{\theta} \dot{\sim} N\left(\theta_*, \Sigma_n\right)$, a (symmetric) asymptotic confidence interval for θ_* is

$$\left[\hat{\theta} - \text{std. err.}(\hat{\theta}) z_{1-(1-\alpha)/2}; \hat{\theta} + \text{std. err.}(\hat{\theta}) z_{1-(1-\alpha)/2}\right],$$

where z_p is the quantile of a $N(0,1)$ of order p.

Beware

□ Confidence intervals are often misinterpreted

□ A wrong interpretation will be to say that

> "The true parameter θ_* belongs to **this** confidence interval with probability α. "👎

□ The right interpretation is rather

> "If we were to replicate our experiment N times independently, i.e., Bernoulli experiments, we will thus have N independent confidence intervals and
>
> $$\frac{1}{N}\sum_{j=1}^{N} 1_{\{\theta_* \text{ belongs to the } j\text{-th confidence interval}\}} \xrightarrow{\text{a.s.}} \alpha, \qquad N \to \infty.$$
>
> "👍

Beware

- Confidence intervals are often misinterpreted
- A wrong interpretation will be to say that

> “The true parameter θ_* belongs to **this** confidence interval with probability α. ”👎

- The right interpretation is rather

> “If we were to replicate our experiment N times independently, i.e., Bernoulli experiments, we will thus have N independent confidence intervals and
>
> $$\frac{1}{N}\sum_{j=1}^{N} 1_{\{\theta_* \text{ belongs to the } j\text{-th confidence interval}\}} \xrightarrow{\text{a.s.}} \alpha, \qquad N \to \infty.$$
>
> ”👍

The first interpretation is that of **credible intervals** and refer to Bayesian statistics.

Application: Hold your breath

Exercise 2. Give a 95% (symmetric) confidence interval for the parameter of the Poisson distribution.

3. K–means

Homework

- □ Get the book `An introduction to Statistical Learning with Applications in R` from this link
- □ Read section 12.4.1 and do the lab of Section 12.5.3

- □ 3 wine makers
- □ 178 italian wines
- □ 13 variables

```
    Alcohol Malic  Ash Alcalinity Magnesium Phenols
48    13.90  1.68 2.12       16.0       101    3.10
66    12.37  1.21 2.56       18.1        98    2.42
101   12.08  2.08 1.70       17.5        97    2.23
159   14.34  1.68 2.70       25.0        98    2.80
36    13.48  1.81 2.41       20.5       100    2.70
156   13.17  5.19 2.32       22.0        93    1.74
    Flavanoids Nonflavanoid Proanthocyanins Color  Hue
48        3.39         0.21            2.14   6.1 0.91
66        2.65         0.37            2.08   4.6 1.19
101       2.17         0.26            1.40   3.3 1.27
159       1.31         0.53            2.70  13.0 0.57
36        2.98         0.26            1.86   5.1 1.04
156       0.63         0.61            1.55   7.9 0.60
    OD280/OD315 of diluted wines Proline
48                          3.33     985
66                          2.30     678
101                         2.96     710
159                         1.96     660
36                          3.47     920
156                         1.48     725
```

K-means

☞ The k–means measures homogeneity using the euclidean distance denoted $d(x,y) = \|x - y\|$.

K-means

☞ The k–means measures homogeneity using the euclidean distance denoted $d(x, y) = \|x - y\|$.

☞ Computing $\|x_i - x_j\|^2$ must thus be sensible:

- quantitatives variables → OK
- categorical variables → KO[3]

☞ The variables must have the same order of magnitude and if not need to work on a scaled version

[3] Well unless you use one-hot encoding but. . .

Optimization problem

We aim at finding K groups that are the most homogeneous using the Euclidean norm, i.e.,

Optimization problem

We aim at finding K groups that are the most homogeneous using the Euclidean norm, i.e.,

$$\pi^* = \underset{\pi \in \mathscr{P}(n,K)}{\arg\min} \frac{1}{2n} \sum_{k=1}^{K} \underbrace{\sum_{i,j=1}^{n} \|x_i - x_j\|^2 1_{\{\pi(i)=\pi(j)=k\}}}_{\text{homogenity of class } k},$$

where $\mathscr{P}(n, K)$ is the set of partitions of n elements from K labels.

Optimization problem

We aim at finding K groups that are the most homogeneous using the Euclidean norm, i.e.,

$$\pi^* = \operatorname*{arg\,min}_{\pi \in \mathscr{P}(n,K)} \frac{1}{2n} \sum_{k=1}^{K} \underbrace{\sum_{i,j=1}^{n} \|x_i - x_j\|^2 1_{\{\pi(i)=\pi(j)=k\}}}_{\text{homogenity of class } k},$$

where $\mathscr{P}(n, K)$ is the set of partitions of n elements from K labels.

☞ OK that's just a discrete (or combinatorial) optimization problem since $\mathscr{P}(n, K)$ is finite! Easy!

Optimization problem

We aim at finding K groups that are the most homogeneous using the Euclidean norm, i.e.,

$$\pi^* = \underset{\pi \in \mathscr{P}(n,K)}{\arg\min} \frac{1}{2n} \sum_{k=1}^{K} \underbrace{\sum_{i,j=1}^{n} \|x_i - x_j\|^2 1_{\{\pi(i)=\pi(j)=k\}},}_{\text{homogenity of class } k}$$

where $\mathscr{P}(n, K)$ is the set of partitions of n elements from K labels.

☞ OK that's just a discrete (or combinatorial) optimization problem since $\mathscr{P}(n, K)$ is finite! Easy!

💣 Well no since $|\mathscr{P}(n, K)|$ induces a combinatorial burden,e.g., $S(11, 5) \approx 2.5 \times 10^5$. It is hopeless to get the global minimum and in practice we stick with a (rather good) local minimum!

LLoyd algorithm

Algorithm 1: Lloyd algorithm.

```
input  : A sample $x_1, \dots, x_n$, number of urns $K$, maximal number of iterations $T_{\text{max}}$,
         initial partitioning $\pi$.
output: An "optimal" partitioning $\pi$

for $t \leftarrow 1$ to $T_{max}$ do
    For each urn, compute its centroid, i.e.,;
        $$\mu_k = \frac{1}{N_k} \sum_{i:\ \pi(i)=k} x_i, \qquad k = 1, \dots, K, \quad N_k = \sum_{i=1}^{n} 1_{\{\pi(i)=k\}}.$$
    For each observation, place it into the urn of the closest centroid, i.e.,
        $$\pi(i) = \underset{k \in \{1,\dots,K\}}{\arg\min} \|x_i - \mu_k\|^2.$$
    if The partitioning $\pi$ has not changed then
        Go outside the loop;
return $\pi$;
```

Application to the Fisher's Iris data

Data 150 measures of length and width of Iris sepals and petals.

Objective Find the Iris species, i.e., `setosa`, `versicolor` or `virginica`.

