Principal Component Analysis

Vincent Guillemot

Dec. 2

First things first

We will need the packages `corrplot`, `FactoMineR` and `factoextra`:

- Check that `corrplot`, `FactoMineR` and `factoextra` are installed
- If not, install them, then load them

```
library(corrplot)
library(FactoMineR)
library(factoextra)
```

We also need to load `dplyr`, `tidyr` and the “fruits” data, :

```
library(dplyr)
library(tidyr)
data("fruits", package = "ReMUSE")
```

Unsupervised/Exploratory vs Supervised Analysis

- In the supervised methods, we observe both a set of features/variables (e.g. gene expression) for each object, as well as a response or outcome variable (e.g. metastasis information or survival information of the patients).
- The goal is then to predict the response using the variables (e.g. which genes predict best or are associated with the survival of the patients).
- Here we instead focus on exploratory analysis, we are not interested (yet) in prediction.

The two exploratory methods we are interested in…

- **Principal Component Analysis**, whose goal is to extract the most "important" sources of variability in the quantitative data
- **Hierachical Agglomerative Clustering**, whose goal is to define *clusters* of samples

Program for this morning

Covariance and correlation :

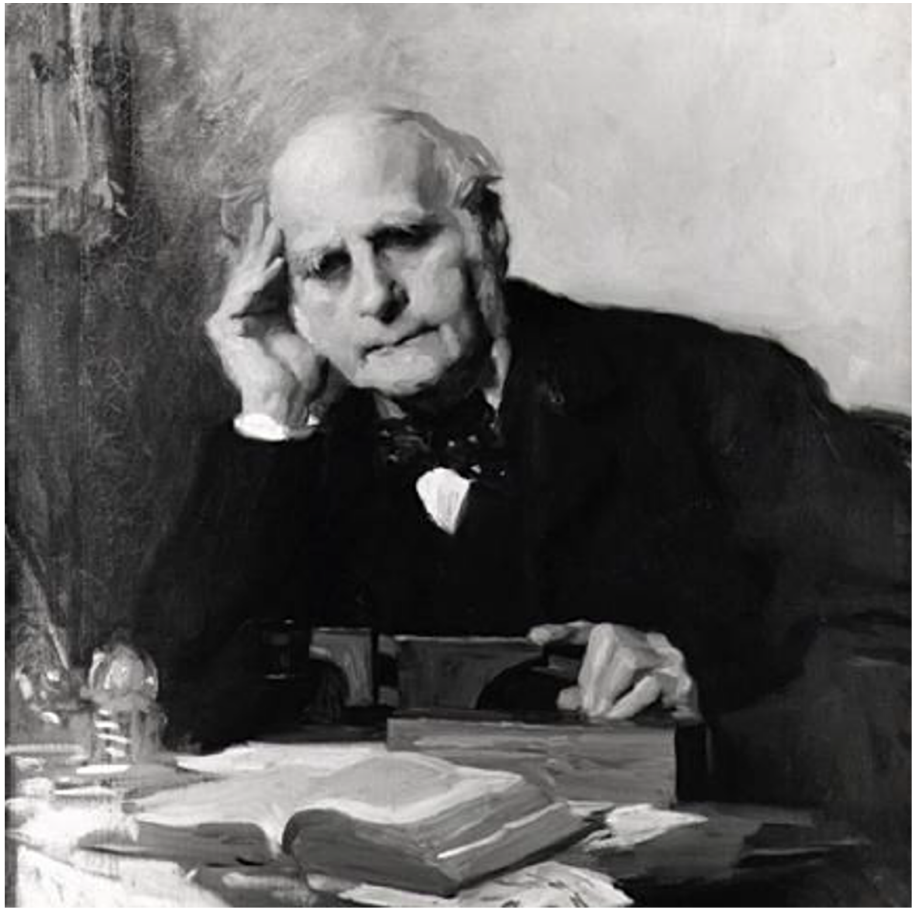
- reminders, and/or not
- visualization

Principal Component Analysis:

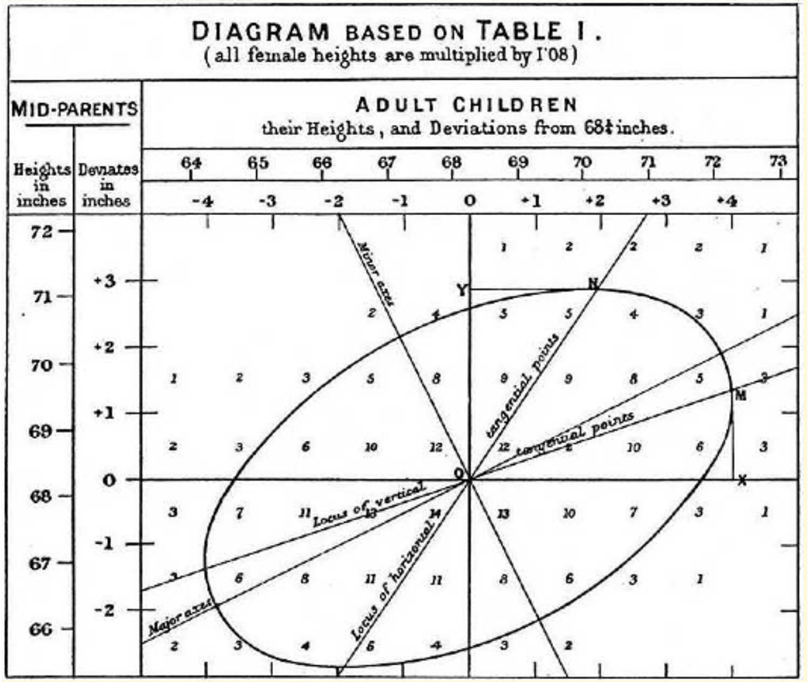
- principle (vocabulary!), method, a little bit of theory
- PCA with `FactoMineR` and visualization with `factoextra`