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width ## <<- I'm lying ;-)
1          5.1         3.5          1.4         0.2
2          4.9         3.0          1.4         0.2
3          4.7         3.2          1.3         0.2
4          4.6         3.1          1.5         0.2
5          5.0         3.6          1.4         0.2
6          5.4         3.9          1.7         0.4
...
```

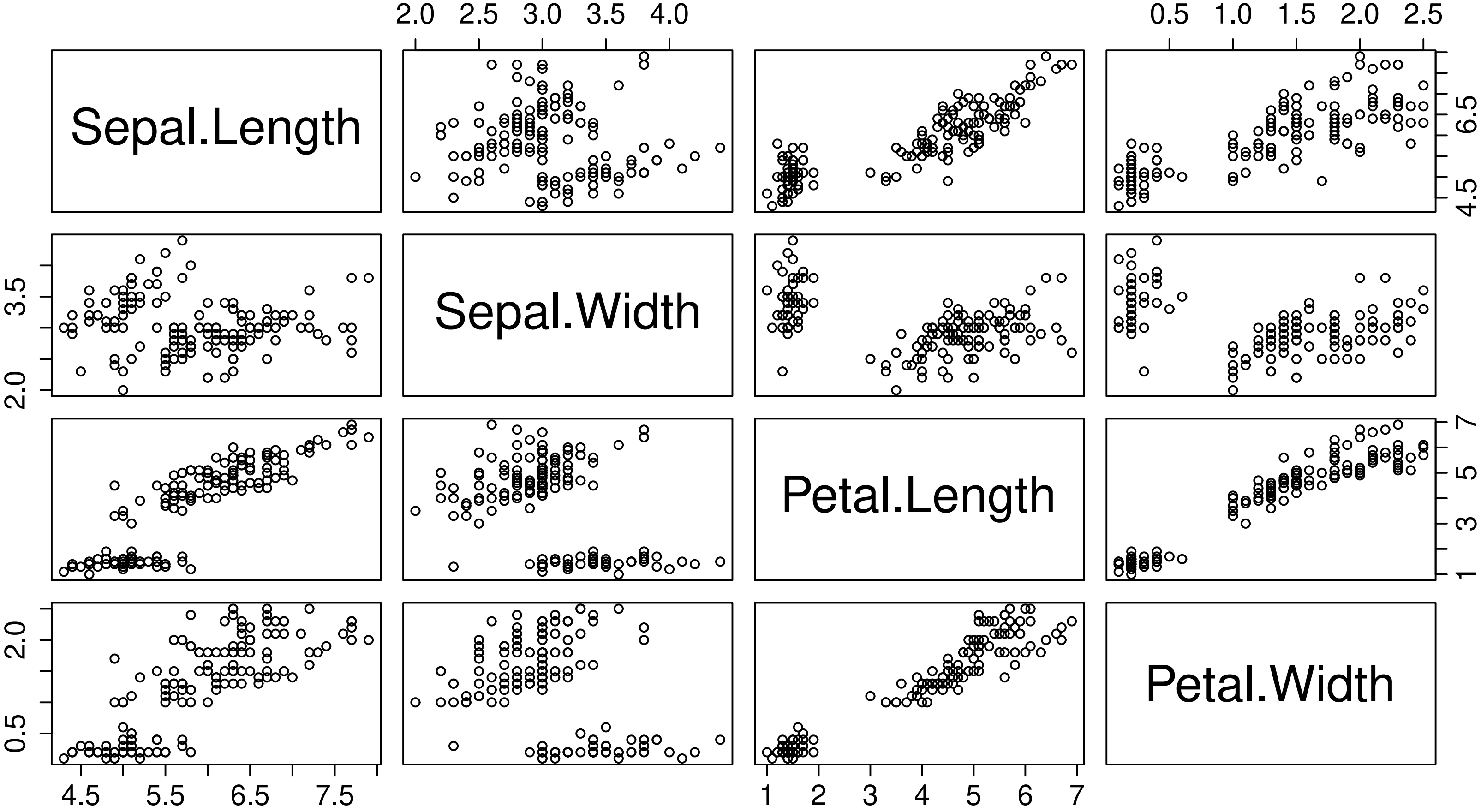


Figure 10: *Scatterplot of the iris dataset.*

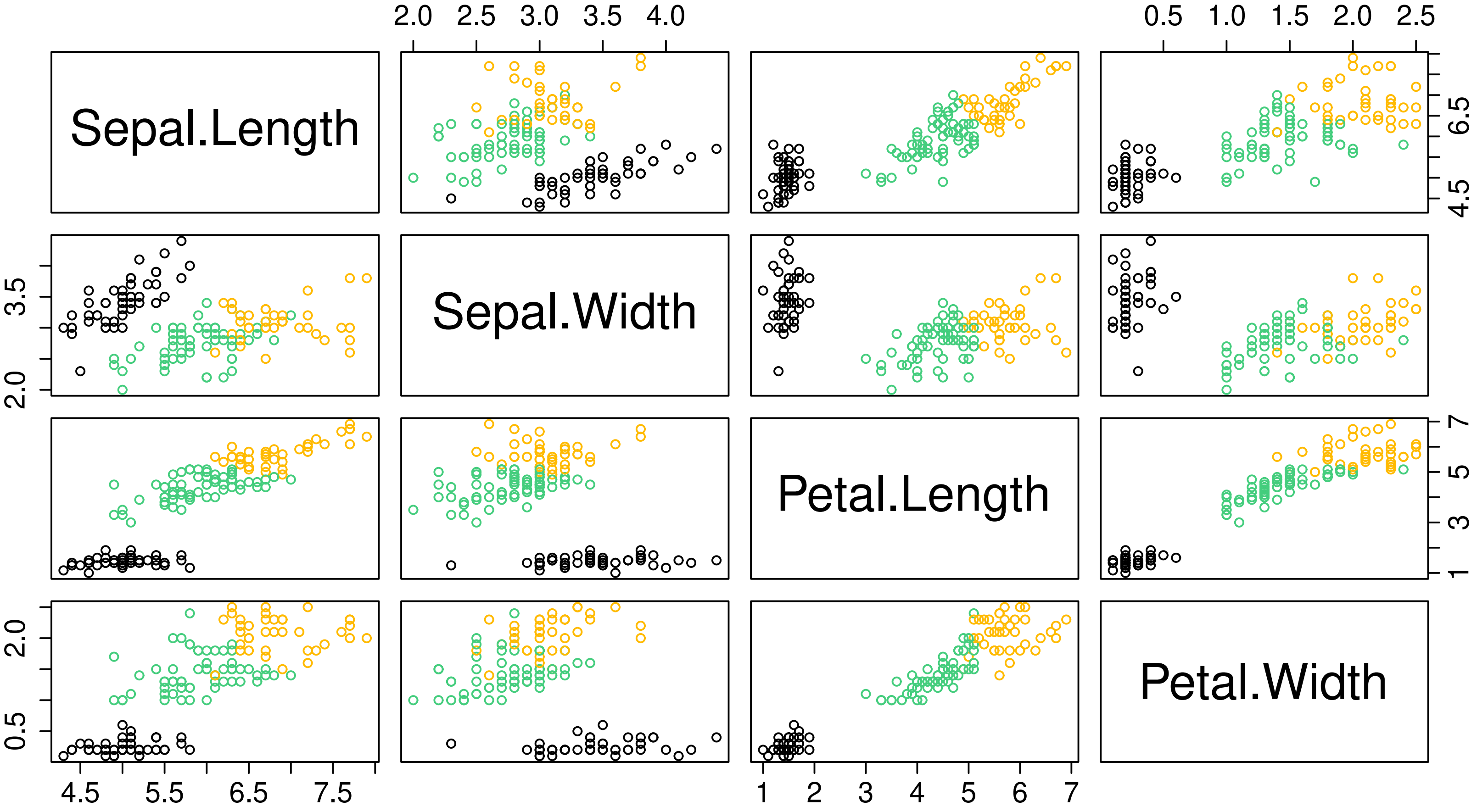


Figure 10: *Scatterplot of the iris dataset.*

Is this a good clustering?

- □ Looking at the previous plots, we may feel rather happy. . .
- □ But it is a bit subjective. What about a more formal way to asses it?
  - Inertia
  - Confusion matrix (if supervised)

Inertia

Definition 6. Consider the following cloud of points $\mathbf{x} = (x_1, \dots, x_n)$, i.e., our observations. The inertia (for the Euclidean norm $\| \cdot \|$) of these points is given by

$$I(\mathbf{x}) = \frac{1}{2n} \sum_{i,j=1}^{n} \|x_i - x_j\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \bar{x}\|^2, \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

It is a dispersion measure of the scatter plot.

Inertia

Definition 6. Consider the following cloud of points $\mathbf{x} = (x_1, \ldots, x_n)$, i.e., our observations. The inertia (for the Euclidean norm $\|\cdot\|$) of these points is given by

$$I(\mathbf{x}) = \frac{1}{2n} \sum_{i,j=1}^{n} \|x_i - x_j\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \bar{x}\|^2, \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

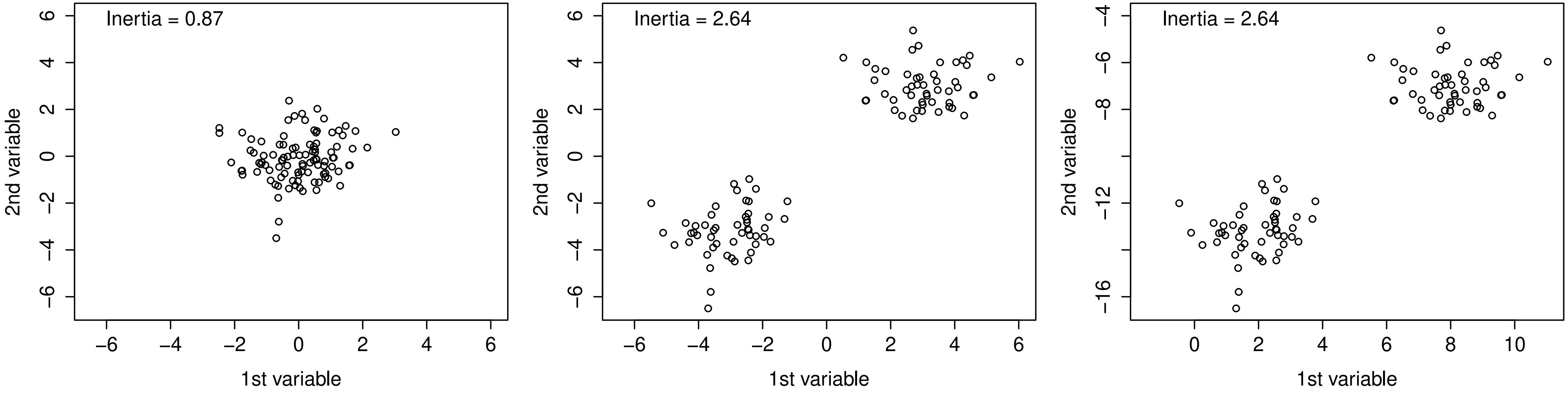


Figure 11: *Inertia computed on three different cloud points.*

Within–Between decomposition: Huygens formula

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a cloud point and π a clustering of it using K classes. Then

$$\begin{aligned}
I(\mathbf{x}) &= \frac{1}{2n} \sum_{i,j=1}^{n} \|x_i - x_j\|^2 \\
&= \frac{1}{2n} \sum_{k=1}^{K} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \|x_i - x_j\|^2 1_{\{\pi(j)=k\}} + \sum_{j=1}^{n} \|x_i - x_j\|^2 1_{\{\pi(j)\neq k\}} \right) 1_{\{\pi(i)=k\}} \\
&= W(\mathbf{x}, \pi) + B(\mathbf{x}, \pi)
\end{aligned}$$

where

$$W(\mathbf{x}, \pi) = \frac{1}{2n} \sum_{k=1}^{K} \sum_{i,j=1}^{n} \|x_i - x_j\|^2 1_{\{\pi(i)=\pi(j)=k\}} \qquad \text{(within)}$$

$$B(\mathbf{x}, \pi) = \frac{1}{2n} \sum_{k=1}^{K} \sum_{i,j=1}^{n} \|x_i - x_j\|^2 1_{\{\pi(i)=k, \pi(j)\neq k\}} \qquad \text{(between)}$$

Why is this useful?

$$I(\mathbf{x}) = W(\mathbf{x}, \pi) + B(\mathbf{x}, \pi)$$

□ The inertia $I(\mathbf{x})$ is independent of the clustering π

□ Our k–means aims at finding π^* minimizing $\pi \mapsto W(\mathbf{x}, \pi)$.

Why is this useful?

$$I(\mathbf{x}) = W(\mathbf{x}, \pi) + B(\mathbf{x}, \pi)$$

□ The inertia $I(\mathbf{x})$ is independent of the clustering π

□ Our k–means aims at finding π^* minimizing $\pi \mapsto W(\mathbf{x}, \pi)$.

☞ Equivalently we aim at maximizing $B(\mathbf{x}, \pi)$ than can be used as a measure of "goodness of clustering"

$$\frac{B(\mathbf{x}, \pi)}{I(\mathbf{x})} \in [0, 1], \qquad \text{the closest to 1, the better!}$$

Why is this useful?

$$I(\mathbf{x}) = W(\mathbf{x}, \pi) + B(\mathbf{x}, \pi)$$

- □ The inertia $I(\mathbf{x})$ is independent of the clustering π
- □ Our k–means aims at finding π^* minimizing $\pi \mapsto W(\mathbf{x}, \pi)$.

☞ Equivalently we aim at maximizing $B(\mathbf{x}, \pi)$ than can be used as a measure of "goodness of clustering"

$$\frac{B(\mathbf{x}, \pi)}{I(\mathbf{x})} \in [0, 1], \qquad \text{the closest to 1, the better!}$$

Remark. Note that

$$W(\mathbf{x}, \pi) = \frac{1}{n} \sum_{k=1}^{K} n_k \underbrace{\frac{1}{2n_k} \sum_{i,j=1}^{n} \|x_i - x_j\|^2 1_{\{\pi(i)=\pi(j)=k\}}}_{W_k(\mathbf{x},\pi)=\text{Inertia of class } k}, \qquad n_k = \sum_{i=1}^{n} 1_{\{\pi(i)=k\}}.$$

Prediction

- Once the clustering is done, one may want to describe each cluster. . .

Prediction

- □ Once the clustering is done, one may want to describe each cluster...
- □ ...but we can also do prediction for any new observation!
- □ Let x_* be a new observation. We will set the label of x_* to that for which its centroid is closest, i.e.,

$$\arg\min_{k \in \{1,\dots,K\}} \|x_* - \mu_k\|^2.$$

Prediction

- □ Once the clustering is done, one may want to describe each cluster…
- □ …but we can also do prediction for any new observation!
- □ Let x_* be a new observation. We will set the label of x_* to that for which its centroid is closest, i.e.,

$$\arg\min_{k \in \{1,\dots,K\}} \|x_* - \mu_k\|^2.$$

☞ It is thus possible to predict the label continuously on the variable space. It corresponds to the Voronoï cells of germ $\mu_1, \dots, \mu_K$, i.e.,

$$\text{Voronoï}(\mu_k) = \{x \in \mathbb{R}^p \colon \|x - \mu_k\| \leq \|x - \mu_\ell\|, \ell = 1, \dots, K\}.$$

Illustration of Voronoï cells and prediction

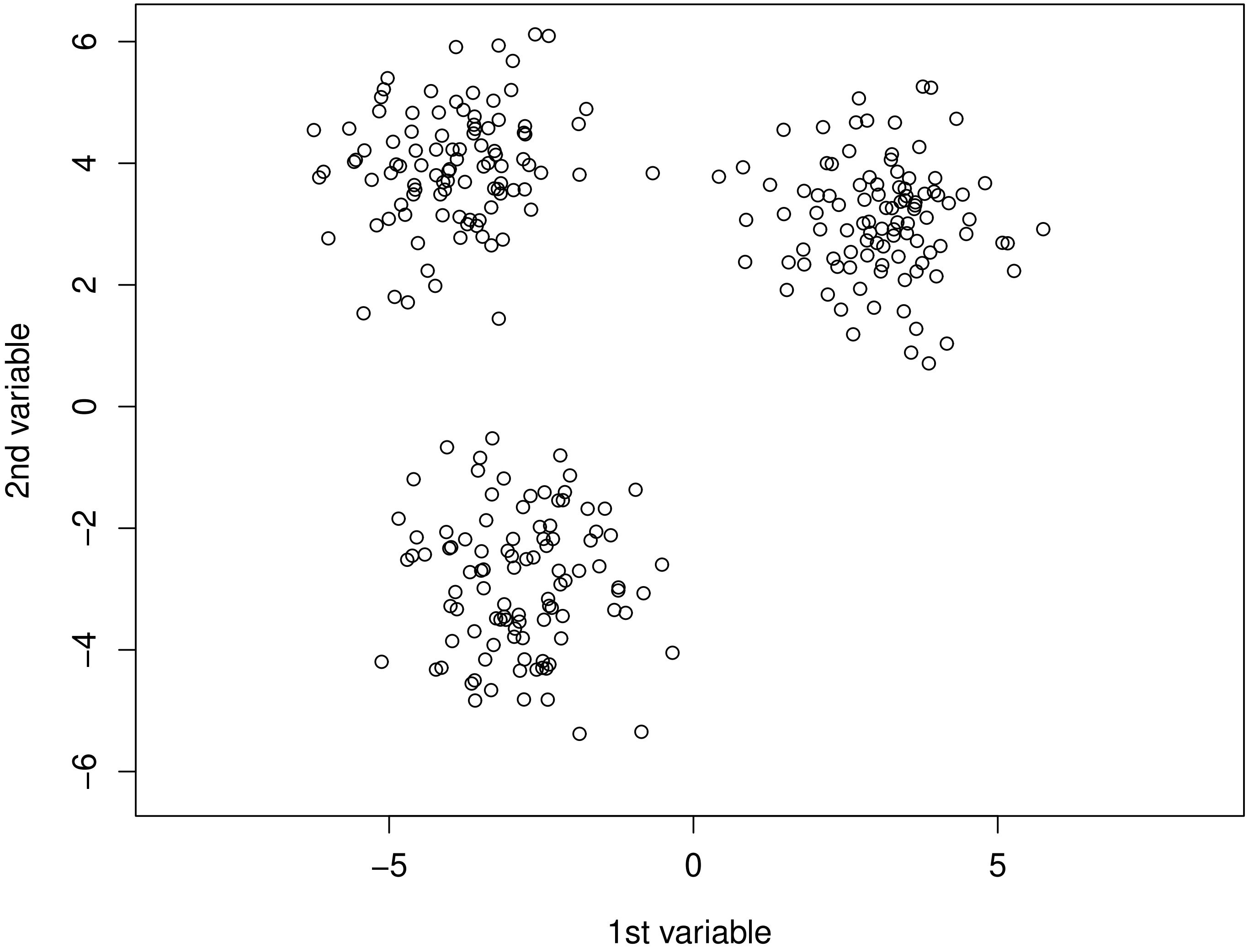


Figure 12: *Illustration of Voronoï cells and prediction.*

Illustration of Voronoï cells and prediction

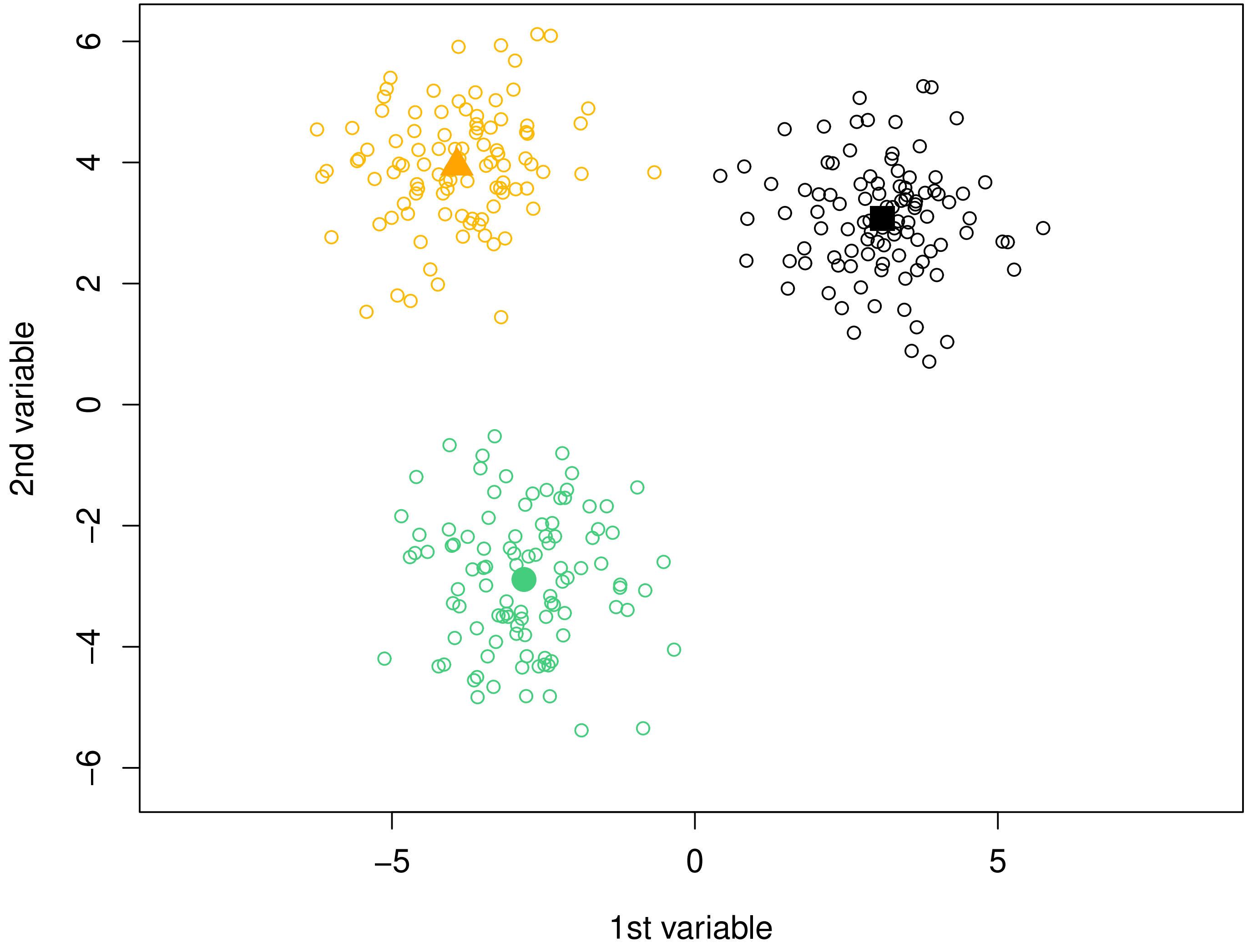


Figure 12: *Illustration of Voronoï cells and prediction.*

Illustration of Voronoï cells and prediction

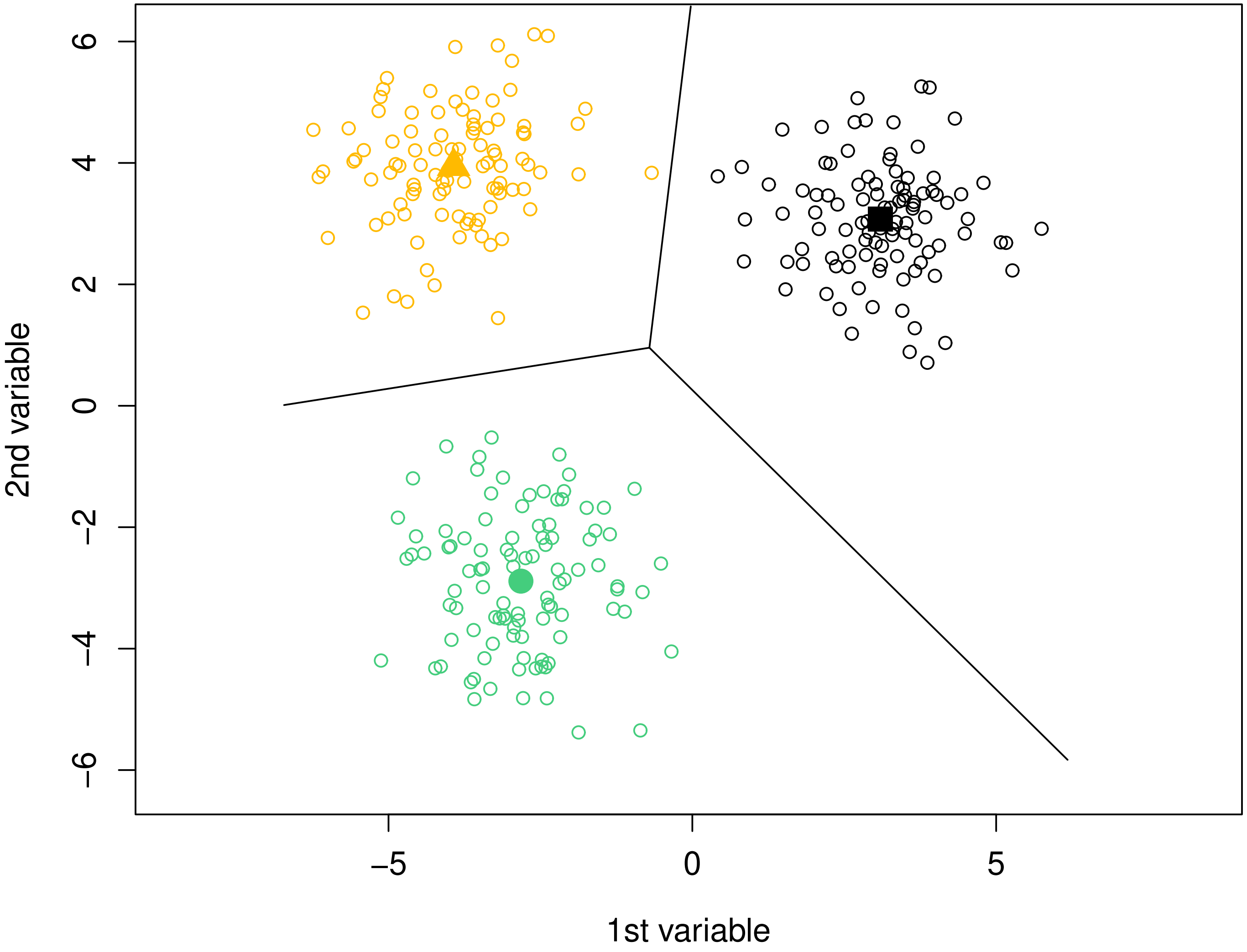


Figure 12: *Illustration of Voronoï cells and prediction.*

How many classes K?

- □ So far we consider that the number of classes was known ($K = 3$ for the iris dataset).
- □ In many situations we have no idea![4]
- □ How do we do?

How many classes K?

□ So far we consider that the number of classes was known ($K = 3$ for the iris dataset).

□ In many situations we have no idea![4]

□ How do we do? The idea is simple but efficient

1. Run multiple k-means with an increasing number of classes, e.g., $K = 2, \ldots, 10$.
2. Stick with the clustering such that adding one more class "doesn't bring nothing", i.e.,

$$\frac{B(\mathbf{x}, \pi)}{I(\mathbf{x})} \text{ doesn't increase much} \iff \frac{W(\mathbf{x}, \pi)}{I(\mathbf{x})} \text{ doesn't decrease much}$$

☞ It is known as the "elbow rule".

[4] Or it can be bad to set it to the number of "known classes", e.g.,MNIST.

Number of classes for the iris dataset

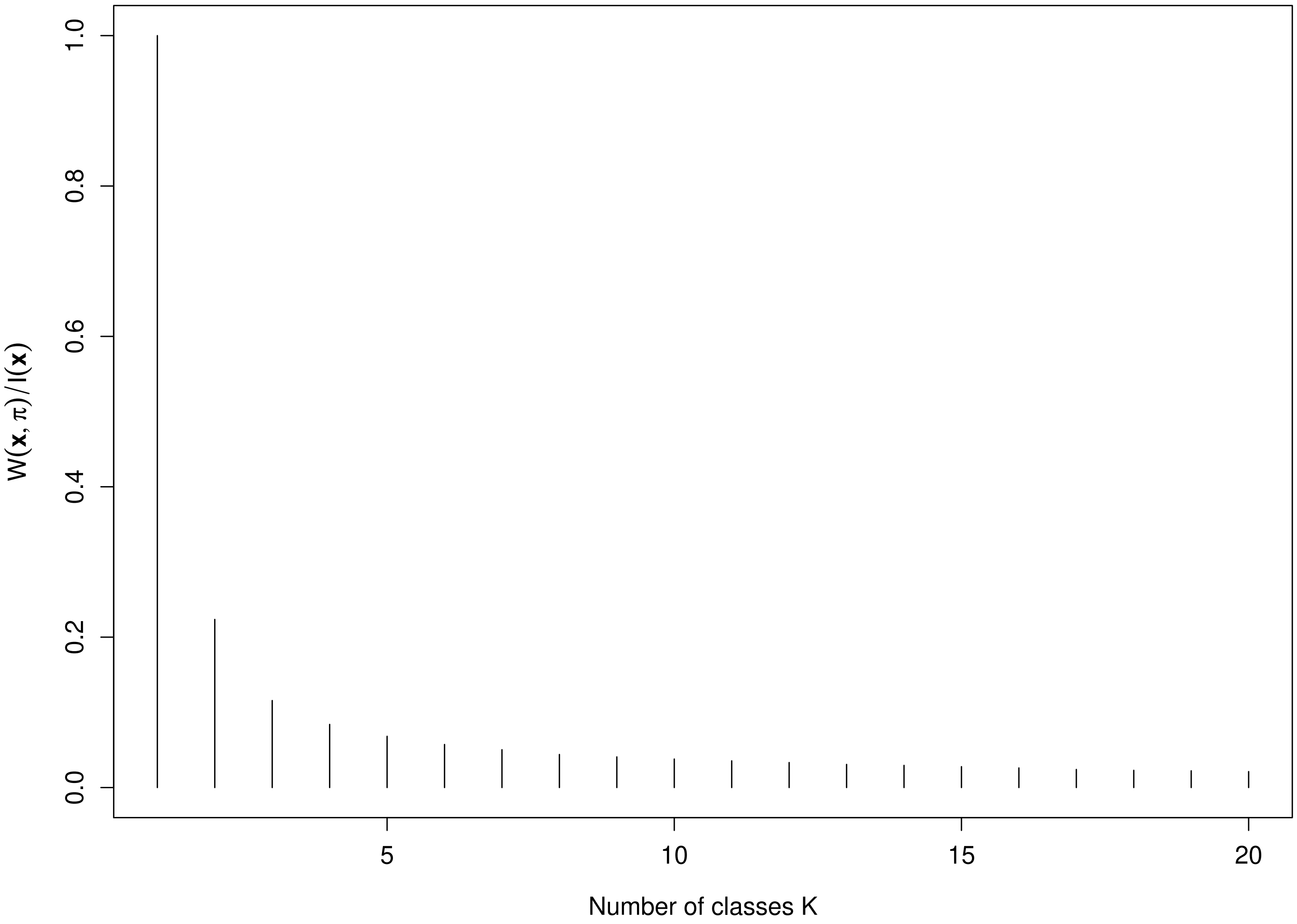


Figure 13: *Identify an appropriate number of classes using the "elbow rule". Here $K = 3$ or 4 seems to be appropriate (rather subjective I confess)*

Impact of initialization

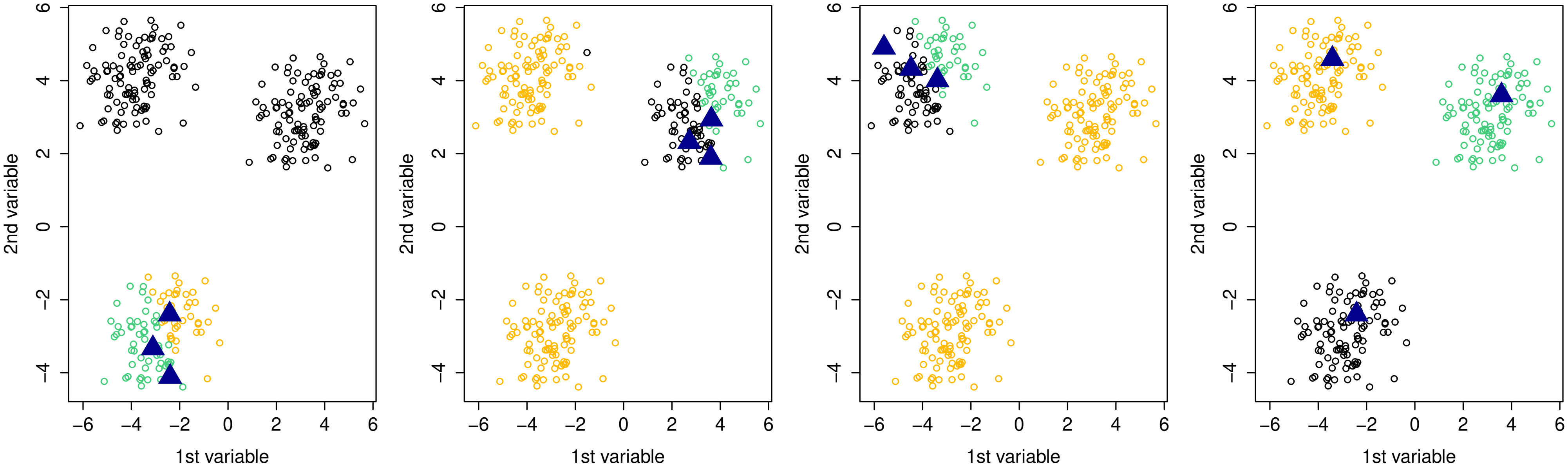


Figure 14: *Illustration on how sensitive is the k–means to initialization. Here 4 different initialization were used.*

Impact of initialization

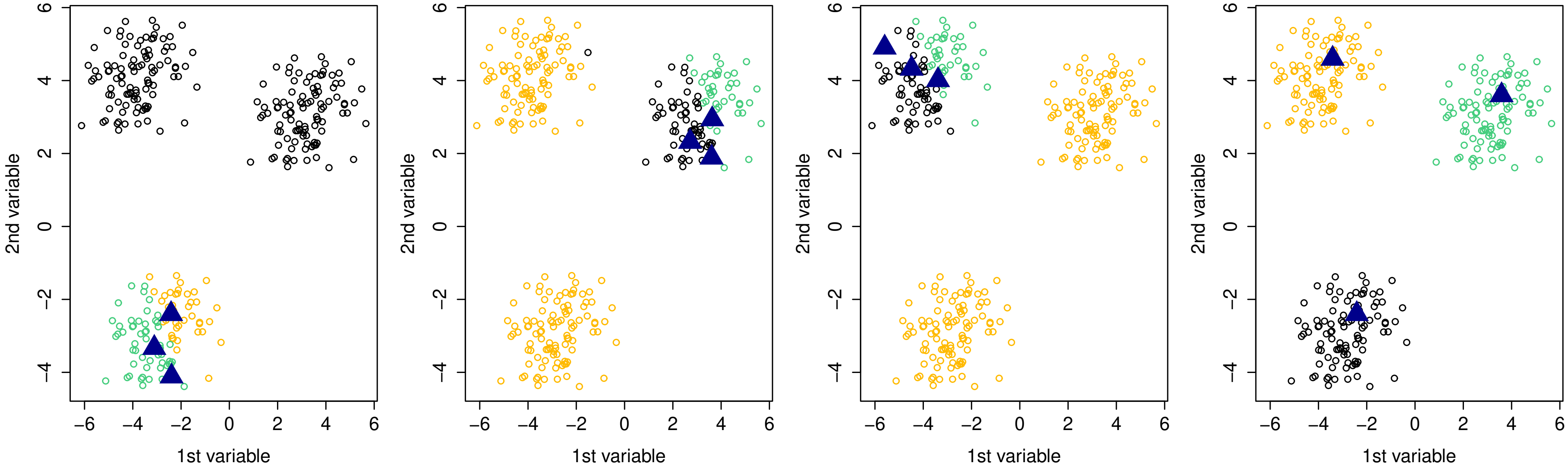


Figure 14: *Illustration on how sensitive is the k–means to initialization. Here 4 different initialization were used.*

☞ It is highly recommended to run several times the algorithm with different (random) initialization and keep the best clustering.

To sum up

Steps

- □ Center and scale the data (if necessary) since computations are based on the Euclidean norm;
- □ Let the number of class K vary and stick with the "best one";
- □ Analyze each class and/or do predictions.

Pros

- □ Scale well with large dimension, i.e., $n \gg 1$. Complexity is $O(nKT_{\text{max}})$[5];
- □ Easy and fast prediction.

Cons

- □ Implicit hypothesis of isotropy and balanced classes[6]
- □ Optimization problem: local minimum, initialization

[5] Since often T_{max} and K are small it is often said that it is a linear algorithm (in n)

[6] The k–means is actually a Gaussian mixture model with very specific assumptions...

4. Principal Component Analysis

Homework

- □ Get the book `An introduction to Statistical Learning with Applications in R` from this link
- □ Read sections 12.1 and 12.2 and ask for details if needed
- □ Work on the lab of Section 12.5

Motivation (1)

- □ Let $\mathbf{X} = (x_{ij} \colon i = 1, \dots, n, j = 1, \dots, p)$ be a data frame.
- □ This data frame is too big, i.e., $p \gg 1$, for what we about to do.
- □ We wish to get a more tractable version of $\mathbf{X}$ without too much loss.

Motivation (1)

□ Let $\mathbf{X} = (x_{ij} : i = 1, \dots, n, j = 1, \dots, p)$ be a data frame.

□ This data frame is too big, i.e., $p \gg 1$, for what we about to do.

□ We wish to get a more tractable version of $\mathbf{X}$ without too much loss.

☞ We need a framework to “compress” the data so that it scales to a following learning algorithm.

Motivation (2)

- □ Let $\mathbf{X} = (x_{ij} : i = 1, \dots, n, j = 1, \dots, p)$ be a data frame.
- □ It is our first time working with these data and there is a pressing need to get "familiarized" with them.
- □ One could be tempted to show pairwise scatterplots
- □ Since the number of pairs is $\binom{p}{2}$, it is hopeless. For example when $p = 10$ we have 45 plots!
- □ Further, it is likely that such plots are limited since dependencies typically involves more than a single variable.

Motivation (2)

- □ Let $\mathbf{X} = (x_{ij} : i = 1, \ldots, n, j = 1, \ldots, p)$ be a data frame.
- □ It is our first time working with these data and there is a pressing need to get "familiarized" with them.
- □ One could be tempted to show pairwise scatterplots
- □ Since the number of pairs is $\binom{p}{2}$, it is hopeless. For example when $p = 10$ we have 45 plots!
- □ Further, it is likely that such plots are limited since dependencies typically involves more than a single variable.

☞ We need a framework to "visualize" these data.

Way of proceeding

Idea Project the data frame $\mathbf{X}$ onto a lower dimensional sub-space.
Why?

a Ideally we aim at a “good” sub-space in a sense to be defined later;
lower To be able to visualize the data and/or use these “compressed” data frame in a subsequent analysis.

Way of proceeding

Idea Project the data frame $\mathbf{X}$ onto a lower dimensional sub-space.

Why?

a Ideally we aim at a “good” sub-space in a sense to be defined later;

lower To be able to visualize the data and/or use these “compressed” data frame in a subsequent analysis.

☞ Beware! From now we suppose that the data frame $\mathbf{X}$ is **centered and scaled**. Most often, software will do that for you.

Theorem 2 (Singular value decomposition)**.**
Let $\mathbf{X} \in \mathbb{C}^{n\times p}$ *be a matrix. There exists a triplet, known as the SVD,* $(U, D, V) \in \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times p} \times \mathbb{C}^{p\times p}$ *such that*

$$\mathbf{X} = UDV^\top,$$

where U *and* V *are orthogonal matrices and* $D = (d_{ij})$ *is such that*

$$d_{ij} = \begin{cases} \lambda_i, & i = j \\ 0, & i \neq j \end{cases}, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0, \quad k = \min(n, p).$$

λ_i *is called the* *i-th singular value.*

A convenient theorem

Definition 7. The Frobenius (matrix) norm, denoted $\|\cdot\|_F$, is given by $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\mathrm{Tr}(A^\top A)}, A \in \mathbb{R}^{n\times p}$.

(You can think about it a the usual ℓ_2 norm where A is now vectorized.)

Theorem 3 (Eckart–Young–Mirsky). *Let $\mathbf{X} \in \mathbb{C}^{n\times p}$ be a complex matrix and $r \in \{1, \dots, \min(n,p)\}$. The solution to the constrained optimization problem*

$$\arg\min_{M\in\mathbb{C}^{n\times p}} \|M - \mathbf{X}\|_F \quad \textit{such that rank}(M) \leq r$$

is given from the SVD of $\mathbf{X}$, denoted (U, D, V), truncated to the order r, i.e.,

$$M_* = U\tilde{D}V^T,$$

where $\tilde{D}$ is identical to D except that $\lambda_{r+1} = \cdots = \lambda_k = 0$.

A convenient theorem

Definition 7. The Frobenius (matrix) norm, denoted $\| \cdot \|_F$, is given by $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\mathrm{Tr}(A^\top A)}, A \in \mathbb{R}^{n \times p}$.

(You can think about it a the usual ℓ_2 norm where A is now vectorized.)

Theorem 3 (Eckart–Young–Mirsky)**.** *Let $\mathbf{X} \in \mathbb{C}^{n \times p}$ be a complex matrix and $r \in \{1, \ldots, \min(n,p)\}$. The solution to the constrained optimization problem*

$$\arg\min_{M \in \mathbb{C}^{n \times p}} \|M - \mathbf{X}\|_F \quad \textit{such that rank}(M) \leq r$$

is given from the SVD of $\mathbf{X}$, denoted (U, D, V), truncated to the order r, i.e.,

$$M_* = U\tilde{D}V^T,$$

where $\tilde{D}$ is identical to D except that $\lambda_{r+1} = \cdots = \lambda_k = 0$.

☞ The closest approximation of X (according to Frobenius norm) is the truncated SVD (with r small enough to help visualization/computation).

Amount of approximation

□ How to choose the cutoff value r?

Amount of approximation

□ How to choose the cutoff value r?

□ Let $\tilde{\mathbf{X}} = U\tilde{D}V^\top$ be the truncated SVD up to order $r \in \{1, \dots, k\}$.

□ The loss of information (according to the Frobenius norm) is

$$\sum_{j=r+1}^{k} \lambda_j^2.$$

□ Equivalently we say that the approximation $\tilde{\mathbf{X}}$ explains

$$100 \times \frac{\sum_{j=1}^{r} \lambda_j^2}{\sum_{j=1}^{k} \lambda_j^2} \%$$

of the variance // inertia.

Illustration (never used for image compression—non linearity of images)

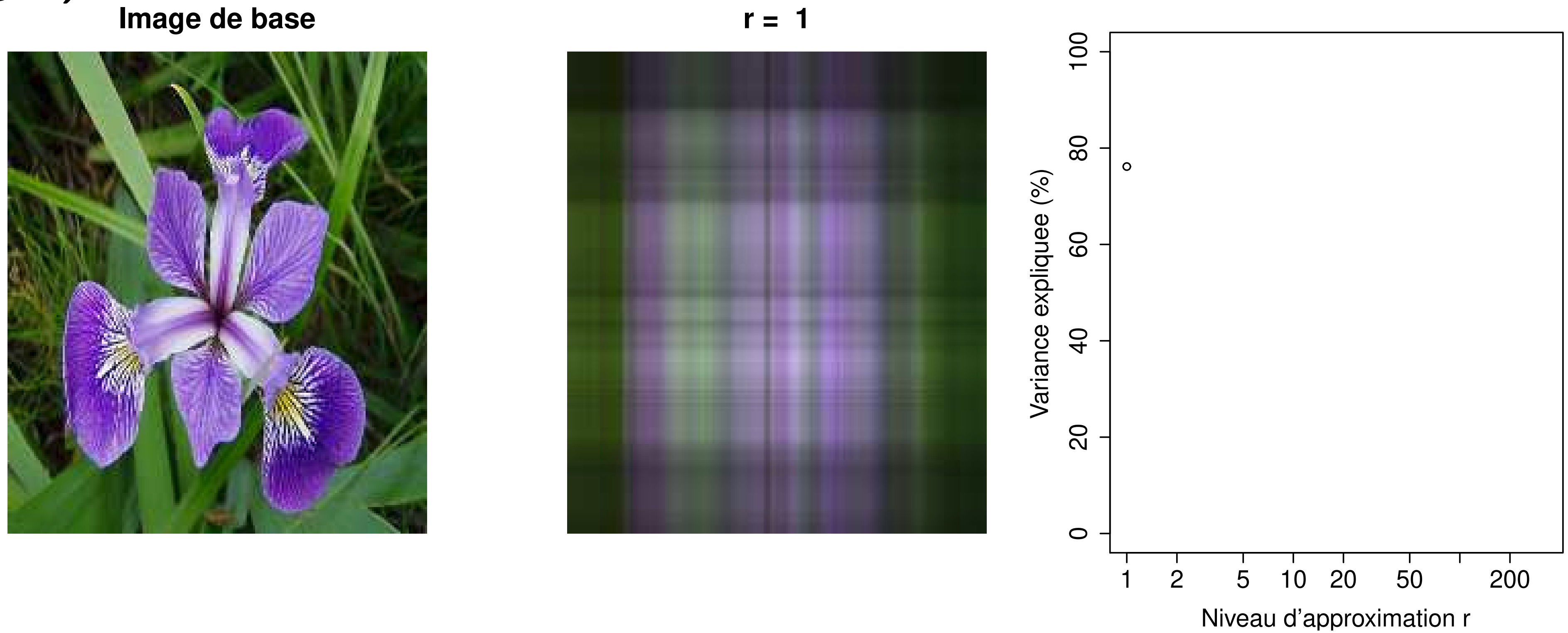


Figure 15: *Degree of approximation of the truncated SVD.*

Illustration (never used for image compression—non linearity of images)

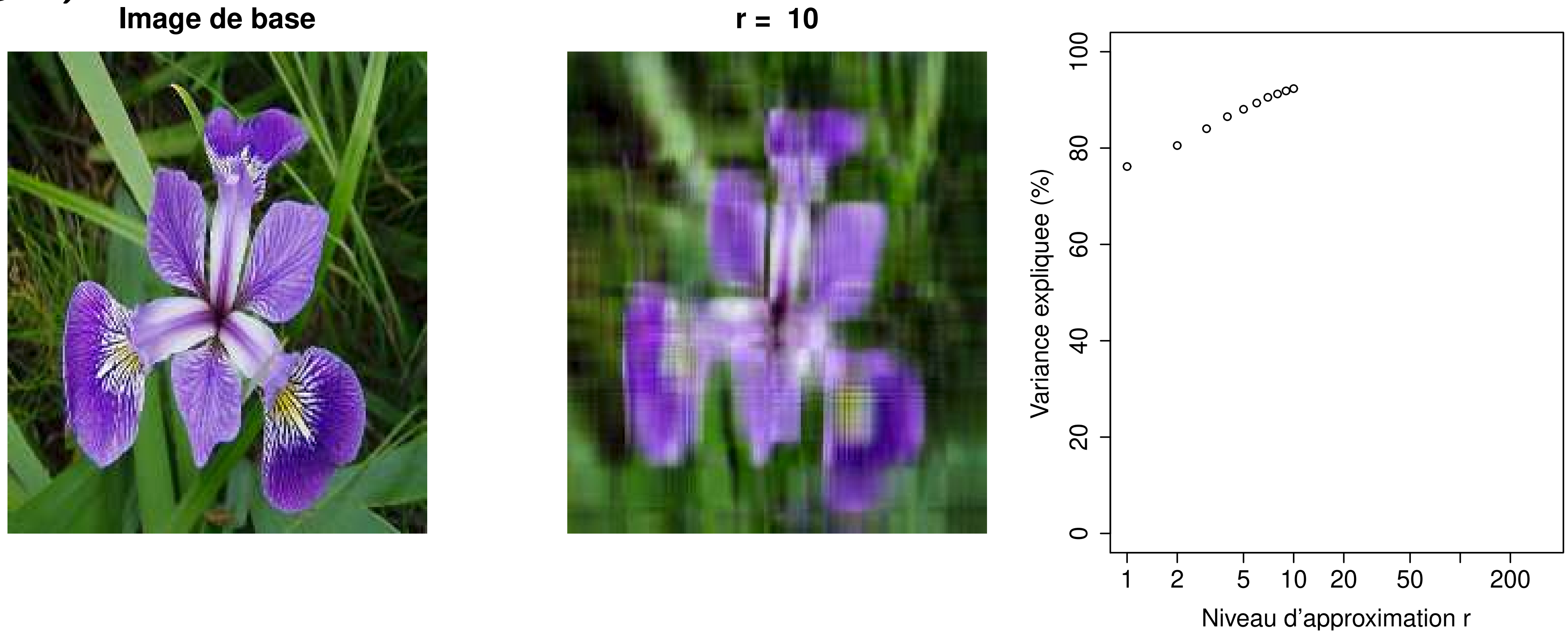


Figure 15: *Degree of approximation of the truncated SVD.*

Illustration (never used for image compression—non linearity of images)

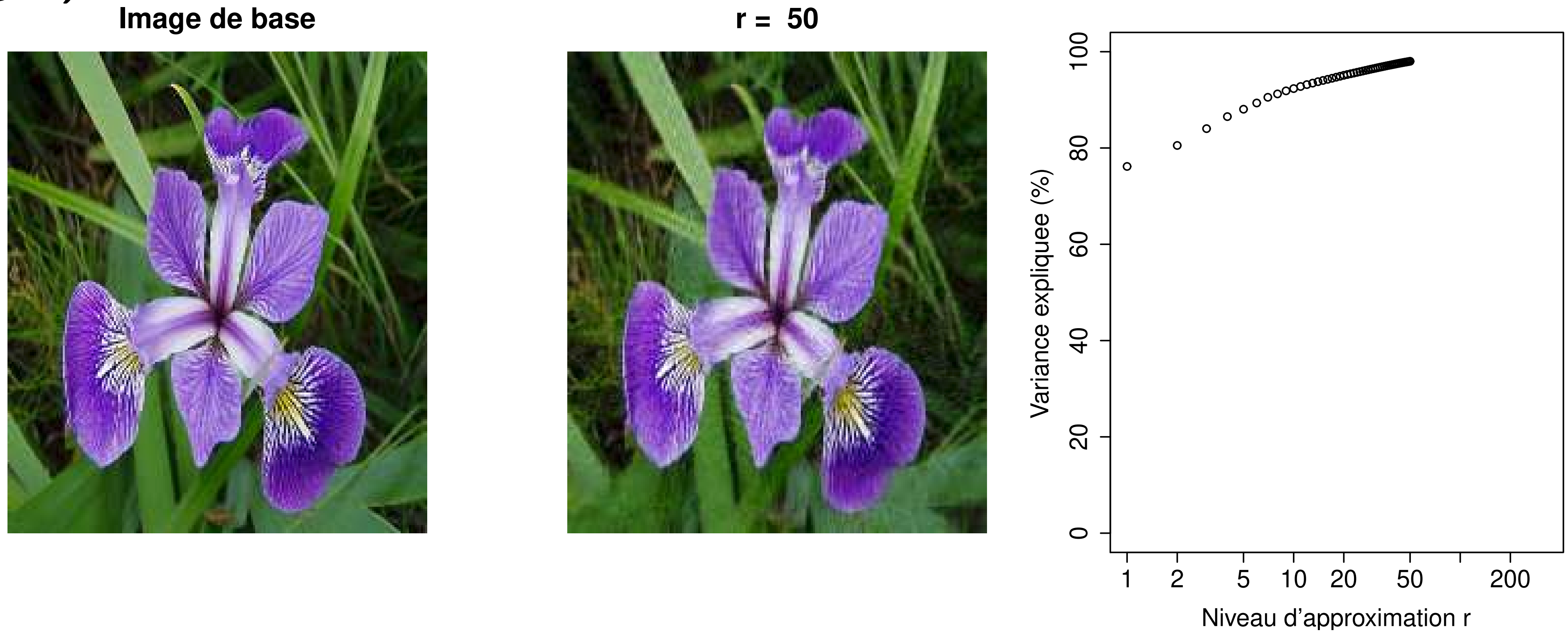


Figure 15: *Degree of approximation of the truncated SVD.*

Illustration (never used for image compression—non linearity of images)

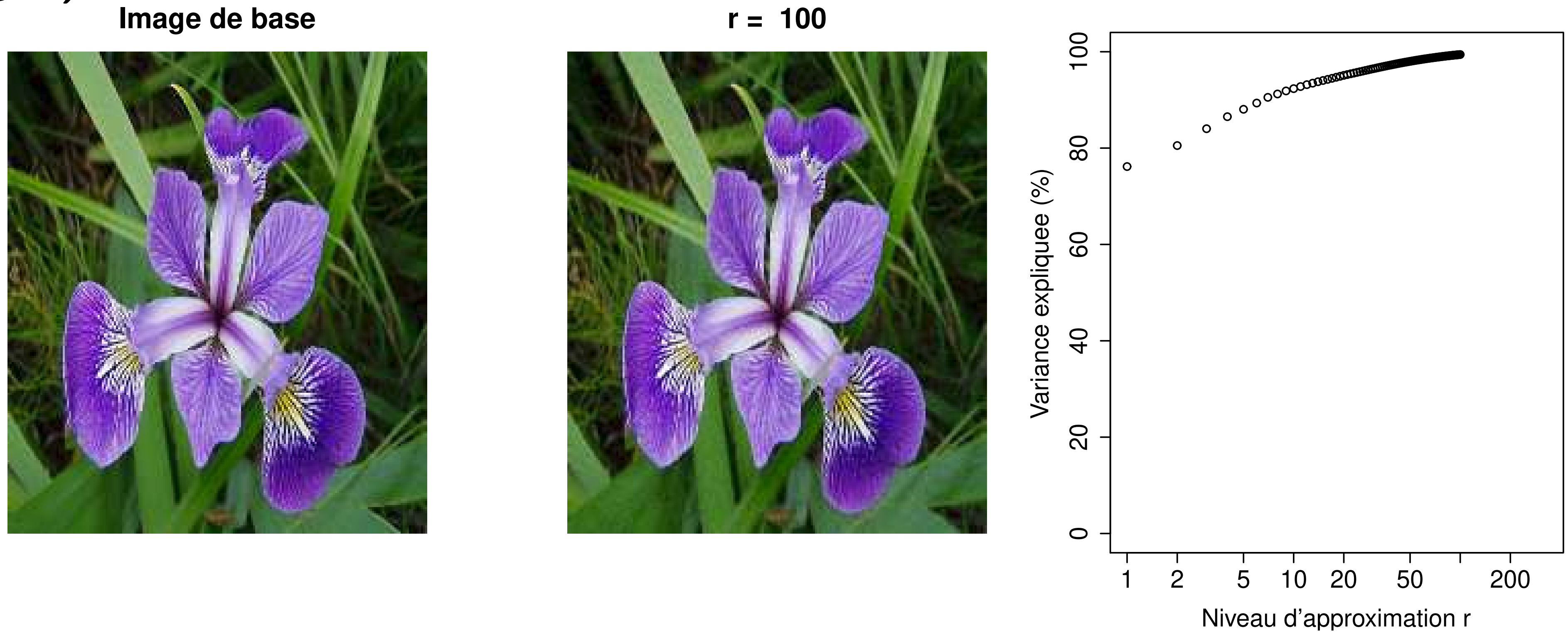


Figure 15: *Degree of approximation of the truncated SVD.*

Illustration (never used for image compression—non linearity of images)

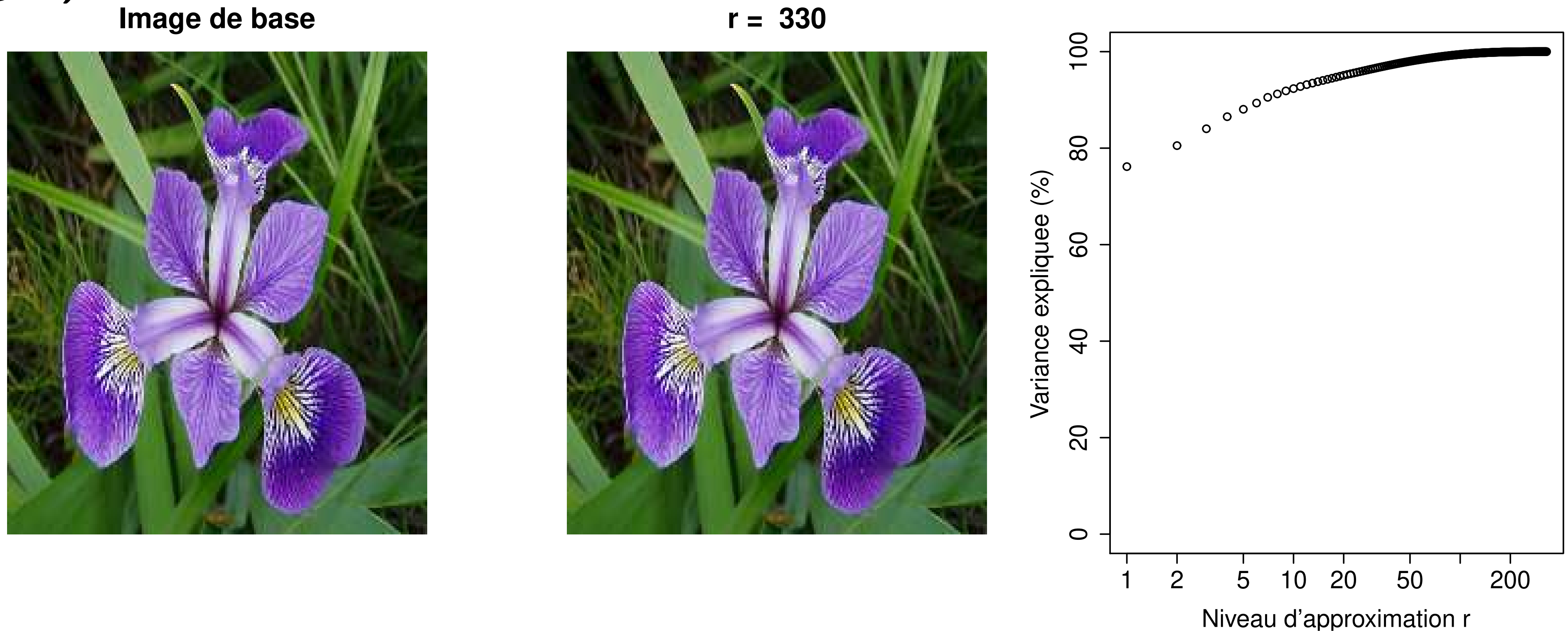


Figure 15: *Degree of approximation of the truncated SVD.*

Illustration (never used for image compression—non linearity of images)

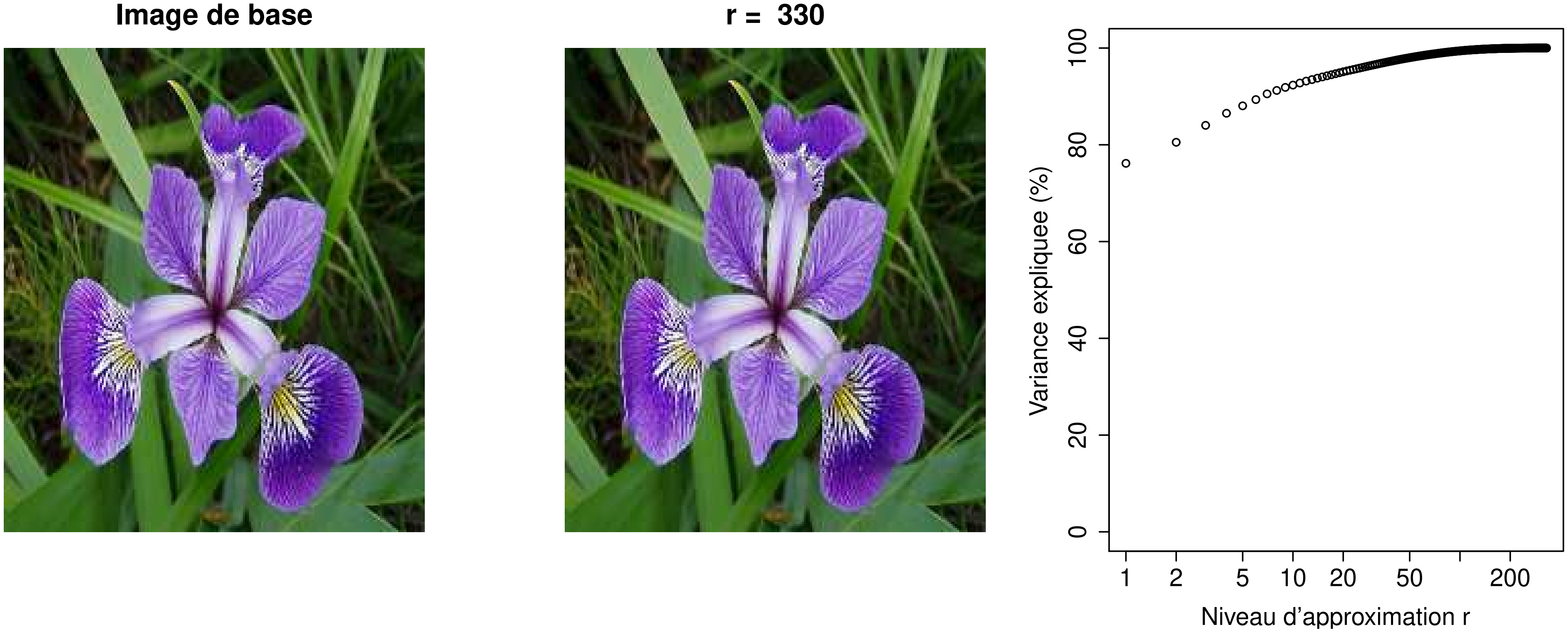


Figure 15: *Degree of approximation of the truncated SVD.*

Table 2: *Size of the compressed image as the cutoff value r varies.*

Rank r	1	10	50	100	Original (330)
Taille (Ko)	10	17	28	31	41
Compression (%)	75	58	31	24	0

Never forget

- □ We will work on an approximation of the data
- □ Degree of precision is related to the cutoff value r
- □ If approximation is poor, then our subsequent conclusions will be just as poor!

PCA as a visualization tool

- □ Let start with our SVD (U, D, V) of $\mathbf{X}$.
- □ Recall that V is an orthogonal matrix and, as so, defines an orthonormal basis:
  - ☞ $\mathbf{X}V$ is the projection of (the rows of) $\mathbf{X}$ onto the basis V, i.e., we have projected individuals on a new subpace.

PCA as a visualization tool

- □ Let start with our SVD (U, D, V) of $\mathbf{X}$.
- □ Recall that V is an orthogonal matrix and, as so, defines an orthonormal basis:
  - ☞ $\mathbf{X}V$ is the projection of (the rows of) $\mathbf{X}$ onto the basis V, i.e., we have projected individuals on a new subpace.
- □ Using traditional PCA phrasing, we say
  - – that the j-th column v_j of V is the j-th factorial axis ;
  - – the points $\mathbf{X}v_j$ are the principal components for the j-th factorial axis.

PCA as a visualization tool

- □ Let start with our SVD (U, D, V) of $\mathbf{X}$.
- □ Recall that V is an orthogonal matrix and, as so, defines an orthonormal basis:

  ☞ $\mathbf{X}V$ is the projection of (the rows of) $\mathbf{X}$ onto the basis V, i.e., we have projected individuals on a new subpace.

- □ Using traditional PCA phrasing, we say
  - that the j-th column v_j of V is the j-th factorial axis ;
  - the points $\mathbf{X}v_j$ are the principal components for the j-th factorial axis.

☞ We will thus visualize projected data rather than raw data.

Illustration on a toy example

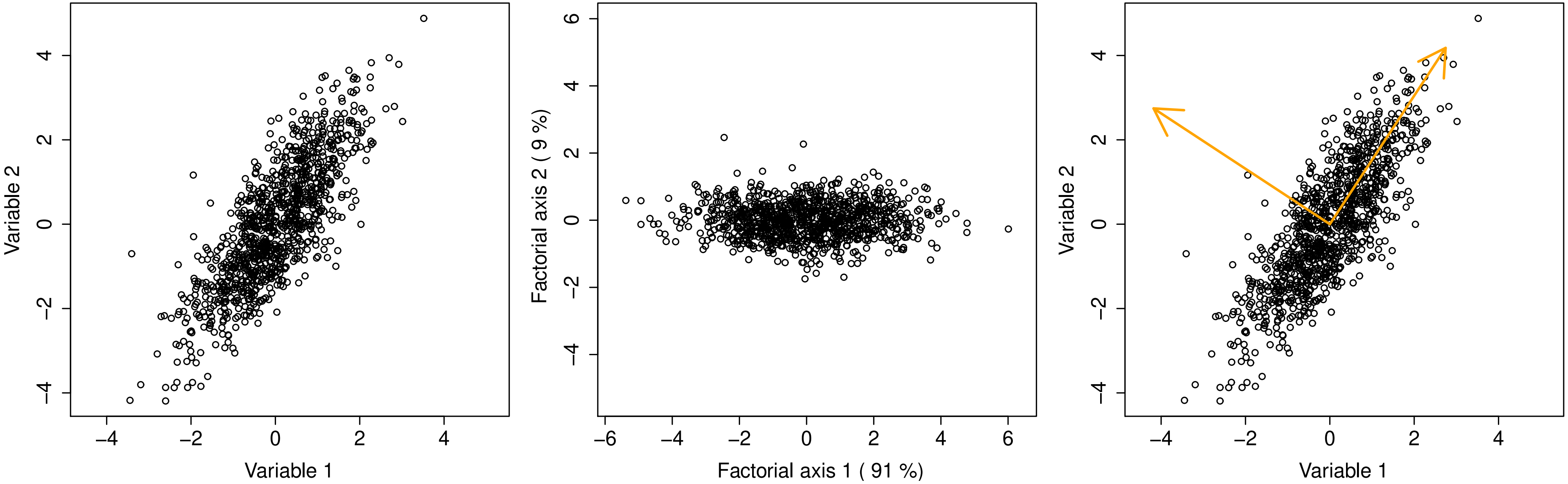


Figure 16: *Illustration of the factorial axis, principal components and proportion of variance explained.*

Illustration on a toy example

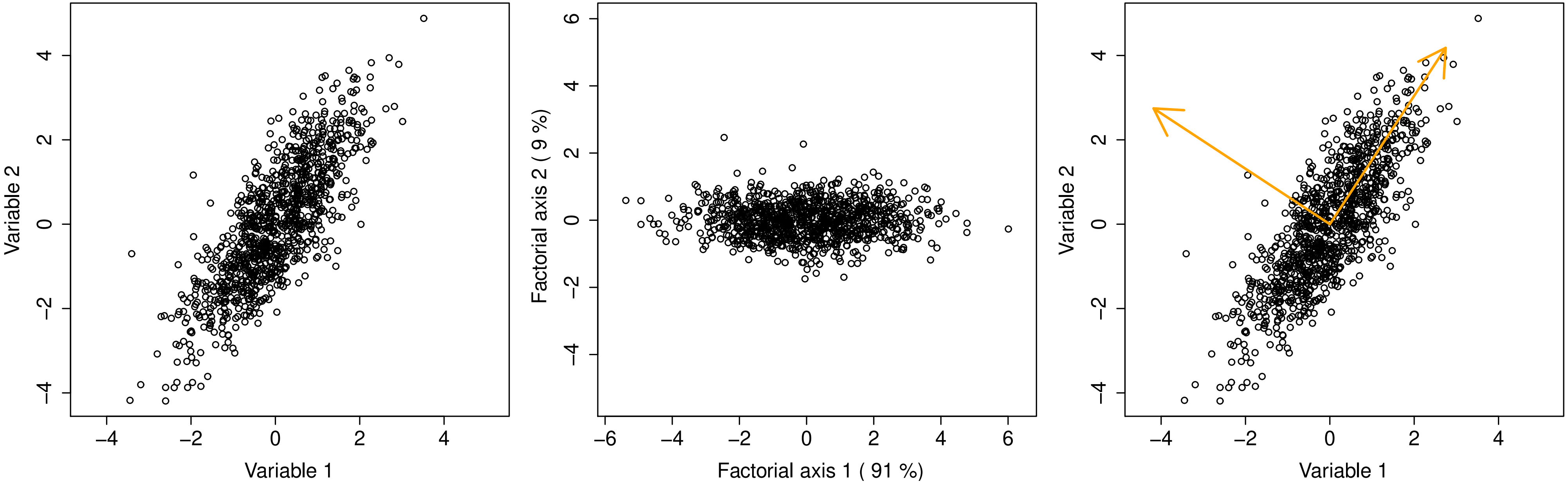


Figure 16: *Illustration of the factorial axis, principal components and proportion of variance explained.*

1st axis explains 91% of the variance and is defined by
Axis 1 $= 0.55 \times$ Variable 1 $+ 0.84 \times$ Variable 2.

2nd axis explains 9% of the variance and is defined by
Axis 2 $= -0.84 \times$ Variable 1 $+ 0.55 \times$ Variable 2.

Beware of projections

- □ The above example is dumb since we start from $\mathbb{R}^2$ to go to $\mathbb{R}^2$
- □ There is thus no loss of information
- □ Most often we will start from $\mathbb{R}^p$ to go to $\mathbb{R}^{p'}$, $p' < p$—typically $p' \in \{2, 3\}$.
- □ There is potentially a (large) information loss.

Beware of projections

- □ The above example is dumb since we start from $\mathbb{R}^2$ to go to $\mathbb{R}^2$
- □ There is thus no loss of information
- □ Most often we will start from $\mathbb{R}^p$ to go to $\mathbb{R}^{p'}$, $p' < p$—typically $p' \in \{2, 3\}$.
- □ There is potentially a (large) information loss.

Example 3. Consider the points $A = (1, 2, 0)$ and $B = (1, 2, 500)$ of $\mathbb{R}^3$. We project them onto the plan $\{(x, y, z) \colon z = 0\}$. Within this plan, A and B are identically while there are very different in $\mathbb{R}^3$.

Accuracy of projection

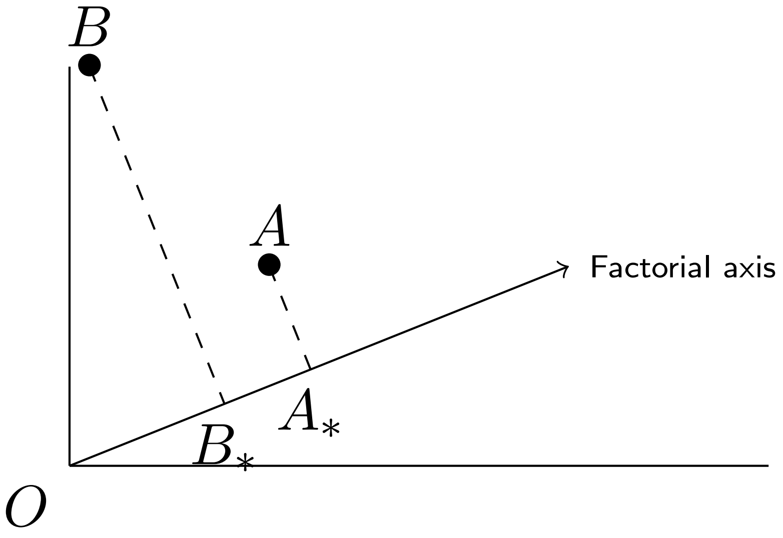


Figure 17: *Illustration of the notion of* $\cos^2$ *as a measure of projection accuracy.*

- □ $OA_* \approx OA \Rightarrow A$ is well represented on the factorial axis;
- □ $OB_* \not\approx OB \Rightarrow B$ is poorly represented on the factorial axis.

Accuracy of projection

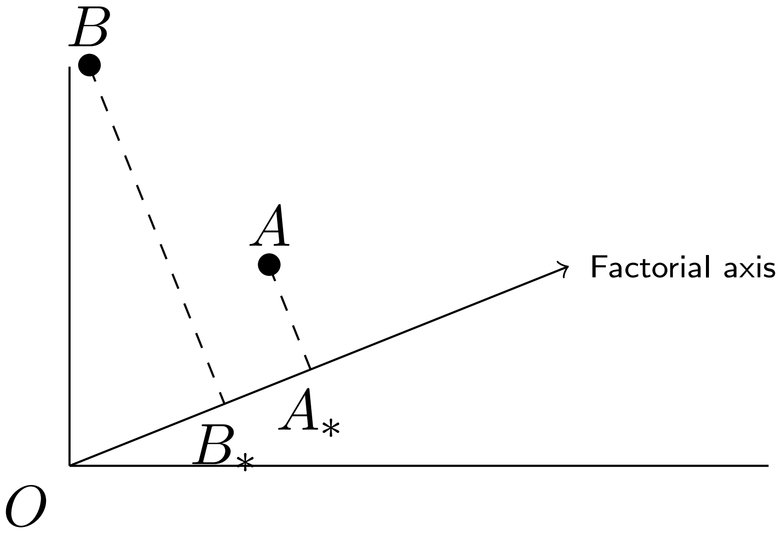


Figure 17: *Illustration of the notion of* $\cos^2$ *as a measure of projection accuracy.*

- □ $OA_* \approx OA \Rightarrow A$ is well represented on the factorial axis;
- □ $OB_* \not\approx OB \Rightarrow B$ is poorly represented on the factorial axis.

☞ The projection accuracy is thus measured by

$$\frac{OA_*^2}{OA^2} = \cos^2 \widehat{AOA_*}.$$

Individual leverage on a factorial axis

- □ Recall that $\|\mathbf{X}\|_F^2 = \sum_{j=1}^p \lambda_j^2$.
- □ The j-th factorial axis has contribution

$$100 \times \frac{\lambda_j^2}{\sum_{\ell=1}^p \lambda_\ell^2} \text{ \% of the variance / inertia.}$$

- □ The i-th individuals contributes to the j–th factorial axis

$$\frac{\|x_{i\cdot} v_j\|^2}{\lambda_j^2}$$

Duality

- □ So far we talked about projected individuals, i.e., rows of $\mathbf{X}$.
- □ It was justified since, from the SVD $\mathbf{X} = UDV^\top$, V is orthogonal.
- □ But wait. . .

Duality

- □ So far we talked about projected individuals, i.e., rows of $\mathbf{X}$.
- □ It was justified since, from the SVD $\mathbf{X} = UDV^\top$, V is orthogonal.
- □ But wait... U is orthogonal too! Just do the same on variables, i.e., columns of $\mathbf{X}$.
- □ This is known under the phrasing duality.

Duality

- □ So far we talked about projected individuals, i.e., rows of $\mathbf{X}$.
- □ It was justified since, from the SVD $\mathbf{X} = UDV^\top$, V is orthogonal.
- □ But wait… U is orthogonal too! Just do the same on variables, i.e., columns of $\mathbf{X}$.
- □ This is known under the phrasing duality.
- □ However this $\mathbf{X}$ is centered and scaled, we have for all $j \in \{1, \dots, p\}$

$$\|\tilde{x}_{\cdot j}\|^2 = 1 \quad \tilde{x}_{\cdot j} = \frac{x_{\cdot j}}{\sqrt{n}}, \qquad \text{since } \frac{1}{n}\|x_{\cdot j}\|^2 = 1,$$

hence the projection of the rescaled variables $\tilde{x}_{\cdot j}$ on any factorial plane (u_{i_1}, u_{i_2}) necessarily lies within the unit circle.

- □ It is known as the correlation circle.
- □ In this setting, the projection accuracy $\cos^2$ simplifies to

$$\frac{OA_*^2}{OA^2} = OA_*^2.$$

A gentle study on a socio-economic dataset

TAN	Growth rate (%)	**M15**	% people under 15
TXN	Birth rate (%)	**P65**	% people over 65
TMI	Child mortality rate (‰)	**PUR**	% urban population (%)
ESV	Life expectancy (years)	**PIB**	annual GDP per capita ($)