Covariance and correlation

A little bit of history

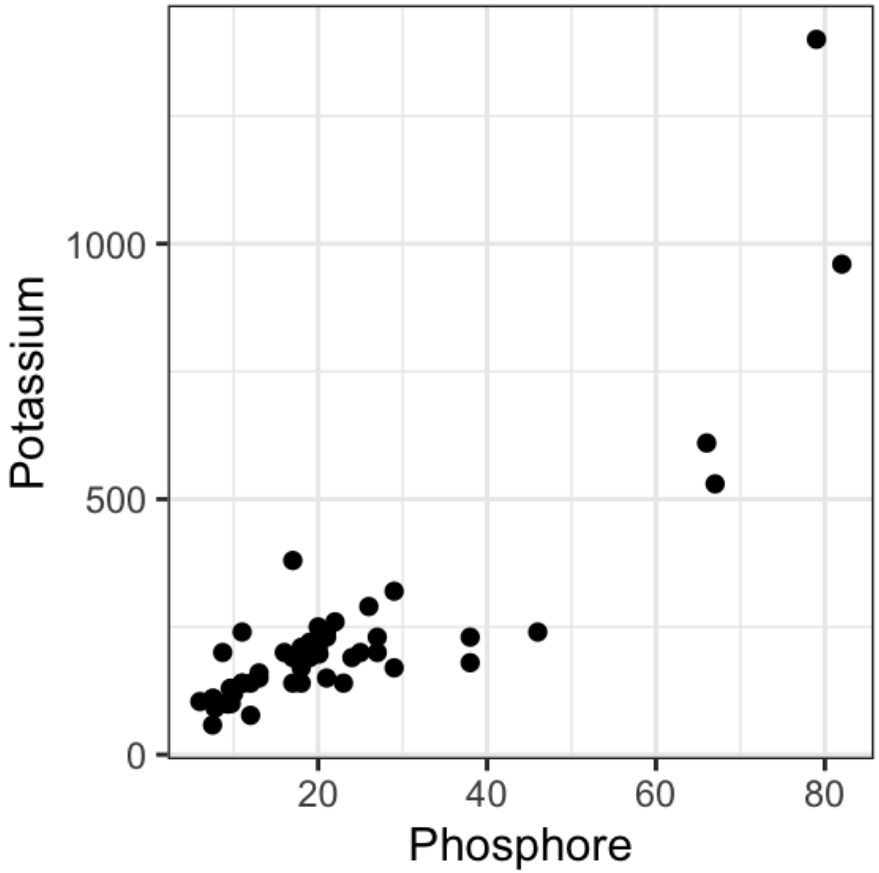


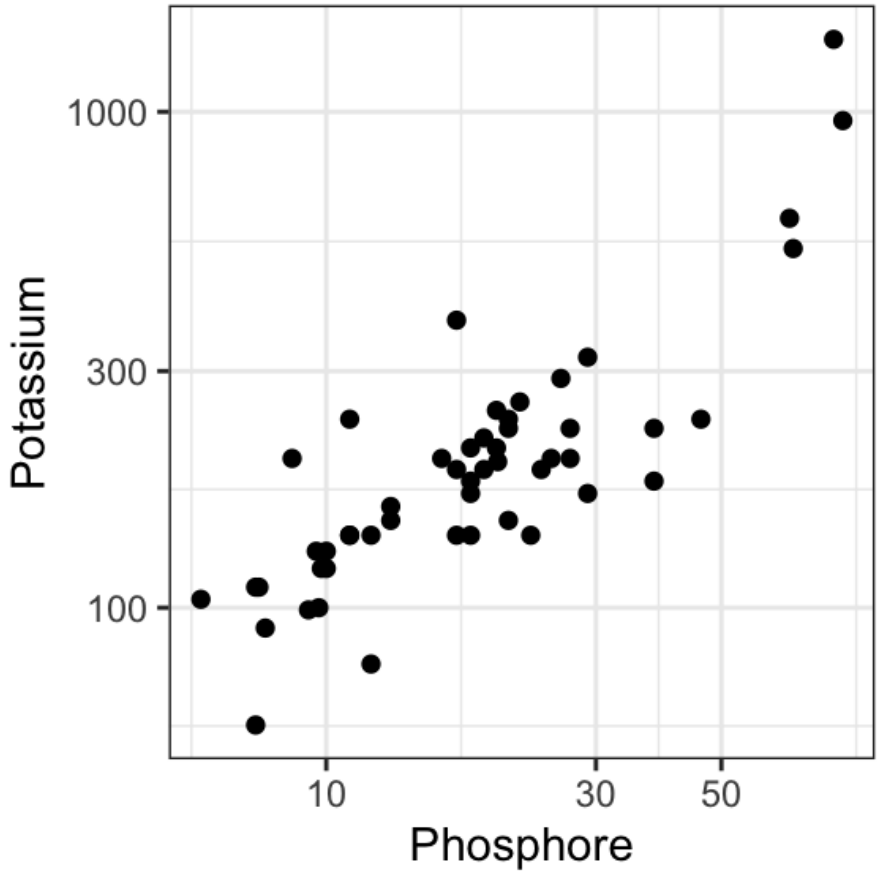
Sir Francis Galton
(1822-1911)