```
              TAN TXN TMI ESV M15 P65  PUR   PIB
Norvege       0.1  12   8  76  20  16 80.3 19500
France        0.4  14   8  75  21  13 77.2 15450
Australie     0.8  16  10  76  24  10 87.0 12000
Japon         0.6  12   6  77  22  10 76.5 19100
USA           0.7  16  11  75  22  12 74.0 18200
Bresil        2.1  29  63  65  36   4 74.0  1980
Pologne       0.8  18  19  71  25   9 60.0  4358
Mexique       2.4  31  50  67  42   4 70.0  1480
Maroc         2.6  36  90  60  42   4 44.0   549
Egypte        2.6  37  93  59  40   4 46.5   770
Albanie       2.0  26  43  71  35   5 34.0   840
Niger         2.9  51 141  44  47   3 16.0   205
Inde          2.1  33 101  55  38   4 25.5   275
Chine         1.3  21  61  66  28   5 21.0   255
ArabieSaoudite 3.2  39  79  63  37   2 73.0  5680
Portugal      0.2  12  17  73  24  12 31.0  3400
```

Explained variance

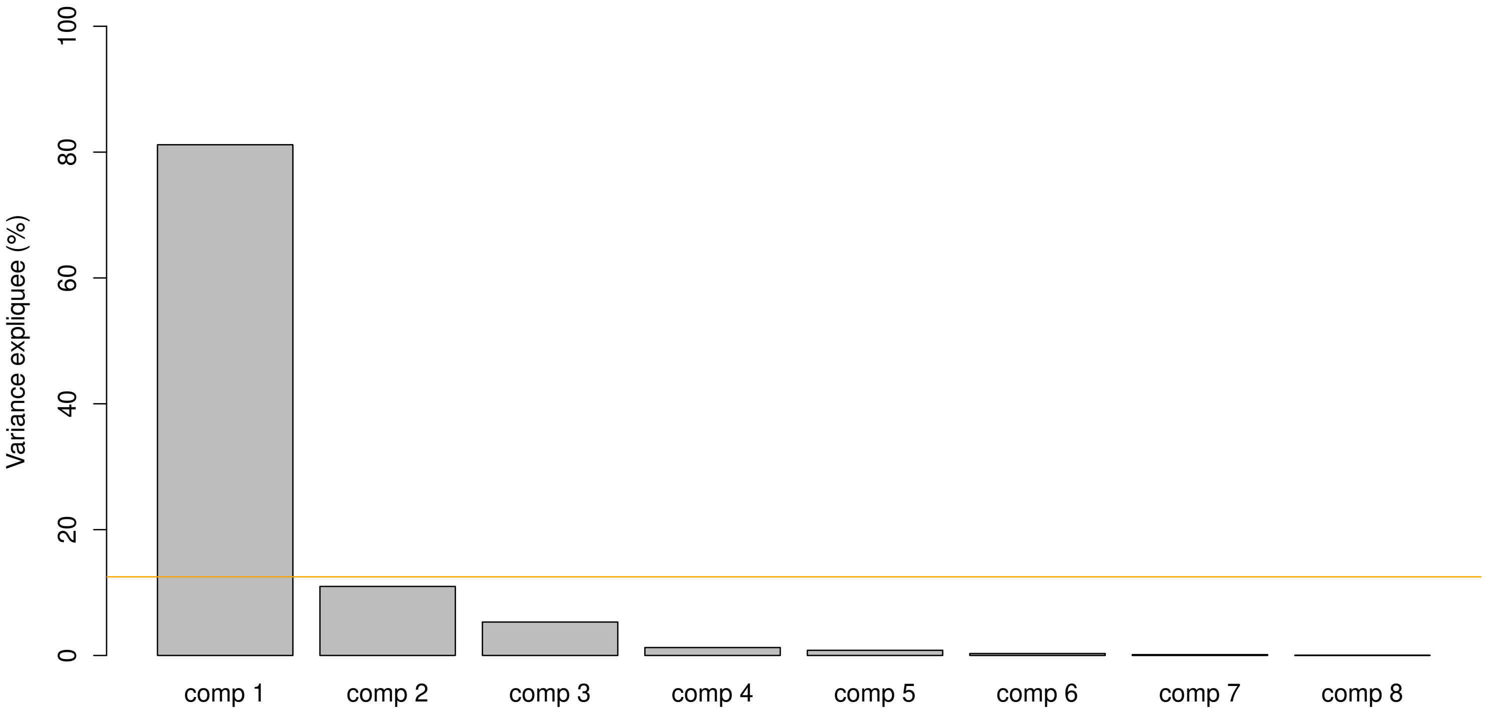


Figure 18: *Percentage of explained variance for each factorial axis. The orange horizontal line ($y = 100/p$) corresponds to a balanced contribution.*

Explained variance

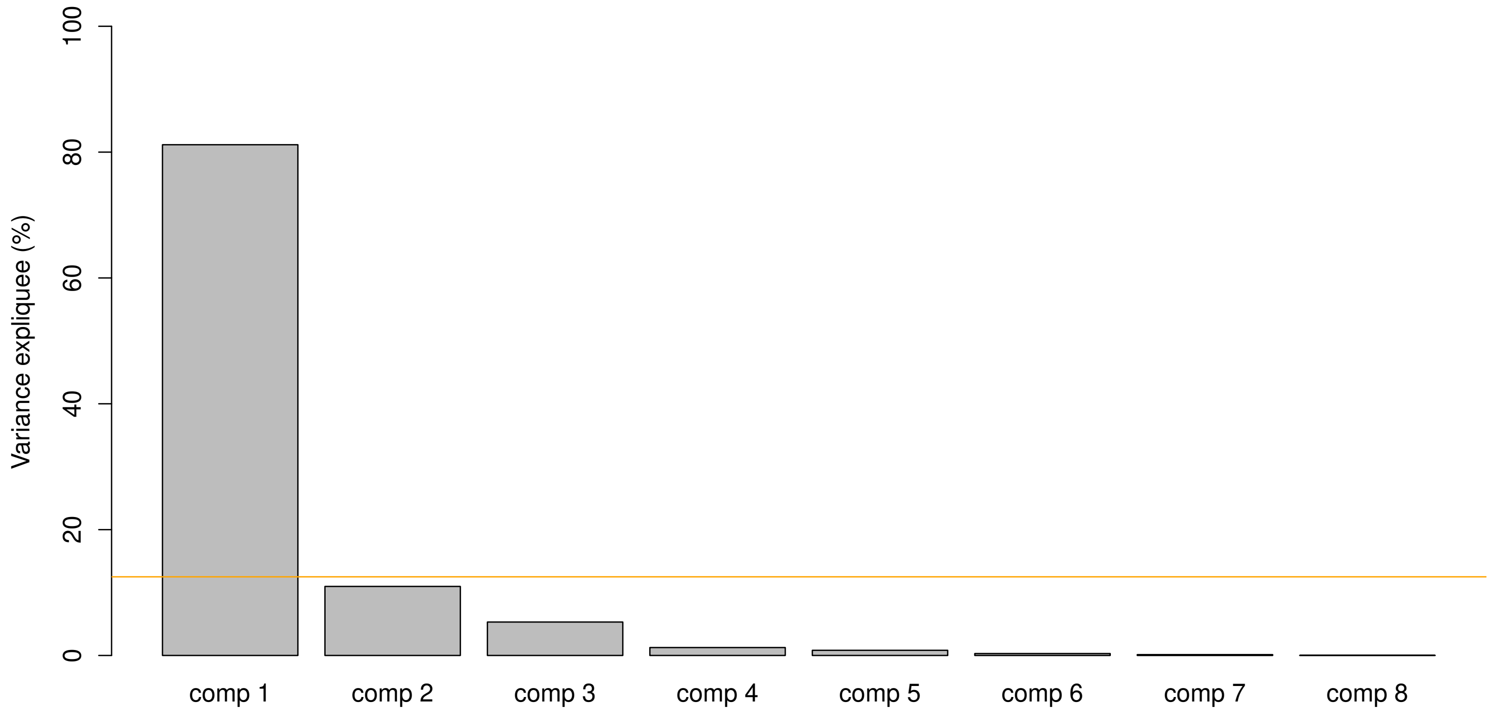


Figure 18: *Percentage of explained variance for each factorial axis. The orange horizontal line ($y = 100/p$) corresponds to a balanced contribution.*

☞ Here we could keep only 2 or 3 factorial axis. With 2 axis, we explain $81+11 = 92\%$ of the variance; adding a 3rd axis will explain $81+11+5 = 97\%$ of the variance.

Interpretation

Step 1 Analyze the variable

□ give a meaning to the axis

□ identify clusters of arrows and give them a meaning

Step 2 Analyze the individuals

□ makes sense of what is the origin

□ look at individuals coordinates and interpret them according to Step 1.

Principal components on the 1st factorial plane (try to interpret it!)

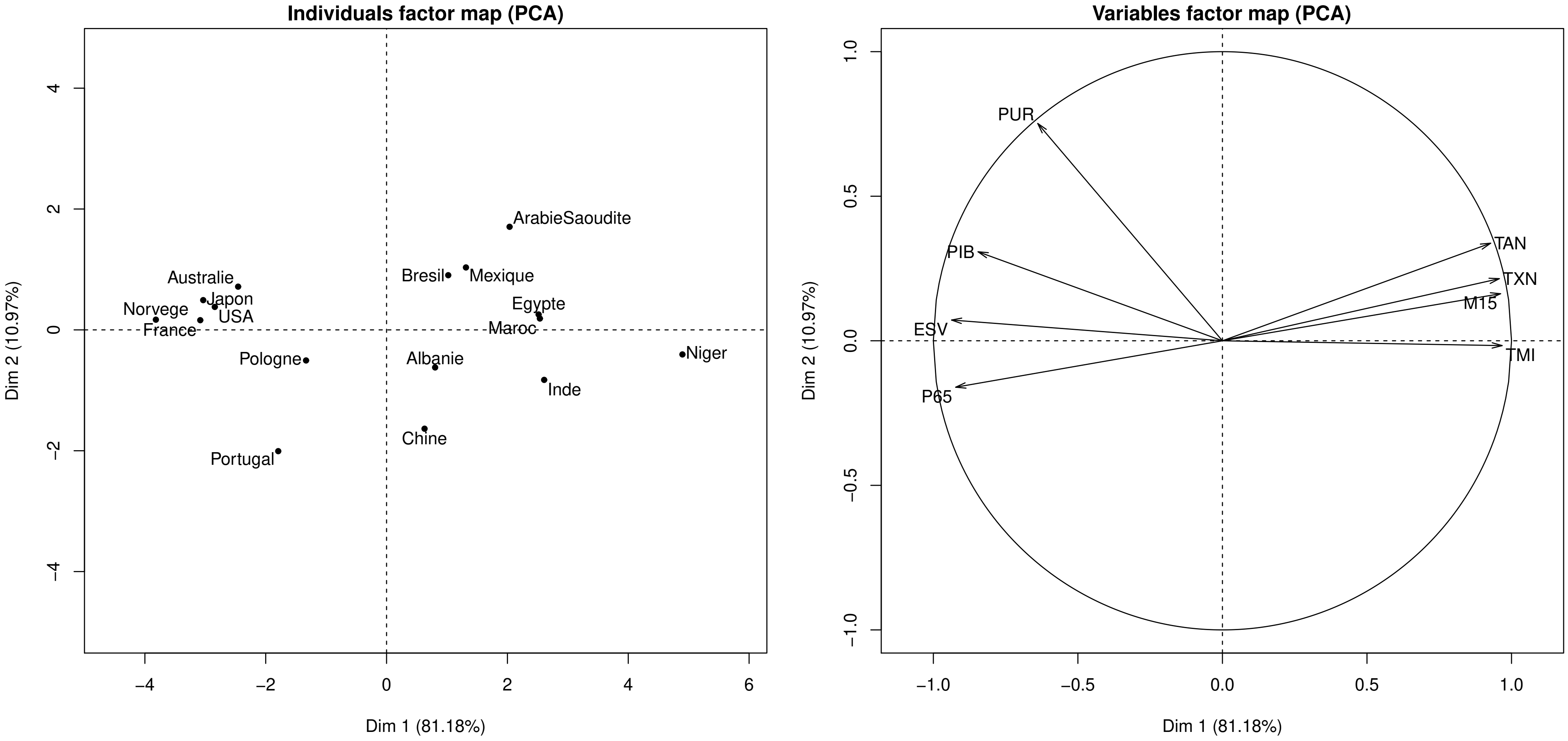


Figure 19: *Principal component on the 1st factorial plane, i.e., axis 1 and 2. Left: individuals. Right: Variables.*

To go a bit further

Supplementary individuals

- □ Let $x_{*\cdot}$ be a new individual.
- □ From our PCA, computed from $\mathbf{X}$ only, we can project $x_{*\cdot}$ onto the basis formed by V, i.e., $x_{*\cdot}V$.
- □ It enables to identify how the new individual $x_{*\cdot}$ relates to our previous conclusions derived from the PCA.
- □ Using duality, we can do the same with a new variable $x_{\cdot *}$, i.e., $x_{\cdot *}^\top U$.

Categorical variables

- □ PCA is limited to quantitative variables
- □ Actually one can use categorical variables as well, but in a different way.
- □ Those categorical variable won't be used for the SVD but rather for visualization purposes.

Why using supplementary individuals?

- □ Recall that the i–th individual contributes to the j–th factorial axis is given by

$$\frac{\|x_{i\cdot} v_j\|^2}{\lambda_j^2}.$$

- □ If the above contribution is too large, i.e., $\gg 1/n$, factorial axis may be too dependent such indivuals.
- □ Recall that the aim of a PCA is to put an emphasis on the general behaviour of individuals not only a few!
- □ We thus may want to treat influential individuals as supplementary.

Why using supplementary variables?

□ You may wonder why not using all possible information in PCA?

□ It may happens that some variables are highly (linearly) dependent

□ We may want to treat a variable as supplementary to see how it relates to other variables.

Supplementary individual // variable

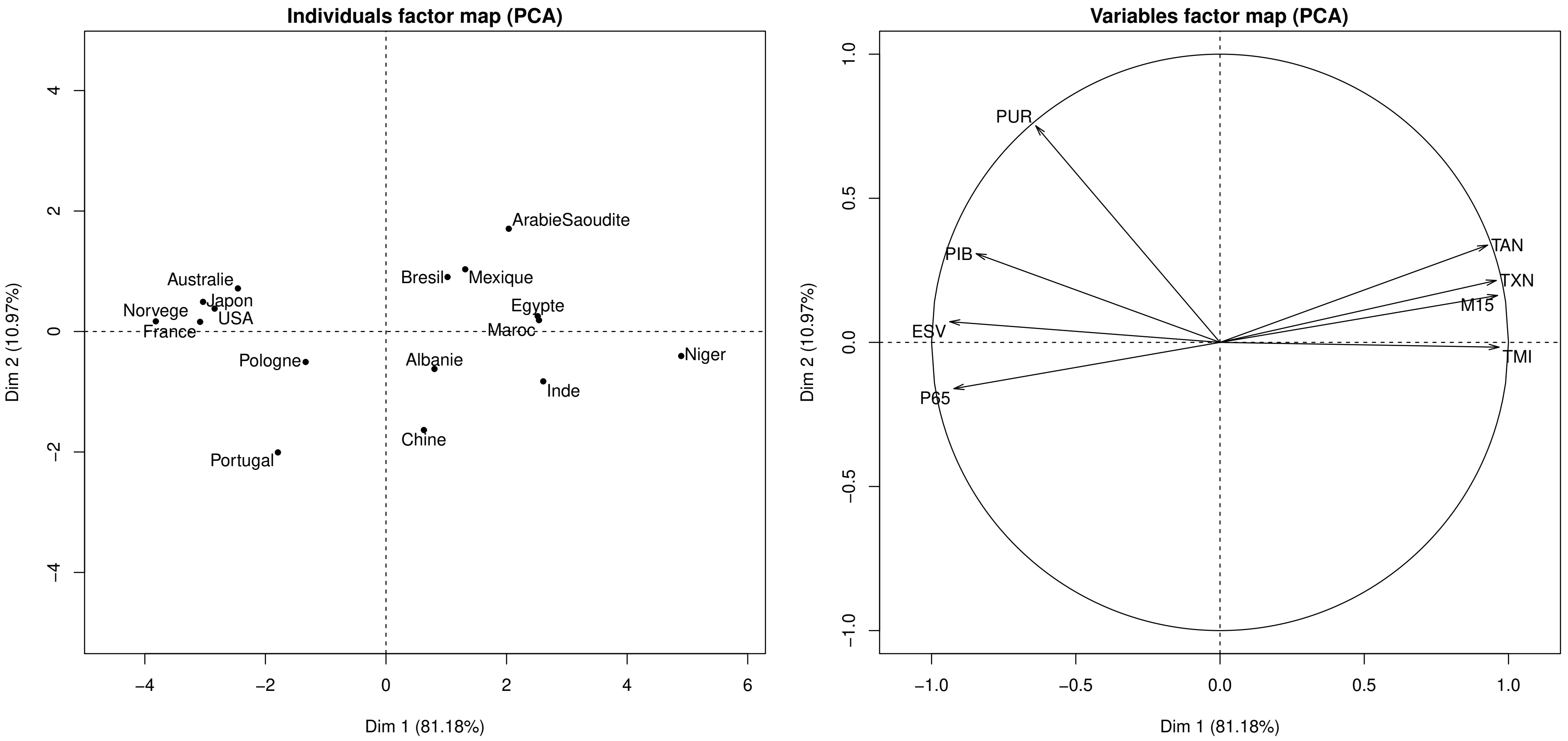


Figure 20: *Illustration of supplementary individuals and variables within a PCA.*

Supplementary individual // variable

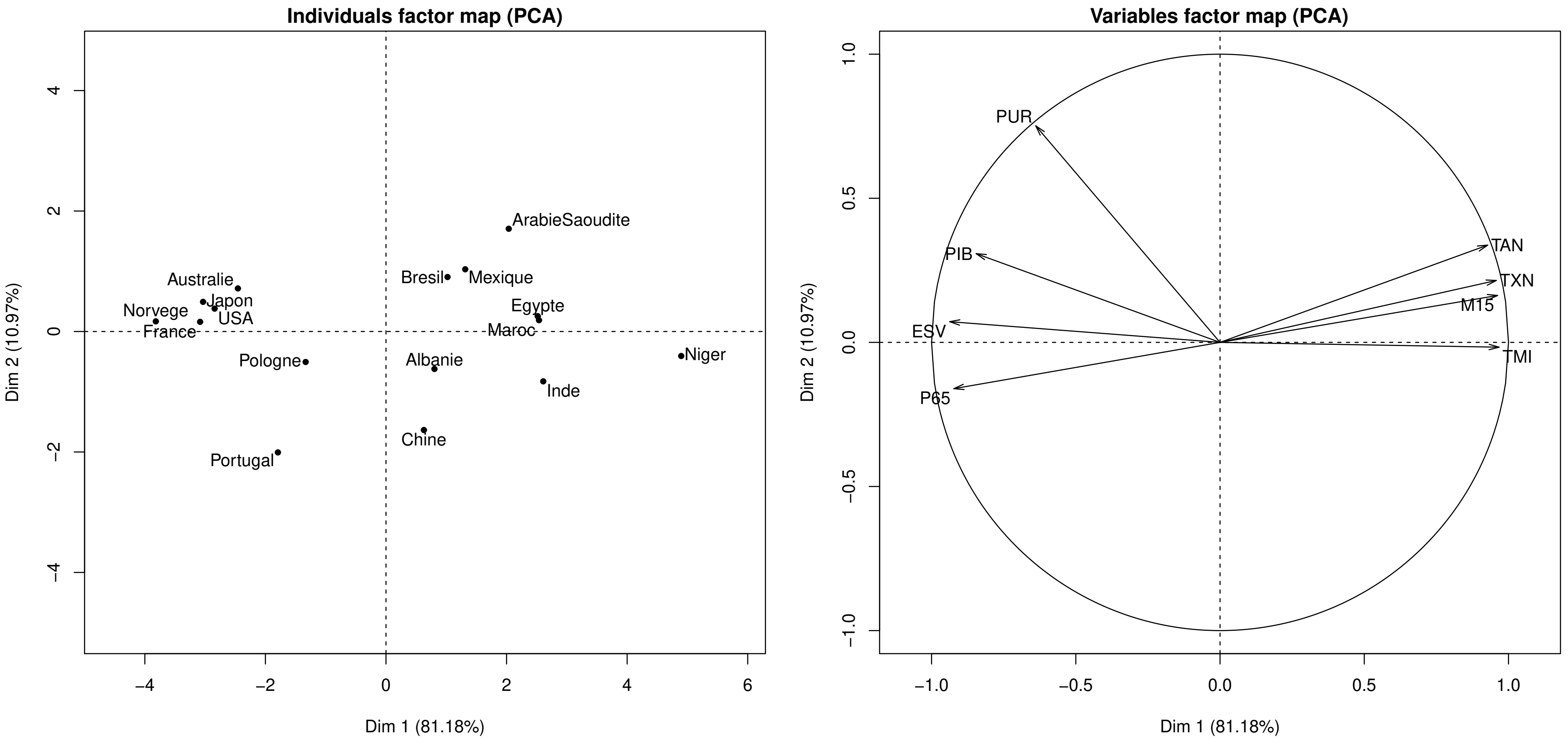


Figure 20: *Illustration of supplementary individuals and variables within a PCA.*

□ Consider the new country “Syldavie”: similar to France but rather rural

□ Consider the new variable “% of smokers”

Supplementary individual // variable

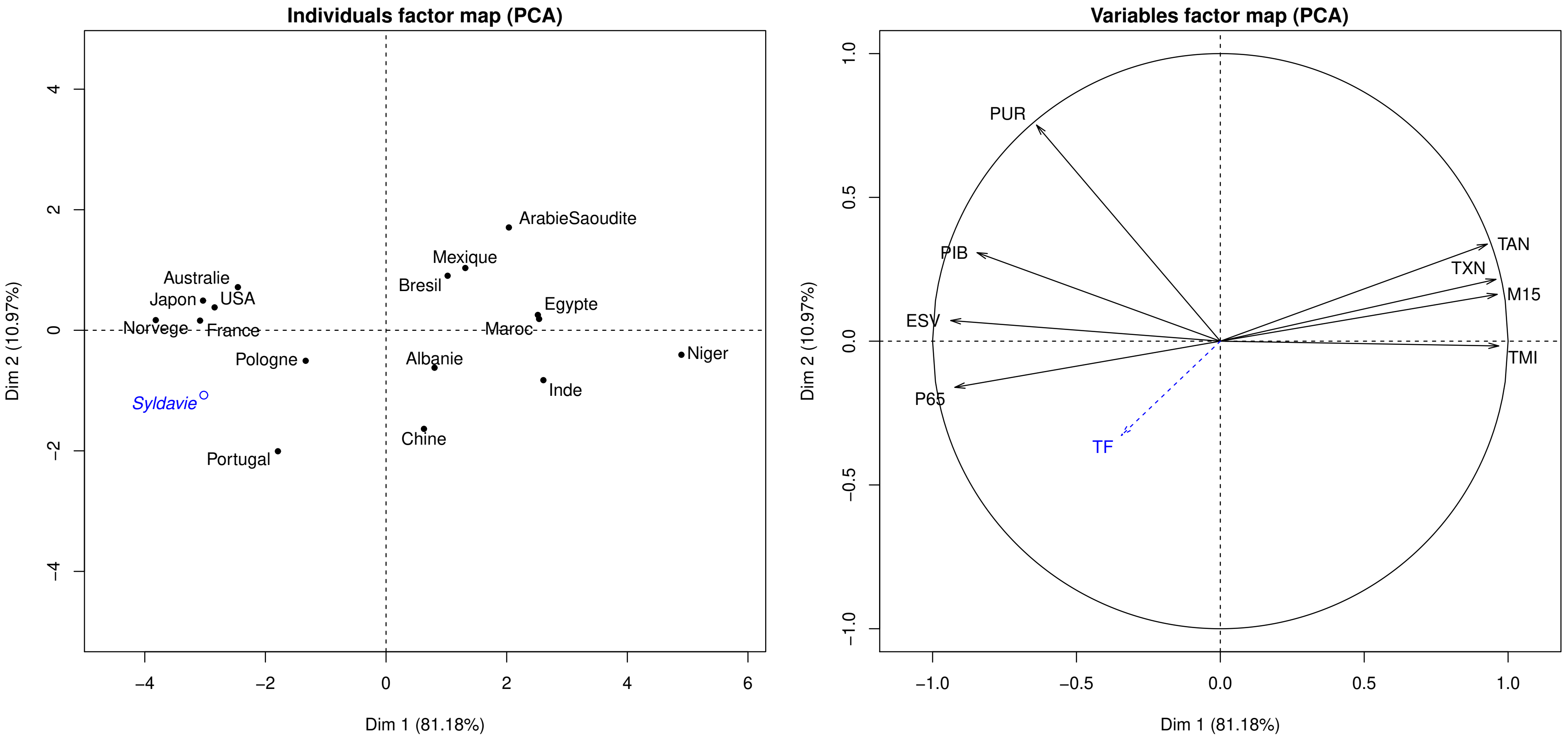


Figure 20: *Illustration of supplementary individuals and variables within a PCA.*

□ Consider the new country "Syldavie": similar to France but rather rural

□ Consider the new variable "% of smokers"

Categorical variable

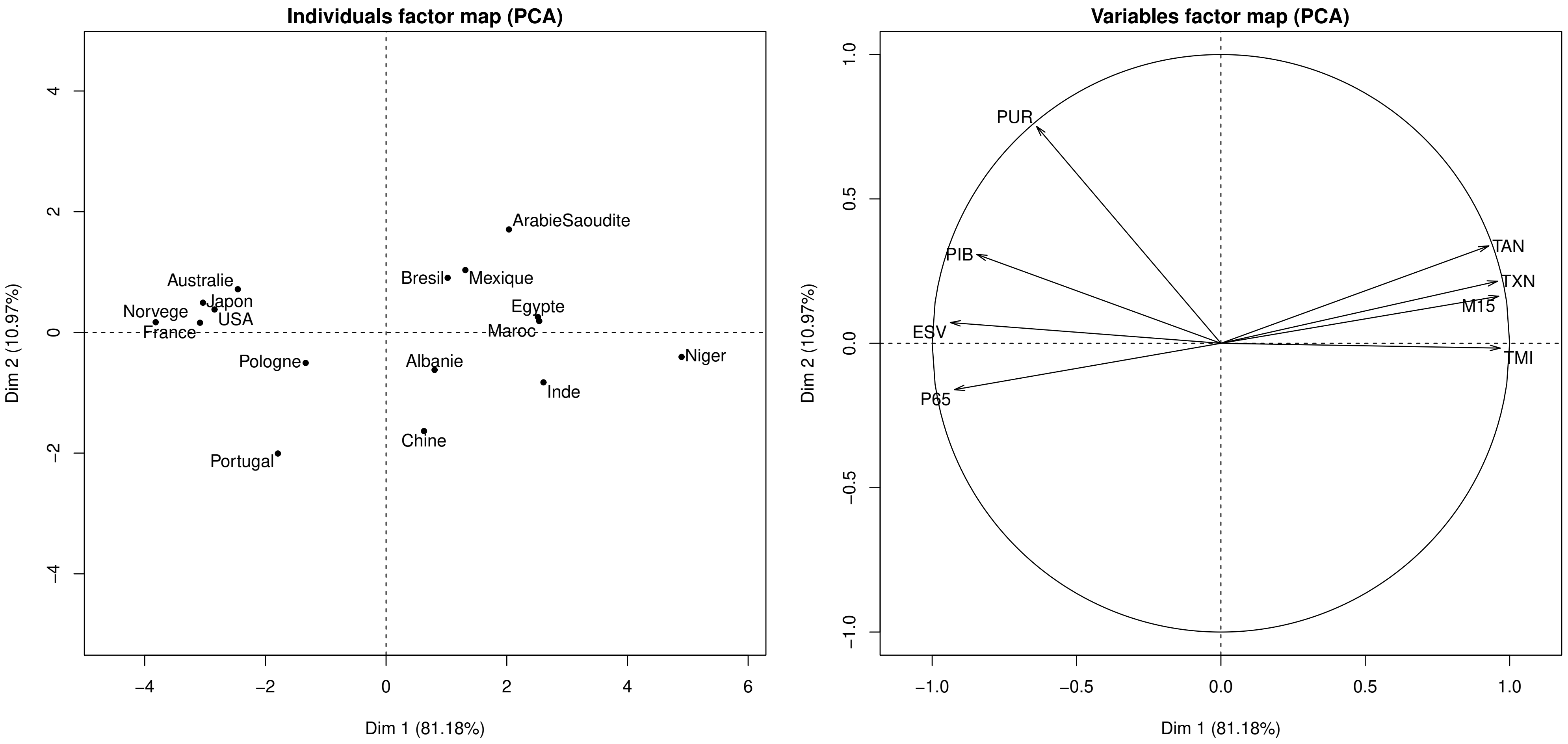


Figure 21: *Illustration of a new categorical variable within a PCA.*

Categorical variable

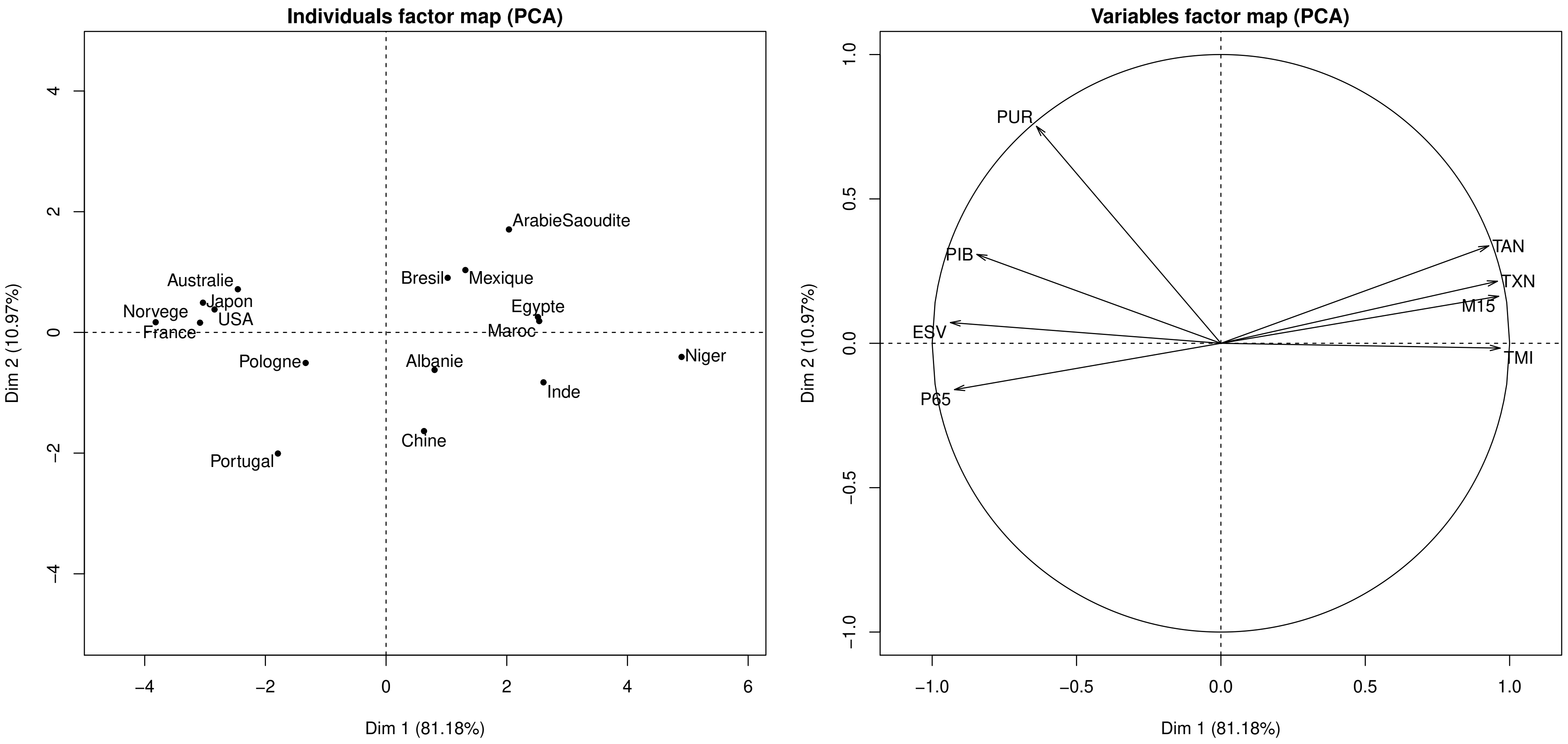


Figure 21: *Illustration of a new categorical variable within a PCA.*

□ Add a new categorical variable $HEM \in \{North, South\}$.

Categorical variable

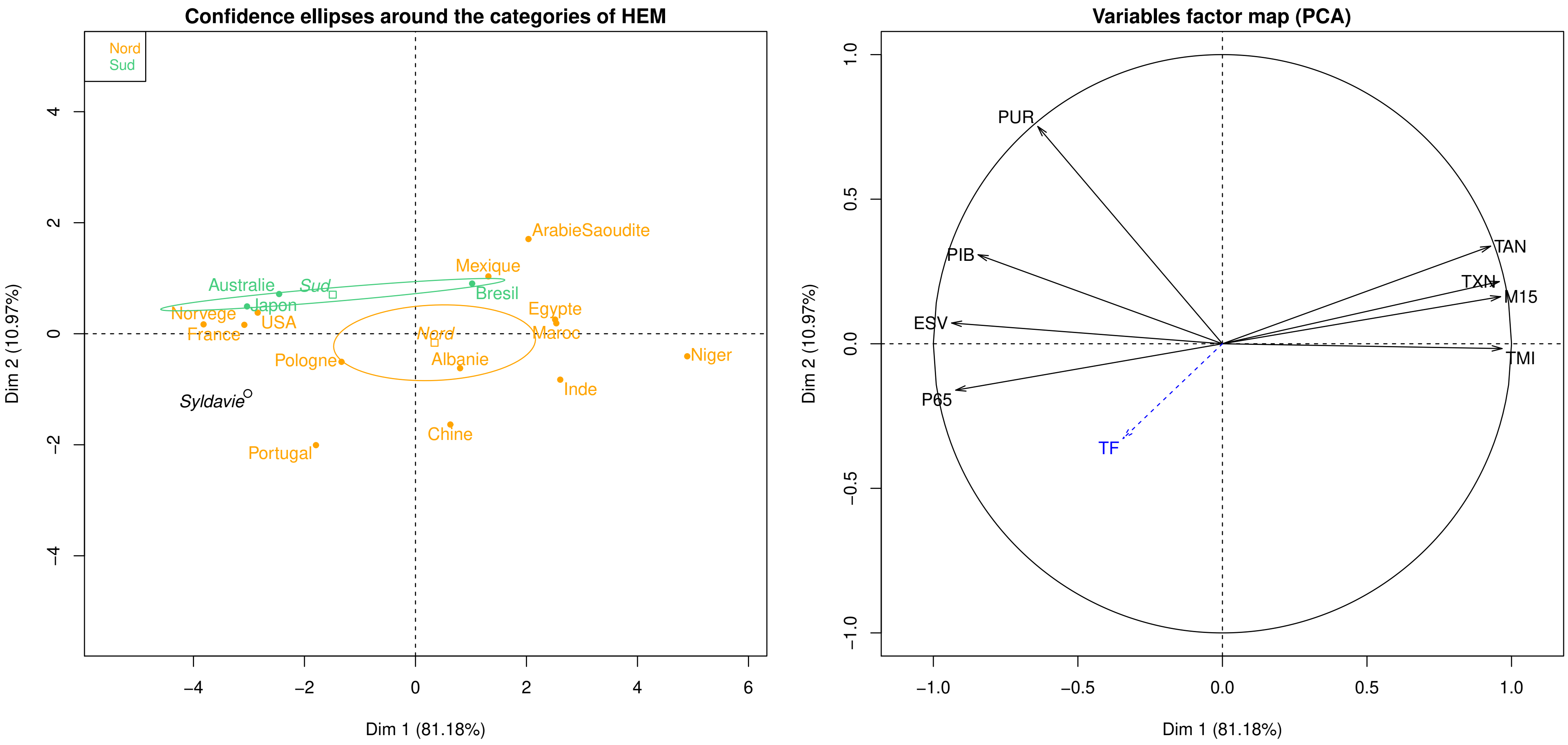


Figure 21: *Illustration of a new categorical variable within a PCA.*

□ Add a new categorical variable $HEM \in \{North, South\}$.

5. Linear models

Homework

□ Get the book `An introduction to Statistical Learning with Applications in R` from this link

□ Read Chapter 3 and do the lab of Section 3.6

- □ Linear models is probably the simple statistical model for regression problem.
- □ Recall that regression problem aims at predicting some numerical value Y with respect to some covariates / features $\mathbf{x} = (x_1, \dots, x_p)\top$.
- □ It is the simple model as extensions are possible such as:

  - – generalized linear models
  - – additive models
  - – generalized additive models
  - – regularized linear model such as ridge, lasso or elastic net.

Linear regression model

Definition 8. Given a sample $\mathscr{D}_n = \{(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^p \colon i := 1, \ldots, n\}$, a statistical model is said to be a (gaussian) linear regression model if we assume

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. More compactly, this can be written (without the Gaussian assumption but only white noise)

$$\mathbb{E}(Y \mid X) = X^\top \boldsymbol{\beta}, \qquad \boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top.$$

Fitting a linear model

□ Having observed a data set $\mathscr{D}_n = \{(Y_i, X_i) \colon i = 1, \ldots, n\}$, we want to fit our linear model, i.e., compute the least square estimator $\hat{\beta}$ for $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta} \right)^2$$

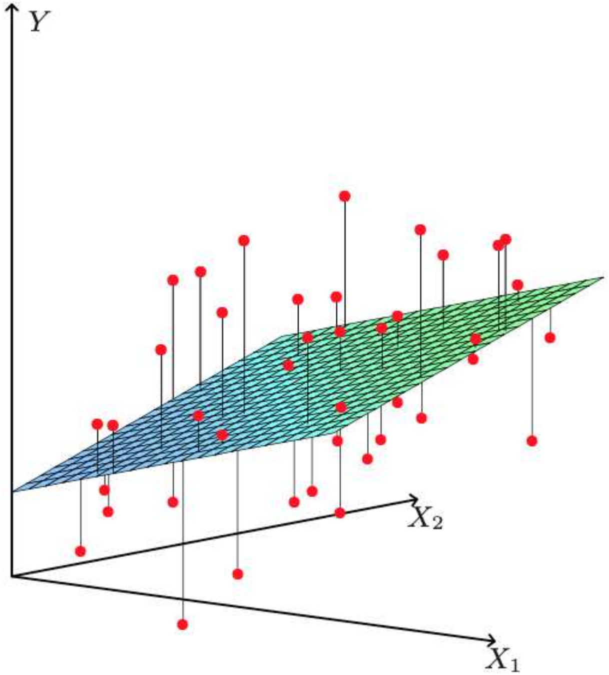


Figure 22: *Linear least square fitting with* $\mathbf{X} \in \mathbb{R}^{n \times 3}$. *[Taken from ESLII]*

Fitting a linear model

- □ Having observed a data set $\mathscr{D}_n = \{(Y_i, X_i) \colon i = 1, \dots, n\}$, we want to fit our linear model, i.e., compute the least square estimator $\hat{\boldsymbol\beta}$ for $\boldsymbol\beta$

$$\hat{\boldsymbol\beta} = \arg\min_{\boldsymbol\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_i^\top \boldsymbol\beta\right)^2$$

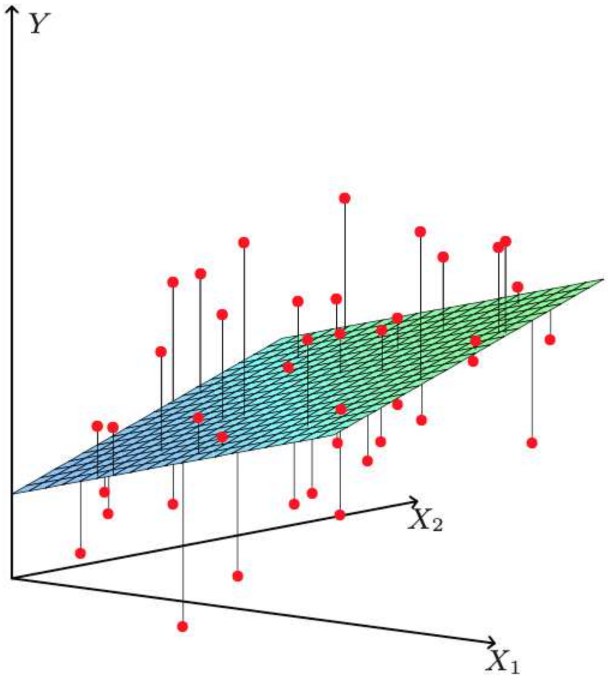


Figure 22: *Linear least square fitting with* $\mathbf{X} \in \mathbb{R}^{n\times 3}$*. [Taken from ESLII]*

- □ One can show that

$$\hat{\boldsymbol\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y},$$

where $\mathbf{X}$ is the design matrix whose i-th row is $\mathbf{x}_i$ and $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$.

- □ This yields to the prediction

$$\hat{\mathbf{Y}} = H\mathbf{Y}, \quad H = \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$$

Least squares as the MLE

Proposition 1. *For Gaussian noise, the MLE for the linear model is the least square solution. Indeed (conditionally on the features X_i) the log–likelihood is*

$$\ell(\theta; \mathcal{D}_n) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - X_i^T\boldsymbol{\beta}\right)^2, \qquad \theta = (\boldsymbol{\beta}, \sigma^2).$$

Consequently, maximizing the above expression w.r.t. $\boldsymbol{\beta}$ consists in the least square problem

$$\mathop{\arg\min}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n}\left(Y_i - X_i^\top\boldsymbol{\beta}\right)^2.$$

Least squares as the MLE

Proposition 1. *For Gaussian noise, the MLE for the linear model is the least square solution. Indeed (conditionally on the features X_i) the log–likelihood is*

$$\ell(\theta; \mathcal{D}_n) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - X_i^T\boldsymbol{\beta}\right)^2, \qquad \theta = (\boldsymbol{\beta}, \sigma^2).$$

Consequently, maximizing the above expression w.r.t. $\boldsymbol{\beta}$ consists in the least square problem

$$\operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n}\left(Y_i - X_i^\top\boldsymbol{\beta}\right)^2.$$

☞ We can use all the properties we know about the maximum likelihood estimator!

Measure of goodness of fit

- □ It is common practice to measure how well the model fits the data.
- □ A common choice is the the coefficient of determination or percentage of variance explained R^2

$$R^2 = 1 - \frac{\sum_{i=1}^n \left(\hat{Y}_i - Y_i\right)^2}{\sum_{i=1}^n \left(\bar{Y} - Y_i\right)^2} = 1 - \frac{\text{residual sum of squares (RSS)}}{\text{total sum of squares (TSS)}}, \quad \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

- □ It measures how your model increases the prediction performance compared to the baseline model, e.g., unknown intercept.
- □ Clearly $R^2 = 1$ for perfect predictions.

Measure of goodness of fit

- □ It is common practice to measure how well the model fits the data.
- □ A common choice is the the coefficient of determination or percentage of variance explained R^2

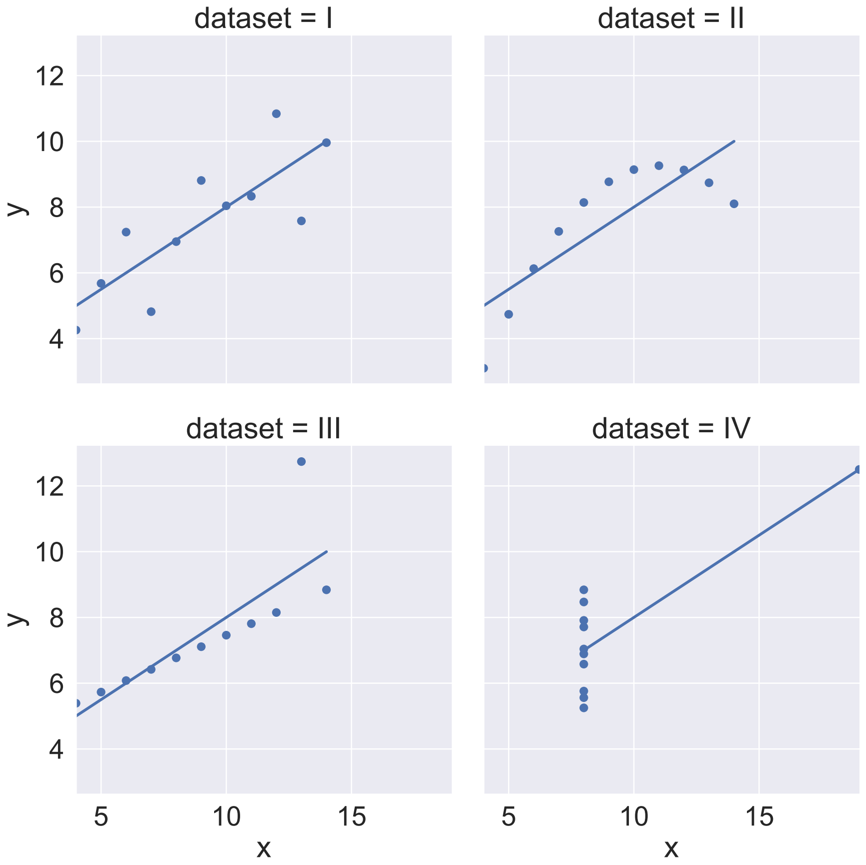
$$R^2 = 1 - \frac{\sum_{i=1}^n \left(\hat{Y}_i - Y_i\right)^2}{\sum_{i=1}^n \left(\bar{Y} - Y_i\right)^2} = 1 - \frac{\text{residual sum of squares (RSS)}}{\text{total sum of squares (TSS)}}, \quad \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

- □ It measures how your model increases the prediction performance compared to the baseline model, e.g., unknown intercept.
- □ Clearly $R^2 = 1$ for perfect predictions.

☞ Watchout if your model has no intercept the above formula is incorrect and one must use $R^2 = \text{corr}(\hat{Y}, Y)^2$.

Warning: Never trust a single numerical value!

```
>>> import seaborn as sns
>>> df = sns.load_dataset("anscombe")
>>> df
   dataset     x      y
0        I  10.0   8.04
1        I   8.0   6.95
     .
     .
     .
42      IV   8.0   7.91
43      IV   8.0   6.89
>>> df.groupby("dataset").corr().iloc[::2,-1]**2
dataset
I         x    0.666542
II        x    0.666242
III       x    0.666324
IV        x    0.666707
Name: y, dtype: float64
```



Species in Galápagos Islands (Faraway, 2014)

Isla Pinta
Isla Marchena
Volcán Wolf
ISLAS GALÁPAGOS
Isla Santiago
Isla Fernandina
Isla Santa Cruz
Puerto Isidro Ayora
Isla Isabela
Puerto Villamil
Puerto Baquerizo Moreno
Isla de San Cristóbal
Isla Floreana
Isla Española

□ Response `Species`: Number of the species found on each of the 30 islands of the Galápagos

□ 5 features : `Elevation`: highest elevation of the island, `Nearest` distance from the nearest island, `Scruz` distance from the Santa Cruz island, `Adjacent` the area of the adjacent island

Interpretation

- □ Suppose we have fitted the following linear model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p,$$

we may wonder what is the meaning of $\hat{\beta}_1$ for instance?

- □ Sometimes (rarely), it is a physical constant but most often it has no real physical meaning as we are just building an empirical model approximating reality.

Interpretation

□ Suppose we have fitted the following linear model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p,$$

we may wonder what is the meaning of $\hat{\beta}_1$ for instance?

□ Sometimes (rarely), it is a physical constant but most often it has no real physical meaning as we are just building an empirical model approximating reality.

□ Naive Interpretation:

> A unit change in X_1 will produce on average a change of $\hat{\beta}_1$ in the response

□ Such a reasoning is correct provided that:

- the model is correct and you are not extrapolating
- covariates are orthogonal—which is typically not the case.

Interpretation

□ Suppose we have fitted the following linear model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p,$$

we may wonder what is the meaning of $\hat{\beta}_1$ for instance?

□ Sometimes (rarely), it is a physical constant but most often it has no real physical meaning as we are just building an empirical model approximating reality.

□ Right Interpretation:

> A unit change in X_1 with the other features held constant will produce on average a change of $\hat{\beta}_1$ in the response

Interpretation

□ Suppose we have fitted the following linear model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p,$$

we may wonder what is the meaning of $\hat{\beta}_1$ for instance?

□ Sometimes (rarely), it is a physical constant but most often it has no real physical meaning as we are just building an empirical model approximating reality.

□ Right Interpretation:

> A unit change in X_1 with the other features held constant will produce on average a change of $\hat{\beta}_1$ in the response

☞ Beware we are talking about correlation but not causality. Think about observing a positive correlation between shoe sizes and reading abilities—we missed lurking variable age of the child! Causality analysis is difficult!

Fitting a linear model (`sklearn`)