(A Kind of) PCA Comparing Mid-Parental to Adult Children Height

On fruits

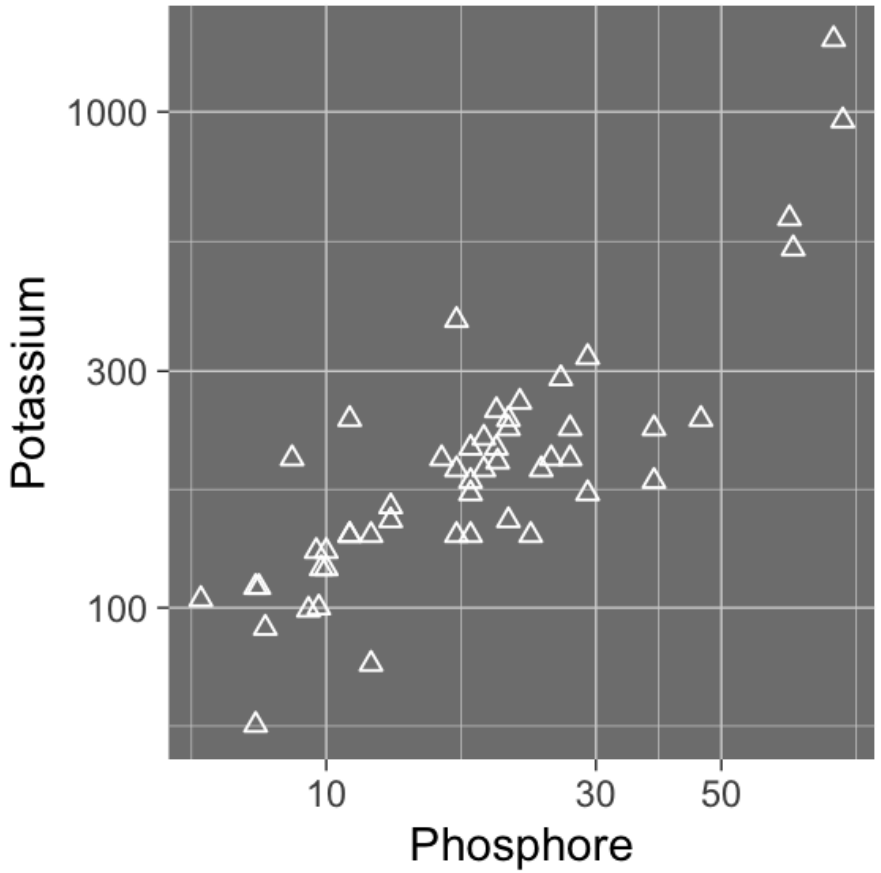

Potassium
1000
500
0
20
40
60
80
Phosphore


Potassium
1000
300
100
10
30
50
Phosphore

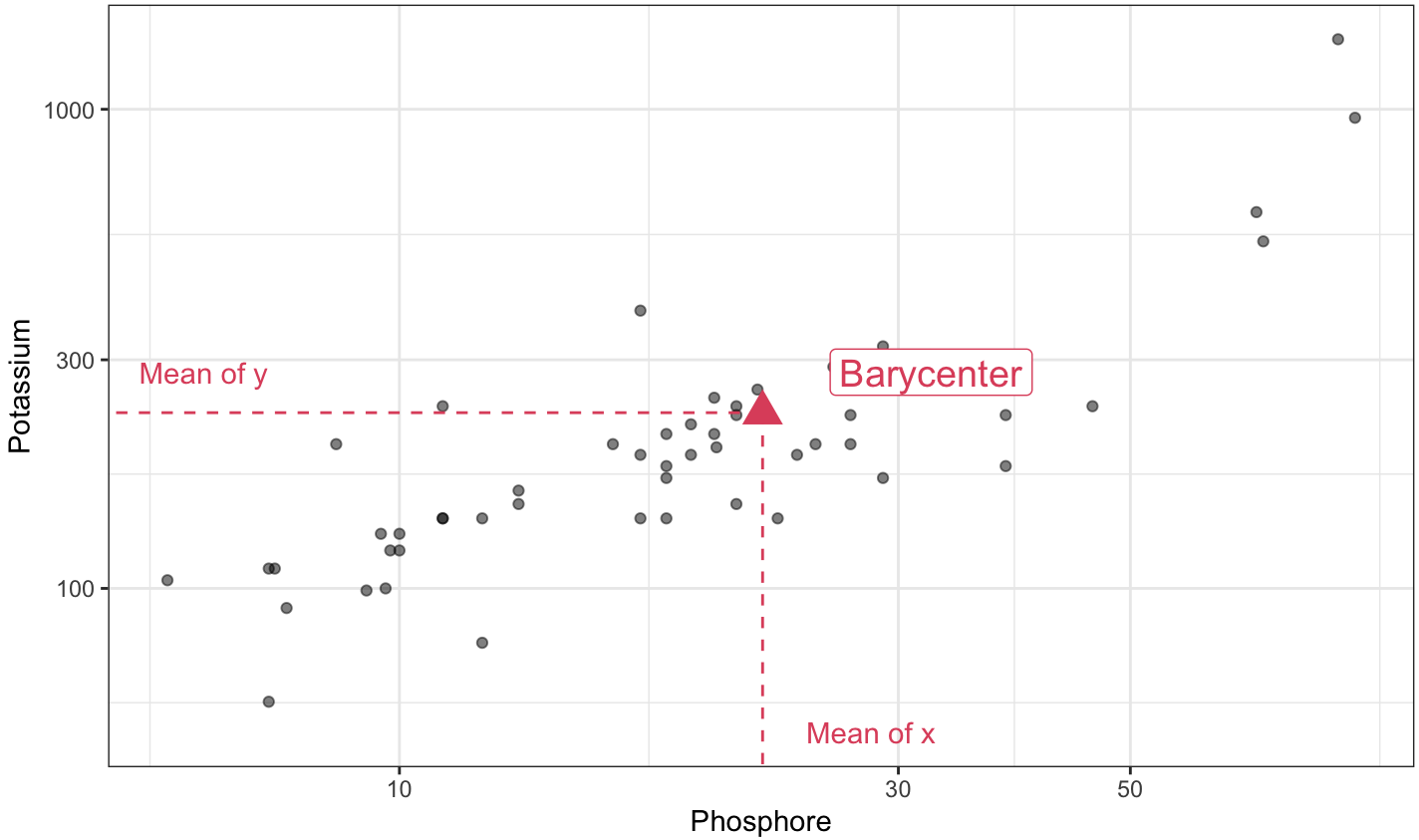
Exercise

Make a graph as similar as the one the right!

Try to focus on the essential.



Barycenter (reminder)



Covariance (reminder)

How far away is a “dot” from the barycenter ? Individual rectangle area :

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

The covariance is (almost) the mean area:

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \times (y_i - \bar{y})$$

Correlation coefficient

Covariance can vary between $-\infty$ and $+\infty$.

Correlation is, by definition, a measure of linear relationship between -1 and +1:

- -1 represents a perfect negative linear relationship,
- 0 represents no linear relationship,
- +1 represents a perfect positive linear relationship.

Pearson's correlation coefficient

$$\operatorname{cor}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

… in short…

$$\operatorname{cor}(x, y) = \frac{\operatorname{cov}(x, y)}{\operatorname{sd}(x)\operatorname{sd}(y)}$$

“Exercises”

Covariance between two variables with R:

```
cov(fruits$Potassium,