```
>>> import faraway.datasets.galapagos ##just for the dataset
>>> galapagos = faraway.datasets.galapagos.load()
>>> galapagos.head()
           Species   Area  Elevation  Nearest  Scruz  Adjacent
Baltra          58  25.09        346      0.6    0.6      1.84
Bartolome       31   1.24        109      0.6   26.3    572.33
Caldwell         3   0.21        114      2.8   58.7      0.78
Champion        25   0.10         46      1.9   47.4      0.18
Coamano          2   0.05         77      1.9    1.9    903.82

>>> X = galapagos.iloc[:, 1:]
>>> Y = galapagos.Species
>>> fit = LinearRegression().fit(X, Y)
>>> fit.coef_
array([-0.02393834,  0.31946476,  0.00914396, -0.24052423, -0.07480483])
```

☞ The analysis we just made is clearly too basic and we need more theory to do it properly.

☞ We will use `statsmodels` rather since `sklearn` is very limited

Fitting a linear model (statsmodels)

```
>>> import statsmodels.formula.api as smf
>>> fit = smf.ols(’Species ˜ Area + Elevation + Nearest + Scruz + Adjacent’, data = galapagos).fit()
>>> fit.summary()
                           OLS Regression Results
==============================================================================
Dep. Variable:                Species   R-squared:                       0.766
Model:                            OLS   Adj. R-squared:                  0.717
Method:                 Least Squares   F-statistic:                     15.70
Date:                Mon, 20 Jun 2022   Prob (F-statistic):           6.84e-07
Time:                        16:28:18   Log-Likelihood:                -162.54
No. Observations:                  30   AIC:                             337.1
Df Residuals:                      24   BIC:                             345.5
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.0682     19.154      0.369      0.715     -32.464      46.601
Area          -0.0239      0.022     -1.068      0.296      -0.070       0.022
Elevation      0.3195      0.054      5.953      0.000       0.209       0.430
Nearest        0.0091      1.054      0.009      0.993      -2.166       2.185
Scruz         -0.2405      0.215     -1.117      0.275      -0.685       0.204
Adjacent      -0.0748      0.018     -4.226      0.000      -0.111      -0.038
==============================================================================
Omnibus:                       12.683   Durbin-Watson:                   2.476
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               13.498
Skew:                           1.136   Prob(JB):                      0.00117
Kurtosis:                       5.374   Cond. No.                     1.90e+03
==============================================================================
```

Fitting a linear model (R)

```
> library(faraway) ## for the dataset
> head(gala[,-2])
            Species  Area Elevation Nearest Scruz Adjacent
Baltra            58 25.09       346     0.6   0.6     1.84
Bartolome         31  1.24       109     0.6  26.3   572.33
Caldwell           3  0.21       114     2.8  58.7     0.78
Champion          25  0.10        46     1.9  47.4     0.18
Coamano            2  0.05        77     1.9   1.9   903.82
Daphne.Major      18  0.34       119     8.0   8.0     1.84