    fruits$Phosphore)
#> [1] 3315.292
```

Correlation between two variables with R:

```
cor(fruits$Potassium,
    fruits$Phosphore)
#> [1] 0.8587945
```

Exercise / challenge !

Compute the correlation between `fruits$Potassium` and `fruits$Phosphore`.

Constraints!

Use **only** the following functions/operations:

- `length` to compute n,
- `*` to multiply, `/` to divide, `-` to substract,
- `mean` to compute the mean,
- `sd` to compute the standard-deviation,
- `sum` for the "$\sum_{i=1}^{n}$"
- as many brackets as you need

Spearman's correlation coefficient

Often noted ρ. Same as Pearson's, but on the ranks! Let:

- r_x be the ranks of x,
- r_y be the ranks of y.

$$\rho(x, y) = \text{cor}(r_x, r_y)$$

Key properties:

- not sensitive to extreme values,
- invariant by monotonous (increasing) transformation (e.g. log, square-root),
- not adapted when there are ex-aequos

“Exercises”

With the argument `method` set to `"spearman`:

```
cor(fruits$Potassium,
    fruits$Phosphore,
    method = "spearman")
#> [1] 0.770185
```

Same, but with the `rank` function:

```
cor(rank(fruits$Potassium),
    rank(fruits$Phosphore))
#> [1] 0.770185
```

Kendall's correlation coefficient

Pairs of points: select two samples i and j.