> fit <- lm(Species ~ Area + Elevation + Nearest + Scruz  + Adjacent,
data=gala)
> fit

Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Coefficients:
(Intercept)         Area    Elevation      Nearest        Scruz     Adjacent
   7.068221    -0.023938     0.319465     0.009144    -0.240524    -0.074805
```

t–test in linear model (statsmodels)

$$H_0\colon \beta_j = 0 \qquad \text{vs} \qquad H_1\colon \beta_j \neq 0$$

□ Under the null H_0 (and with a gaussian noise), one can show that the test statistic statisfies

$$T = \frac{\hat\beta_j - 0}{\text{std. err.}(\hat\beta_j)} \sim t_{n-p-1}$$

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.0682     19.154      0.369      0.715     -32.464      46.601
Area          -0.0239      0.022     -1.068      0.296      -0.070       0.022
Elevation      0.3195      0.054      5.953      0.000       0.209       0.430
Nearest        0.0091      1.054      0.009      0.993      -2.166       2.185
Scruz         -0.2405      0.215     -1.117      0.275      -0.685       0.204
Adjacent      -0.0748      0.018     -4.226      0.000      -0.111      -0.038
==============================================================================
```

t–test in linear model (R)

```
> summary(fit)

Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Analysis of variance (ANOVA) (`statsmodels`)

$$H_0\colon \beta_1 = \cdots = \beta_p = 0 \qquad \text{against} \qquad H_1\colon \beta_j \neq 0 \text{ for some } j \in \{1, \dots, p\}$$

□ Under the null H_0 (and with a gaussian noise), one can show that the test statistic statisfies

$$T = \frac{(\mathsf{TSS} - \mathsf{RSS})/(p-1)}{\mathsf{RSS}/(n-p)} \sim F_{p-1,n-p}$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Species   R-squared:                       0.766
Model:                            OLS   Adj. R-squared:                  0.717
Method:                 Least Squares   F-statistic:                     15.70
Date:                Mon, 20 Jun 2022   Prob (F-statistic):           6.84e-07
Time:                        16:28:18   Log-Likelihood:                -162.54
No. Observations:                  30   AIC:                             337.1
Df Residuals:                      24   BIC:                             345.5
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
```

Analysis of variance (ANOVA) `(R)`

$$H_0\colon \beta_1 = \cdots = \beta_p = 0 \qquad \text{against} \qquad H_1\colon \beta_j \neq 0 \text{ for some } j \in \{1, \ldots, p\}$$

□ Under the null H_0 (and with a gaussian noise), one can show that the test statistic statisfies

$$T = \frac{(\text{TSS} - \text{RSS})/(p-1)}{\text{RSS}/(n-p)} \sim F_{p-1,n-p}$$

```
Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Anova (II) (statsmodels)

$H_0\colon \beta_{\text{Area}} = \beta_{\text{Adjacent}} = 0$ against $H_1\colon$ at least one of the two is non null

```
>>> import faraway.datasets.galapagos
>>> import statsmodels.api as sm
>>> import statsmodels.formula.api as smf
>>>
>>> galapagos = faraway.datasets.galapagos.load()
>>>
>>> form = ’Species ˜ Area + Elevation + Nearest + Scruz + Adjacent’
>>> form0 = ’Species ˜ Elevation + Nearest + Scruz’
>>> fit = smf.ols(form, galapagos).fit()
>>> fit0 = smf.ols(form0, galapagos).fit()

>>> sm.stats.anova_lm(fit0, fit)
  df_resid            ssr  df_diff       ss_diff         F   Pr(>F)
0      26.0  158291.628568      0.0           NaN       NaN      NaN
1      24.0   89231.366330      2.0  69060.262238  9.287352  0.00103
```

Anova (II) (R)

```
> library(faraway)
> data(gala)
> fit <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent, data = gala)
> fit0 <- lm(Species ~ Elevation + Nearest + Scruz, data = gala)
> anova(fit, fit0)
Analysis of Variance Table

Model 1: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
Model 2: Species ~ Elevation + Nearest + Scruz
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     24  89231
2     26 158292 -2    -69060 9.2874 0.00103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Information criterion

- □ Rather than using hypothesis test, one could rely on information criterion.
- □ Information criterion is just a numeric value that summarizes the overall quality of a fitted model. The lower the better.
- □ Two widely used information criterion are:
  - – The Akaike Information Criterion (AIC)

$$AIC(\mathcal{M}) = \underbrace{-2\ell(\hat\theta)}_{\text{goodness of fit}} + \underbrace{2\dim(\hat\theta)}_{\text{model complexity}}, \qquad \hat\theta \text{ MLE of model } \mathcal{M}.$$

  - – The Baysesian/Schwarz Information Criterion (BIC)

$$BIC(\mathcal{M}) = -2\ell(\hat\theta) + \dim(\hat\theta)\log n, \qquad \hat\theta \text{ MLE of model } \mathcal{M}.$$

Information criterion

- □ Rather than using hypothesis test, one could rely on information criterion.
- □ Information criterion is just a numeric value that summarizes the overall quality of a fitted model. The lower the better.
- □ Two widely used information criterion are:
  - – The Akaike Information Criterion (AIC)

$$AIC(\mathcal{M}) = \underbrace{-2\ell(\hat{\theta})}_{\text{goodness of fit}} + \underbrace{2\dim(\hat{\theta})}_{\text{model complexity}}, \qquad \hat{\theta} \text{ MLE of model } \mathcal{M}.$$

  - – The Baysesian/Schwarz Information Criterion (BIC)

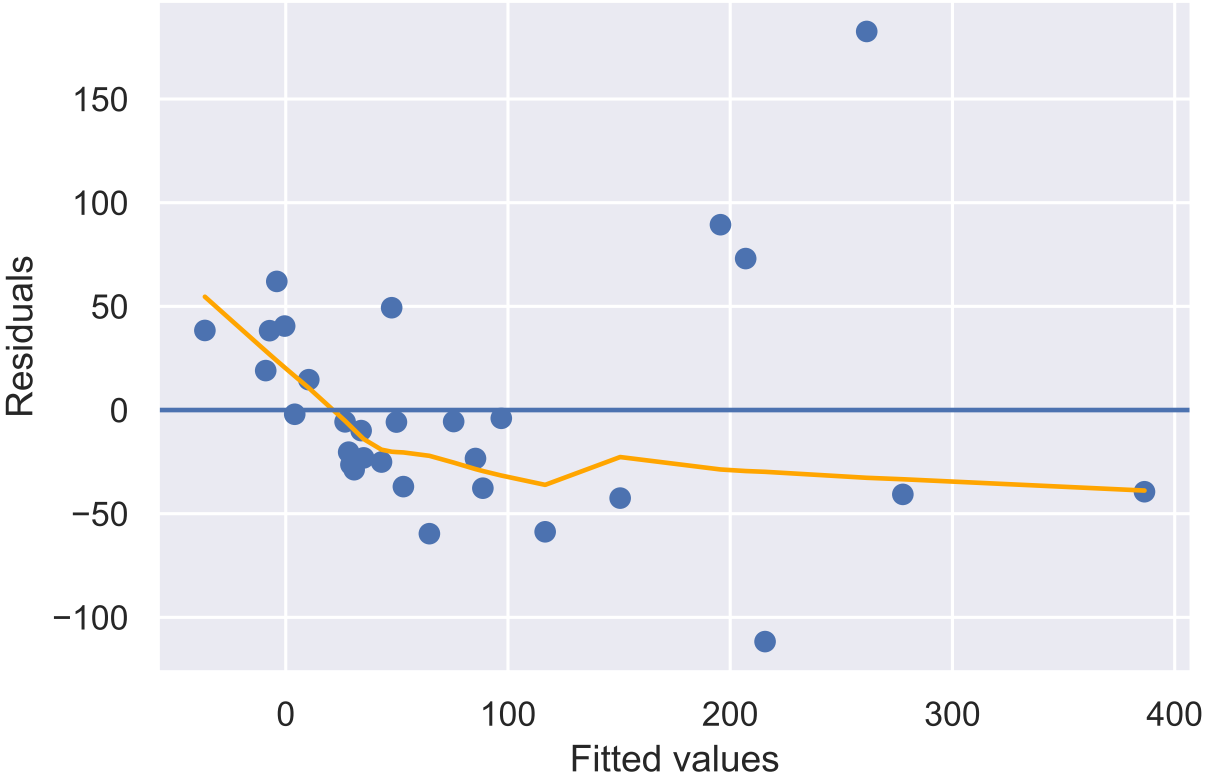
$$BIC(\mathcal{M}) = -2\ell(\hat{\theta}) + \dim(\hat{\theta}) \log n, \qquad \hat{\theta} \text{ MLE of model } \mathcal{M}.$$

☞ AIC and BIC have the advantage that it can be applied to non nested models! But beware AIC is not consistent while BIC is.

Residuals analysis

□ Typically we check if the model assumptions are valid using plots :

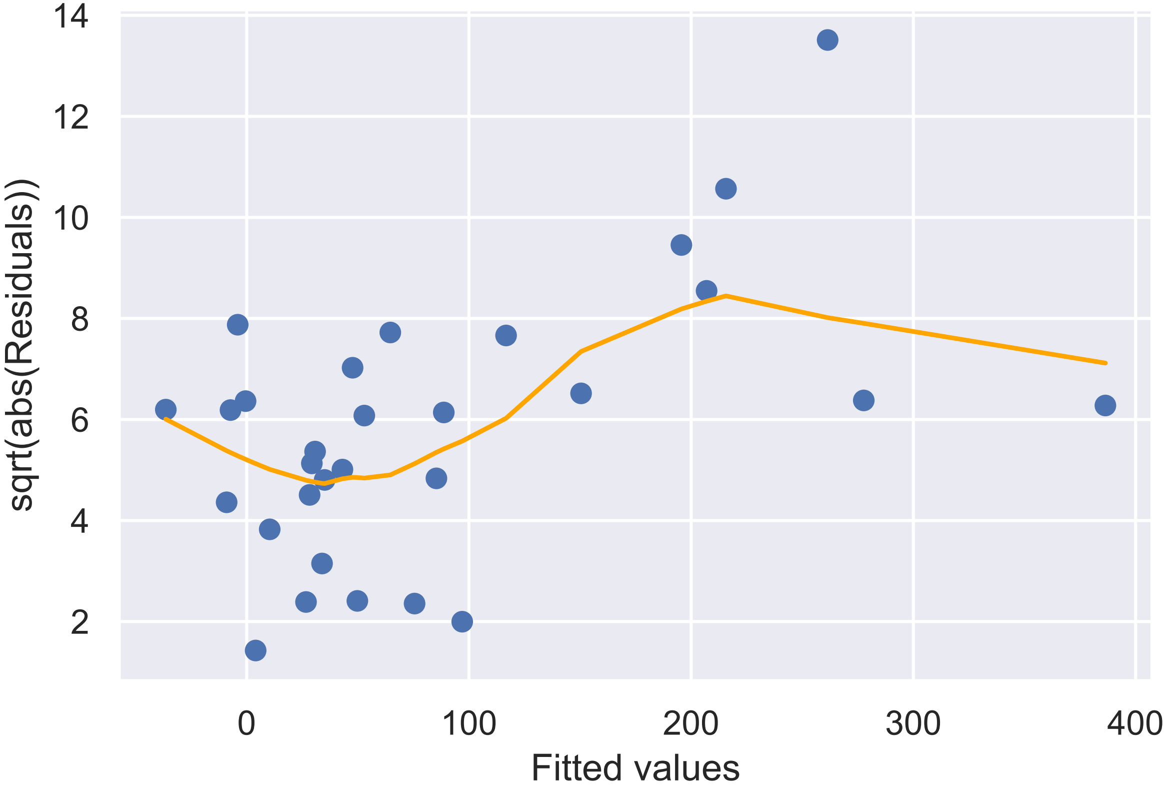
- white noise → plot residuals vs fitted values;
- homoscedasticity → plot $\sqrt{|\text{residuals}|}$ vs fitted values;
- Normality (if gaussian noise) using quantile-quantile plots



Residuals analysis

□ Typically we check if the model assumptions are valid using plots :

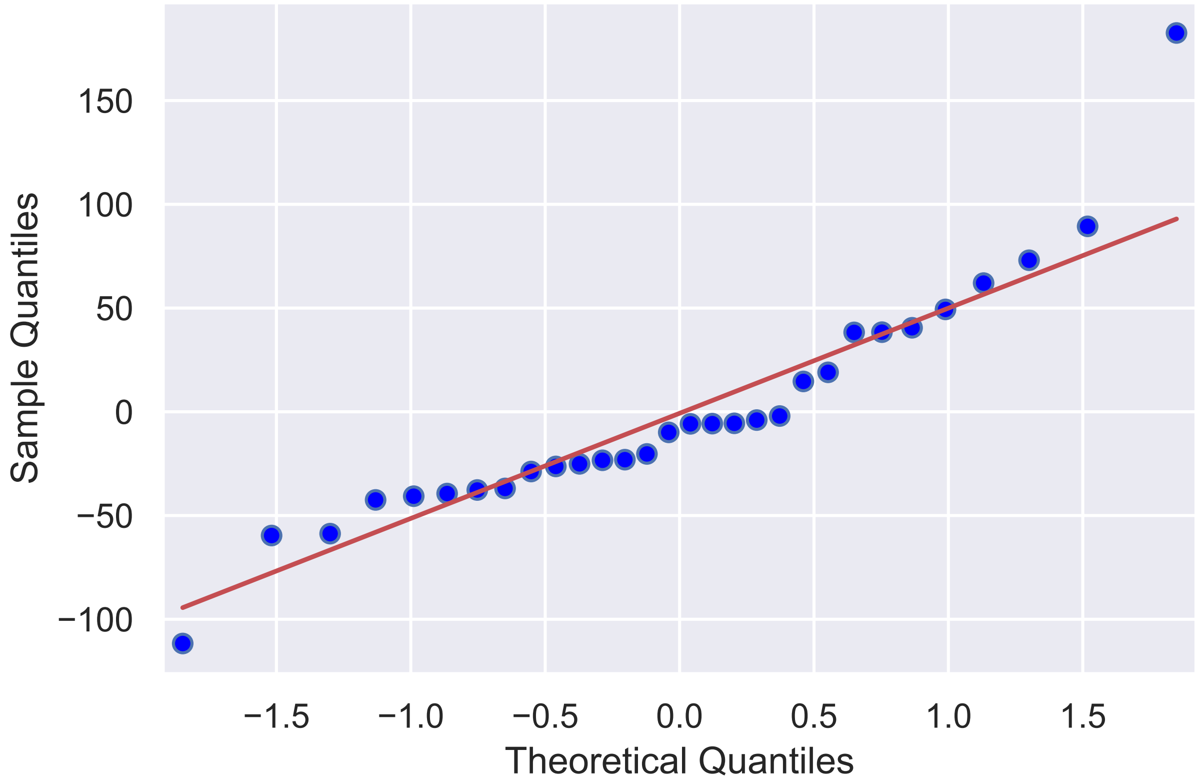
- white noise → plot residuals vs fitted values;
- homoscedasticity → plot $\sqrt{|\text{residuals}|}$ vs fitted values;
- Normality (if gaussian noise) using quantile-quantile plots



Residuals analysis

□ Typically we check if the model assumptions are valid using plots :

- white noise → plot residuals vs fitted values;
- homoscedasticity → plot $\sqrt{|\text{residuals}|}$ vs fitted values;
- Normality (if gaussian noise) using quantile-quantile plots

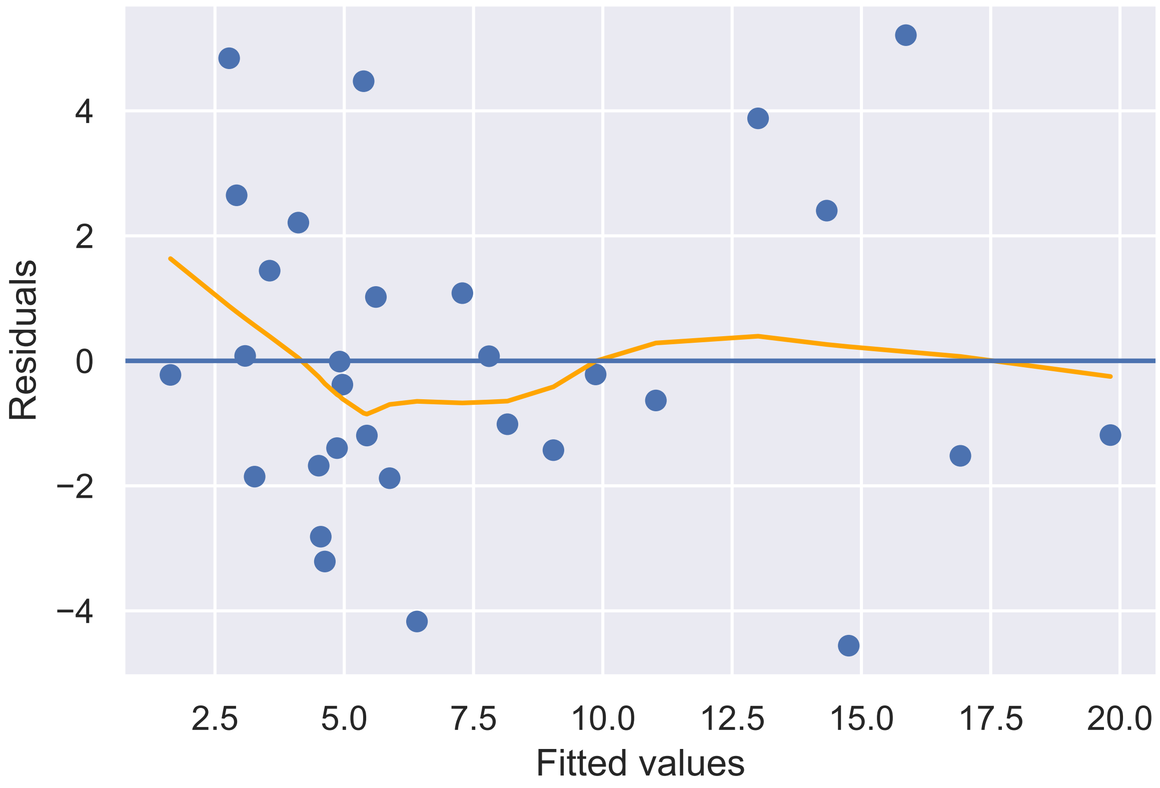


Galapagos revisited

- □ The two first diagnostic plots suggest problems.
- □ One way to fix it is to transform the response variable.
- □ Theory tells that a sensible transformation for counts is $y \mapsto \sqrt{y}$.

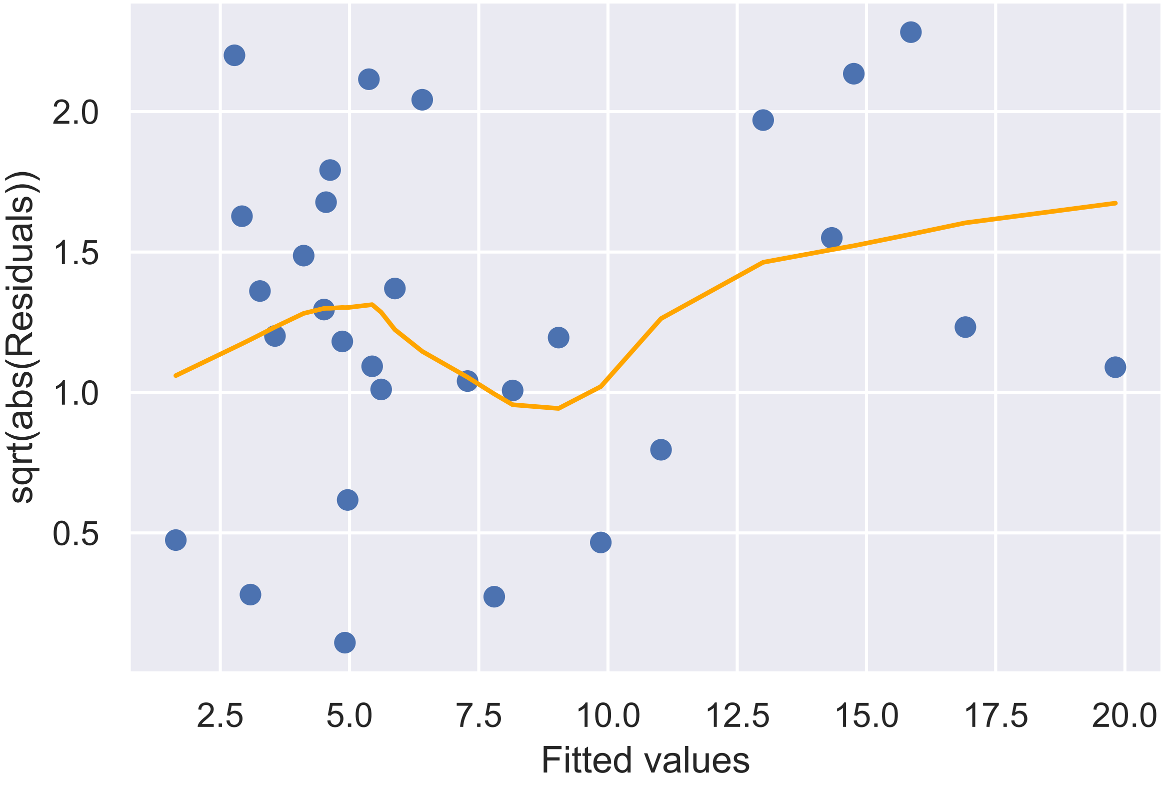
Galapagos revisited

- □ The two first diagnostic plots suggest problems.
- □ One way to fix it is to transform the response variable.
- □ Theory tells that a sensible transformation for counts is $y \mapsto \sqrt{y}$.