- Concordant Pair : $\left(x_i < x_j \text{ et } y_i < y_j\right)$ OU $\left(x_i > x_j \text{ et } y_i > y_j\right)$
- Discordant Pair : $\left(x_i < x_j \text{ et } y_i > y_j\right)$ OU $\left(x_i > x_j \text{ et } y_i < y_j\right)$

$$\tau(x, y) = \frac{n_C - n_D}{n_0},$$

where n_C is the number of concordant pairs, n_D the number of discordant pairs and n_0 total number of pairs.

“Exercise”: compare and comment

Pearson: nice linear relationship.

```
cor(fruits$Potassium,
    fruits$Phosphore,
    method = "pearson")
#> [1] 0.8587945
```

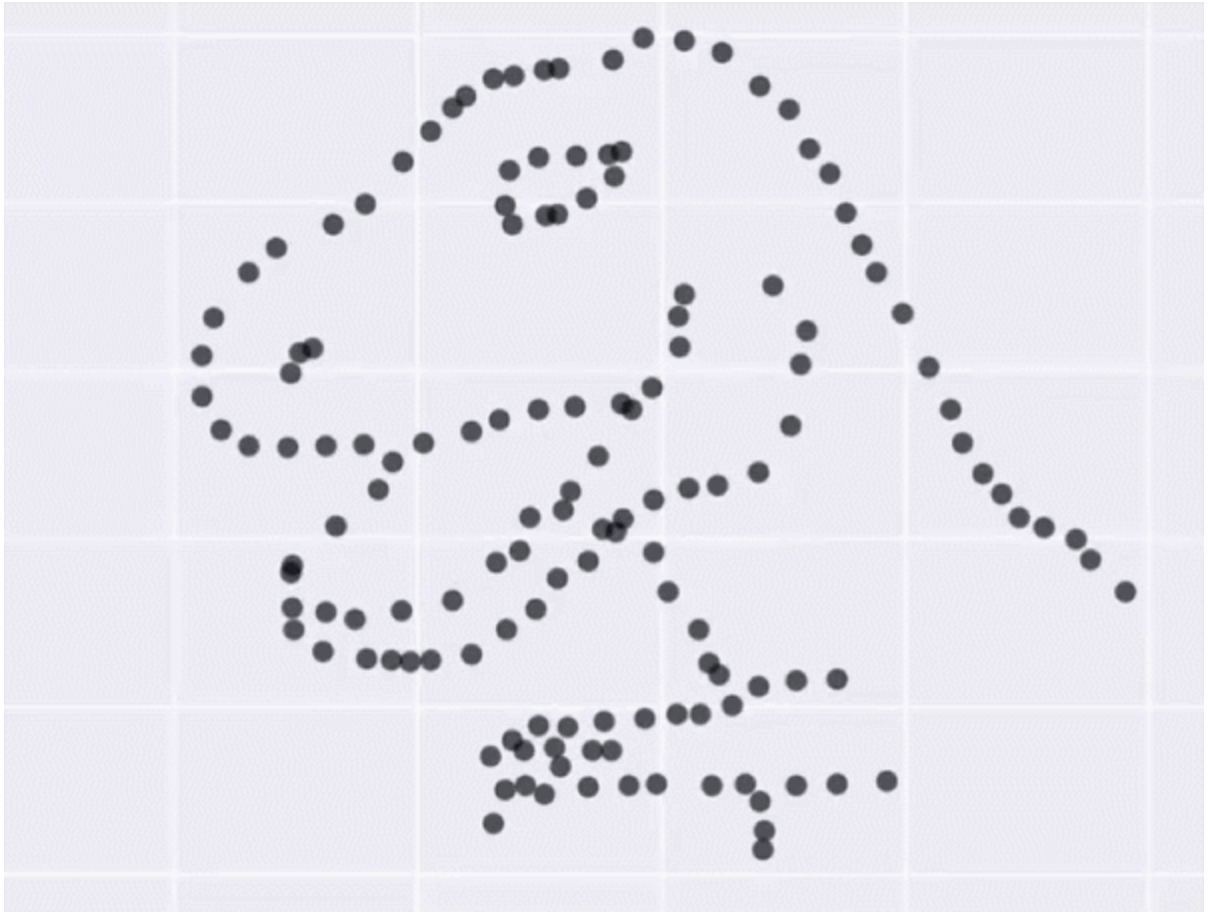
Kendall: ex-aequos.

```
cor(fruits$Potassium,
    fruits$Phosphore,
    method = "kendall")
#> [1] 0.6227486
```

Spearman: relationship is non-linear but monotonous.

```
cor(fruits$Potassium,
    fruits$Phosphore,
    method = "spearman")
#> [1] 0.770185
```

Beware of naked numbers!



Datasaurus

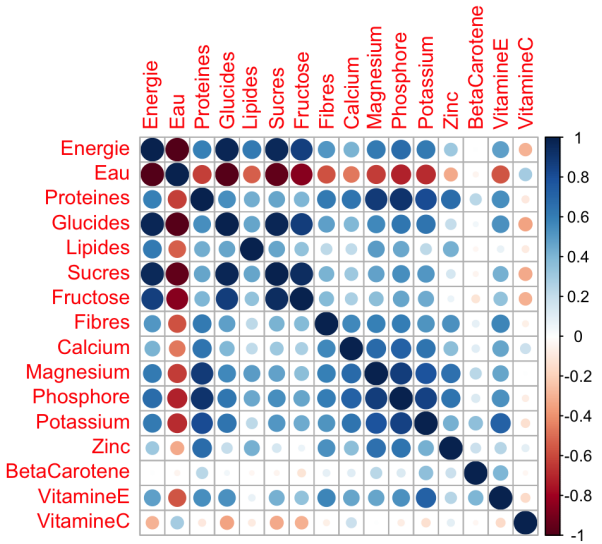
Plot the correlation

To compute all correlations:

```
cormat <- cor(fruits[, -(1:2)])
```

To make a “correlogram”:

```
corrplot(cormat)
```



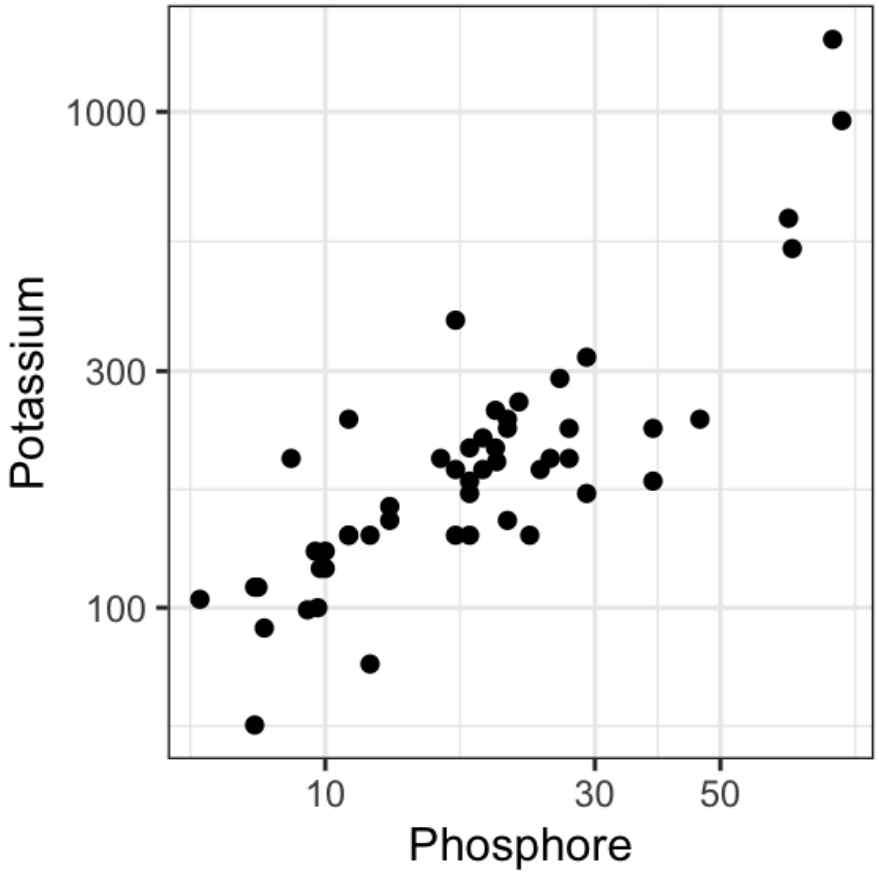
Exercise

Make a correlogram on the fruit data and change the color of the labels.