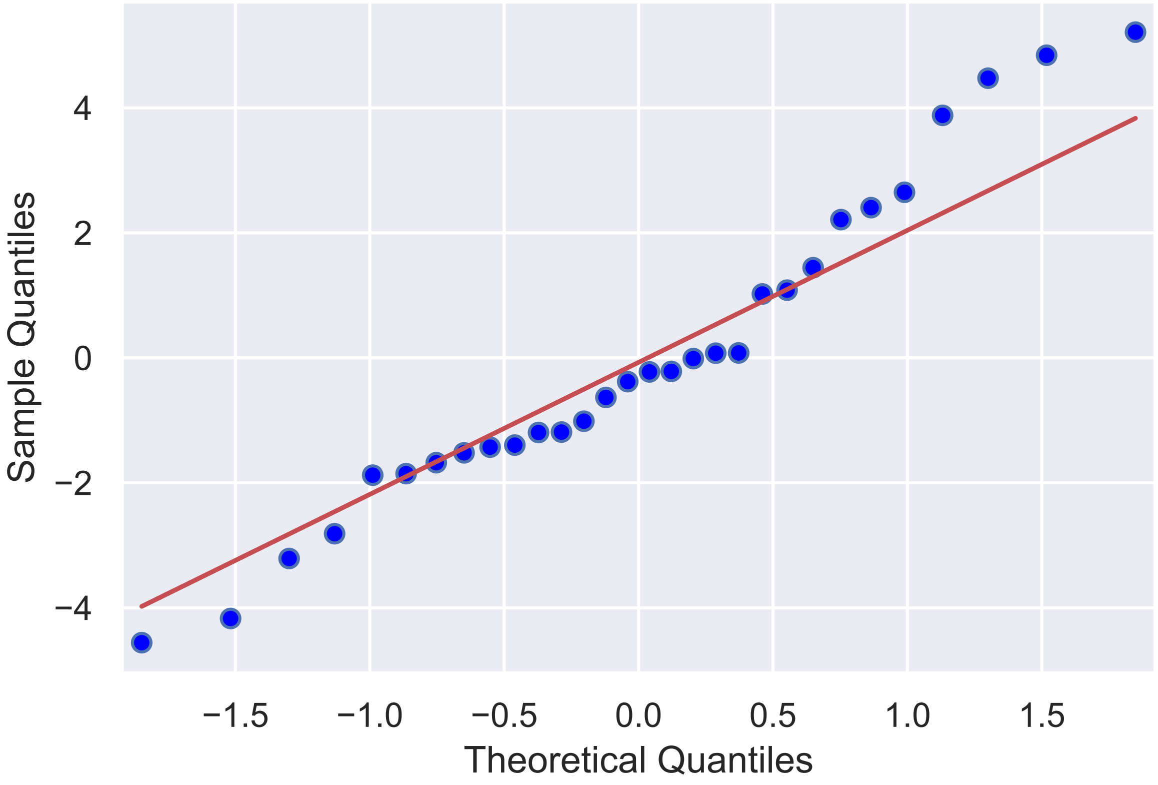
Galapagos revisited

- □ The two first diagnostic plots suggest problems.
- □ One way to fix it is to transform the response variable.
- □ Theory tells that a sensible transformation for counts is $y \mapsto \sqrt{y}$.



Galapagos revisited

- □ The two first diagnostic plots suggest problems.
- □ One way to fix it is to transform the response variable.
- □ Theory tells that a sensible transformation for counts is $y \mapsto \sqrt{y}$.



6. Logistic regression

Logistic regression

- □ Logistic regression is, to some extent, very similar to linear regression except that the response is binary, i.e., $Y \in \{0, 1\}$.
- □ Why this situation deserves a close attention?

Logistic regression

- □ Logistic regression is, to some extent, very similar to linear regression except that the response is binary, i.e., $Y \in \{0, 1\}$.
- □ Why this situation deserves a close attention?
- □ Because in many situations one want to have a binary response such as:
  - email is spam or not spam;
  - should I bring my jacket or not today?
  - should a bank grant a loan to you or not?
- □ Logistic regression is therefore often considered as a supervised classifier.

Logistic regression

- □ Logistic regression is, to some extent, very similar to linear regression except that the response is binary, i.e., $Y \in \{0, 1\}$.
- □ Why this situation deserves a close attention?
- □ Because in many situations one want to have a binary response such as:
  - email is spam or not spam;
  - should I bring my jacket or not today?
  - should a bank grant a loan to you or not?
- □ Logistic regression is therefore often considered as a supervised classifier.

☞ Logistic regression could be extented to more than 2 classes but most often different approaches are used in such situations.

Let's build the model together

□ The response Y is binary and a sensible choice to model Y is thus the Bernoulli(p) distribution whose p.m.f. is

$$m(y) = p^y(1-p)^{1-y}, \qquad y \in \{0,1\}, \quad p = \Pr(Y=1) = \mathbb{E}(Y) \in [0,1]$$

□ Now since it is sensible to let the probability of "success" p depends on some covariates x, we now have

$$Y \mid X = x \sim \text{Bernoulli}(p(x)).$$

□ Working in a parametric setting and paralleling the linear regression model, we may assume the linear form

$$p(x) = x^\top \beta.$$

☞ Clearly not relevant since $x^\top \beta \in \mathbb{R}$!

□ To bypass this hurdle we thus need to define a one–one mapping η such that

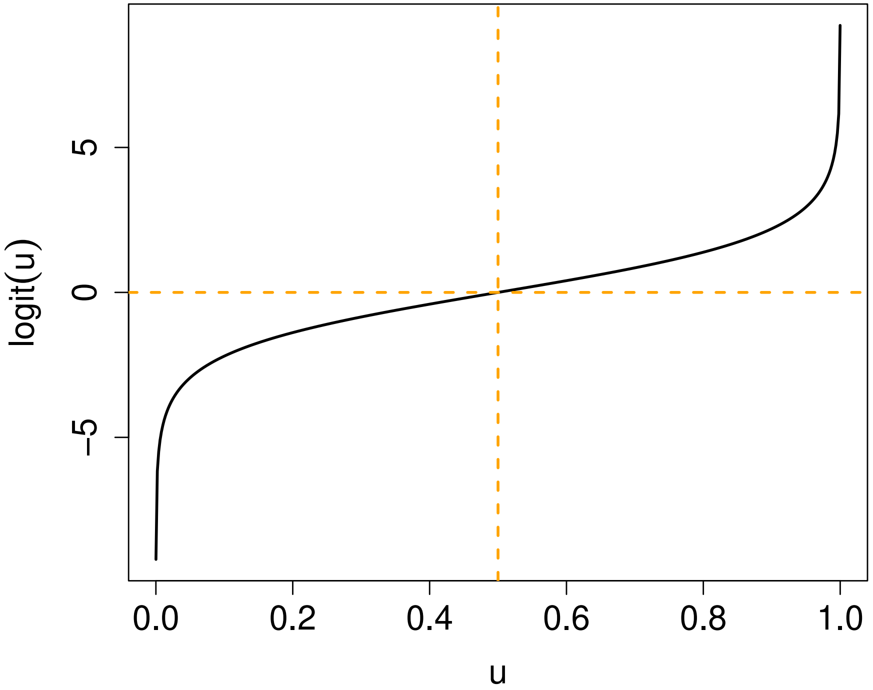
$$\begin{aligned} \eta \colon (0,1) &\longrightarrow \mathbb{R} \\ u &\longmapsto \eta(u) \end{aligned}$$

and set $\eta(p(x)) = x^\top \beta$.

□ Clearly the linear assumption on $\eta(p(x))$ now makes sense.

□ The logistic regression model assumes that η is the logit function, i.e.,

$$\begin{aligned} \operatorname{logit} \colon (0,1) &\longrightarrow \mathbb{R} \\ u &\longmapsto \log \frac{u}{1-u} \end{aligned}$$



An aside: Sigmoid function

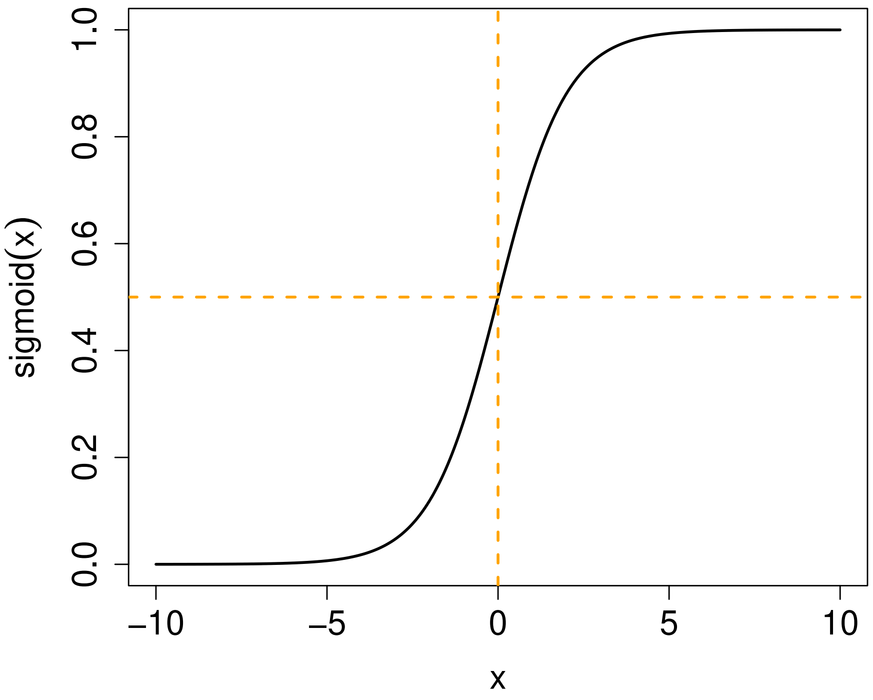
□ We just defined the logistic function

$$\begin{aligned} \operatorname{logit}\colon (0,1) &\longrightarrow \mathbb{R} \\ u &\longmapsto \log \frac{u}{1-u}. \end{aligned}$$

□ The reciprocal of the logit function is nowadays very popular due to the hype of Neural Networks.

□ It is known as the sigmoid function

$$\begin{aligned} \operatorname{sigmoid}\colon \mathbb{R} &\longrightarrow (0,1) \\ x &\longmapsto \log \frac{\exp(x)}{1+\exp(x)}. \end{aligned}$$



An aside: Generalized Linear Models

□ Actually both the linear and logistic regression models are special cases of Generalized Linear Models (GLM), i.e.,

$$\eta \left\{\mathbb{E}(Y \mid X)\right\} = x^\top \beta,$$

where η is the link function.

□ Here are some example of link functions and the corresponding model:

Linear $\eta(u) = u$
Logistic $\eta(u) = \operatorname{logit} u$
Poisson $\eta(u) = \log u$
Gamma $\eta(u) = -u^{-1}$

Fitting a logistic regression model

- □ Apart from the trivial case $\operatorname{logit} p(x) = \beta_0$, there is no closed form expression for the MLE;
- □ Gradient based optimization is typically used—most often Newton–Raphson that makes use of the Hessian matrix, i.e.,

$$\theta_{t+1} = \theta_t + \left\{\nabla^2_\theta \ell(\theta_t; \mathscr{D}_n)\right\}^{-1} \nabla_\theta \ell(\theta_t; \mathscr{D}_n),$$

  where $\nabla^2_\theta \ell(\theta_t; \mathscr{D}_n)$ is the Hessian matrix of $\ell(\theta_t; \mathscr{D}_n)$.
- □ Where for this particular model we have

$$\begin{aligned} \nabla_\theta \ell(\theta; \mathscr{D}_n) &= \mathbf{X}^\top \{\mathbf{Y} - p(\mathbf{X})\} \\ \nabla^2_\theta \ell(\theta; \mathscr{D}_n) &= -\mathbf{X}^\top \mathbf{W} \mathbf{X}, \end{aligned}$$

  where $\mathbf{W}$ is a diagonal matrix whose diagonal is $p(\mathbf{X})\{1 - p(\mathbf{X})\}$.

☞ The above algorithm is known as the Fisher's scoring algorithm.

Predictions

□ There are two types of predictions in a logistic regression model:

response predictor which estimates $p(x)$ using $\hat{p}(x) = \text{sigmoid}(x^\top \hat{\beta})$;
linear predictor which predicts $\text{logit}\, p(x) = x^\top \beta$ using $x^\top \hat{\beta}$.

□ Both can serve as a guideline to predict the outcome Y given $X = x$.

□ More precisely we use the following (binary) classifier

$$\hat{Y} \mid \{X = x\} = 1_{\{\hat{p}(x) > u\}} = 1_{\{x^\top \hat{\beta} > \text{logit}\, u\}},$$

where u is a given threshold, i.e., most often but not invariably $u = 0.5$.

☞ In some cases you might not want to have to many “false alarms”, i.e, $\hat{Y} = 1$ while $Y = 0$. Think about a spam filter. You can achieve this by increasing u, e.g., $u = 0.8$.

Residuals analysis

□ Recall that residuals are given by

$$r_i = Y_i - \hat{Y}_i, \qquad i = 1, \dots, n$$

□ However since Y is binary, we thus have $r_i \in \{-1, 0, 1\}$ which is unfortunate to do diagnostic plots (but see later).

□ Hence for logistic regression we rather define residuals as

$$r_i = Y_i - x^\top \hat{\beta}.$$

□ Note however that there is still a side effect since

$$r_i = \begin{cases} 1 - x^\top \hat{\beta}, & Y_i = 1 \\ -x^\top \hat{\beta}, & Y_i = 0 \end{cases}$$

and thus provides artificial patterns.

What are the odds?

Definition 9. Given a probability p of some events, the associated odds are given by

$$\text{odds}(p) = \frac{p}{1-p} \in (0, \infty).$$

- □ The odds helps in dermining if an event having probability p to occcur is likely or not.
- □ More precisely,
  - $\text{odds}(p) > 1$ indicates the event is more likely to occur than it does not;
  - $\text{odds}(p) < 1$ indicates the event is less likely to occur than it does not.

Odds in a logistic regression model: quantitative case

□ Recall that in logistic regression we have $p(x) = \Pr(Y = 1 \mid X = x)$.

□ Hence

$$\text{odds}(p(x)) := \text{odds}(x) = \frac{p(x)}{1 - p(x)} = \exp\{\operatorname{logit} p(x)\} = x^\top \beta.$$

□ A typical interpretation of odds is when you add a “one unit increase” in quantitative covariate x_j and state how increased/decreased the odds.

□ Indeed let $x_* = x$ except for, say, the p–th element which is $x_{*,p} = x_p + 1$. We get

$$\text{odds}(x_*) = \exp\left(x_*^\top \beta\right) = \exp\left(x^\top \beta + \beta_p\right) = \text{odds}(x) \exp(\beta_p).$$

☞ Depending on the sign of β_p, and all other covariates being fixed, we can tell if one unit increase in x_p increases the odds or not and even quantify the change from $\exp(\beta_p)$.

Odds in a logistic regression model: qualitative case

□ For categorical variables the "one unit increase" has no sense, think about "blue + 1"!

□ We can however still interpret the effect of a categorical variable, say x_p, on the odds.

□ To this aim we first fix a reference level for x_p, e.g., blue.[7]

□ Recall that if x_p has m levels, i.e., $x_p \in \{1, \ldots, m$, $x^\top \beta_p$ actually reads

$$\beta_{p,2} 1_{\{x_p=2\}} + \cdots + \beta_{p,m} 1_{\{x_p=m\}}.$$

□ In the above expression level 1 is the baseline level.

☞ As previously, $\exp(\beta_{p,\ell})$ quantify the changes on the odds as we switch from baseline level to the ℓ–th one.

[7] It is also needed to ensure identifiability of the model parameters.

South African Heart Disease (Rousseauw et al., 1983)

□ Coronary risk factor study survey carried out in 3 rural areas of the Western Cape in South Africa

□ Aim: Establish the intensity of coronary heart disease (`chd`) factors in that high incidence region

□ Data : While males between 15 and 64 and response variable is the presence or absence of myocardal infarction (MI)

□ Overall prevalence in this region is 5.1%

□ There are 160 cases in our data set and a sample of 302 controls.

□ The main motivation with this study was to educate people to have a balanced diet

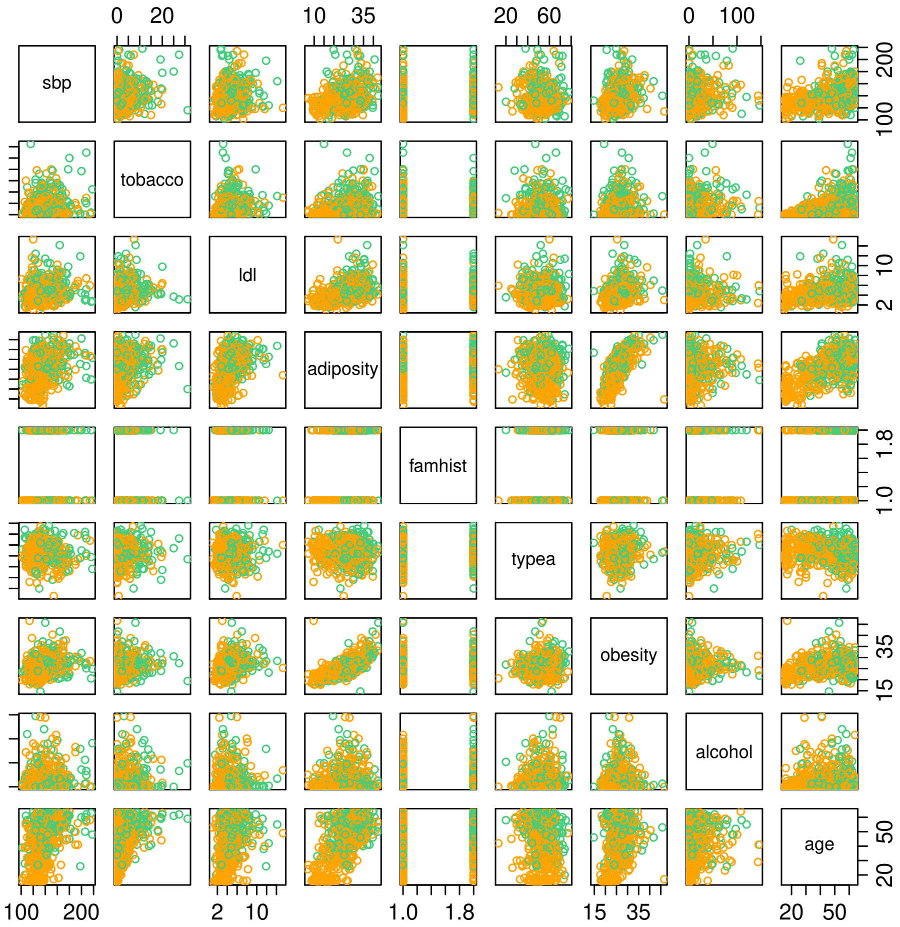


Figure 23: *Scatterplot of the South African heart disease dataset. Green: MI; col1: Control;* `famhist`*: 1 if family history of heart disease.*

```
> fit <- glm(chd ˜ ., data = data, family = binomial)
> summary(fit)

Call:
glm(formula = chd ˜ ., family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7781  -0.8213  -0.4387   0.8889   2.5435

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.1507209  1.3082600  -4.701 2.58e-06 ***
sbp             0.0065040  0.0057304   1.135 0.256374
tobacco         0.0793764  0.0266028   2.984 0.002847 **
ldl             0.1739239  0.0596617   2.915 0.003555 **
adiposity       0.0185866  0.0292894   0.635 0.525700
famhistPresent  0.9253704  0.2278940   4.061 4.90e-05 ***
typea           0.0395950  0.0123202   3.214 0.001310 **
obesity        -0.0629099  0.0442477  -1.422 0.155095
alcohol         0.0001217  0.0044832   0.027 0.978350
age             0.0452253  0.0121298   3.728 0.000193 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.14  on 452  degrees of freedom
AIC: 492.14

Number of Fisher Scoring iterations: 5
```

□ The results sound a bit weird. . .

- □ The results sound a bit weird. . . Systolic blood pressure (`sbp`) is not significant??!!!???
- □ Idem for `obesity` which in addition is negative!

- □ The results sound a bit weird. . . Systolic blood pressure (`sbp`) is not significant??!!!???
- □ Idem for `obesity` which in addition is negative!
- □ This is a consequence of correlation between covariates. Need proof?

□ The results sound a bit weird. . . Systolic blood pressure (`sbp`) is not significant??!!!???

□ Idem for `obesity` which in addition is negative!

□ This is a consequence of correlation between covariates. Need proof?

```
> summary(glm(chd ~ obesity, data = data, family = binomial))

Call:
glm(formula = chd ~ obesity, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3396  -0.9257  -0.8558   1.4021   1.7116

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.92831    0.61692  -3.126  0.00177 **
obesity      0.04942    0.02318   2.132  0.03302 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 591.53  on 460  degrees of freedom
AIC: 595.53

Number of Fisher Scoring iterations: 4
```

Lesson to be learned

- You should interpret with caution non-significance of group of covariates.
- Ideally you should remove sequentially the least significant covariate until you couldn’t drop anything
- Or, if you’re a bit reckless, use `stepAIC` or variants

```
> library(MASS)
> fit.step <- stepAIC(fit)
> summary(fit.step)

Call:
glm(formula = chd ~ tobacco + ldl + famhist + typea + age, family = binomial,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9165  -0.8054  -0.4430   0.9329   2.6139

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.44644    0.92087  -7.000 2.55e-12 ***
tobacco         0.08038    0.02588   3.106  0.00190 **
ldl             0.16199    0.05497   2.947  0.00321 **
famhistPresent  0.90818    0.22576   4.023 5.75e-05 ***
typea           0.03712    0.01217   3.051  0.00228 **
age             0.05046    0.01021   4.944 7.65e-07 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Interpretation

```
> summary(fit.step)

Call:
glm(formula = chd ~ tobacco + ldl + famhist + typea + age, family = binomial,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9165  -0.8054  -0.4430   0.9329   2.6139

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.44644    0.92087  -7.000 2.55e-12 ***
tobacco         0.08038    0.02588   3.106  0.00190 **
ldl             0.16199    0.05497   2.947  0.00321 **
famhistPresent  0.90818    0.22576   4.023 5.75e-05 ***
typea           0.03712    0.01217   3.051  0.00228 **
age             0.05046    0.01021   4.944 7.65e-07 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- How do we interpret for instance `famhistPresent`?
- If a patient has a family history heart disease, it increases the odds of coronary heart disease of $\exp(0.90818) \approx 2.5$ or equivalently 150%.
- And a 95% confidence interval for this odds ratio is

$$\exp(0.90818 \pm 1.96 \times 0.22576) \approx [2, 3].$$

Logistic regression as a binary classifier

- Remember that the outcome Y for the logistic regression is binary.
- We suppose as well as the probability of “success”, i.e., having $Y = 1$, depends on some covariates x as follows

$$\operatorname{logit} p(x) = x^T \beta, \qquad p(x) = \Pr(Y = 1 \mid X = x).$$

- Given some features x_*, how could we say that Y should be 1 or 0?
- One widely used way is to take

$$\hat{Y} = \begin{cases} 1, & p(x) \geq 0.5 \\ 0, & p(x) < 0.5. \end{cases}$$

Remark. The cutoff value $u = 0.5$ is arbitrary[8] and, depending on the application, one could use different levels $u \in (0, 1)$. Think about fraud detection.

[8] but has theoretical justifications