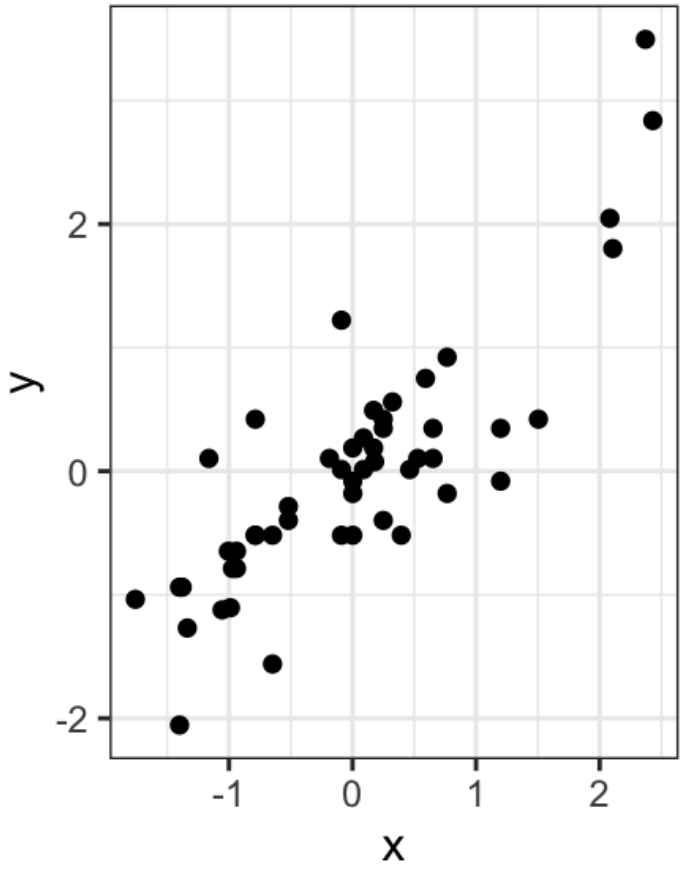
PCA on a two-dimensional data-set

Recall the first graph

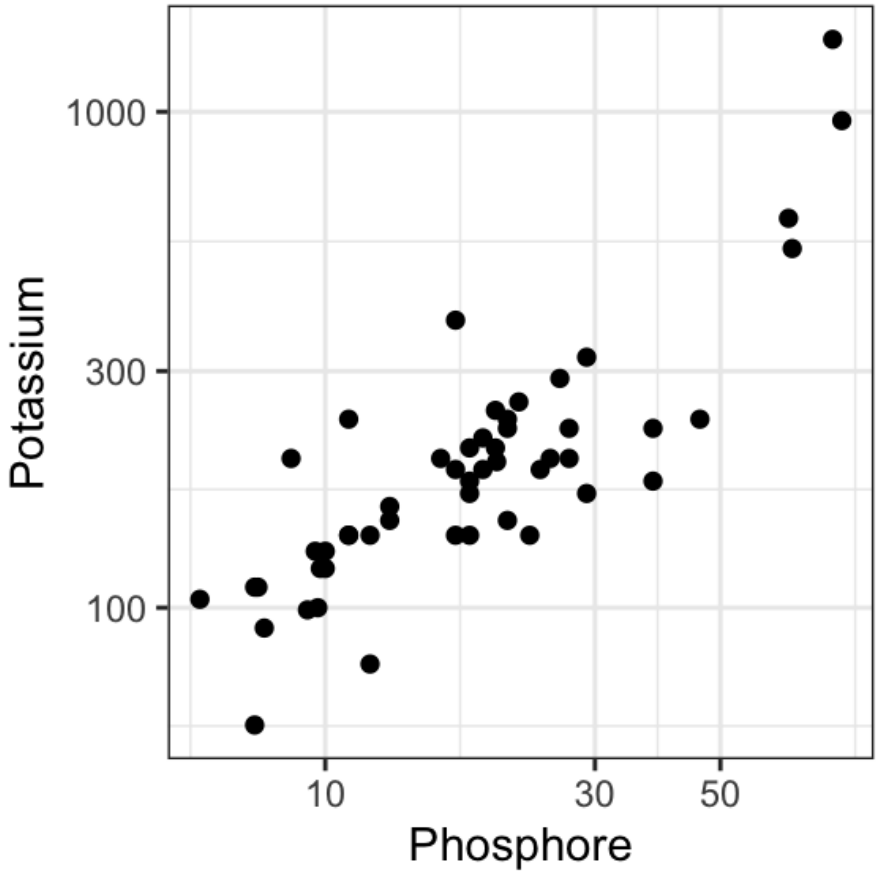
The real data:

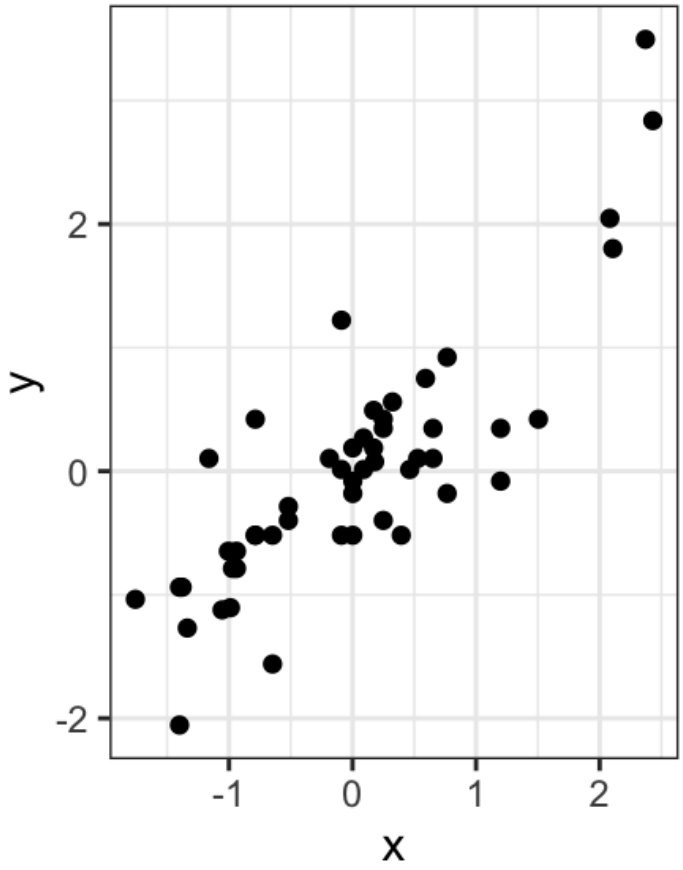


To make my life easier



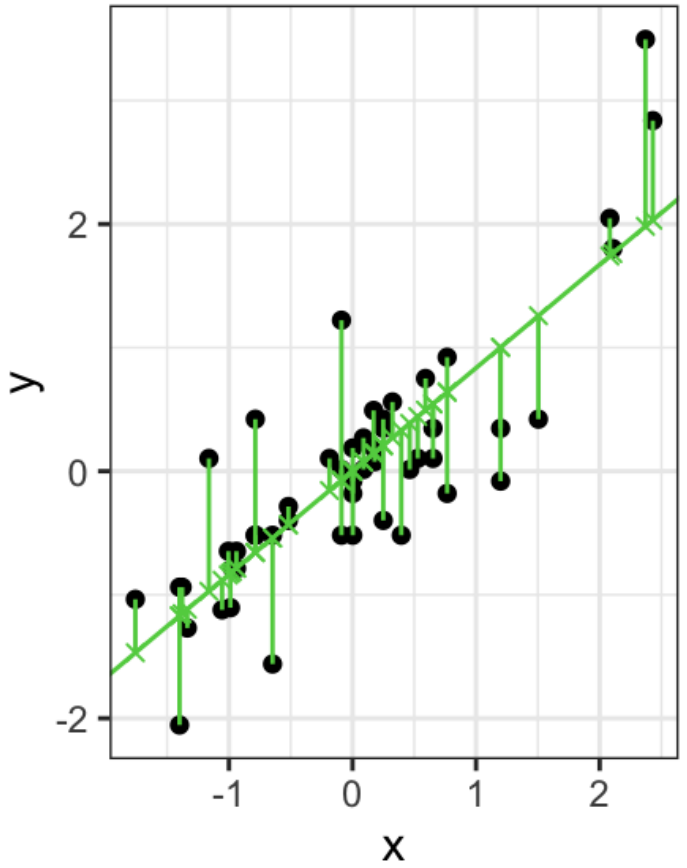
Exercise / discussion: What did I do?





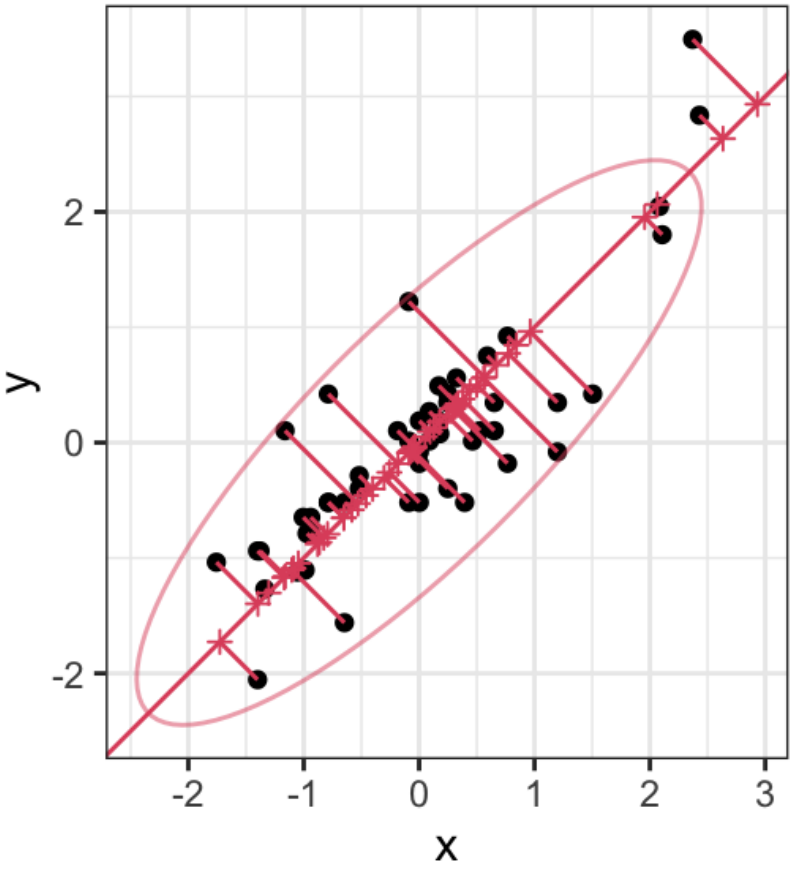
Linear Regression, aka Linear Model

- Model: $y = ax + b$
- $a = \frac{\text{cov}(x,y)}{\text{var}(x)}$ is the slope of the line,
- $b = \bar{y} - a\bar{x}$ is the intercept.

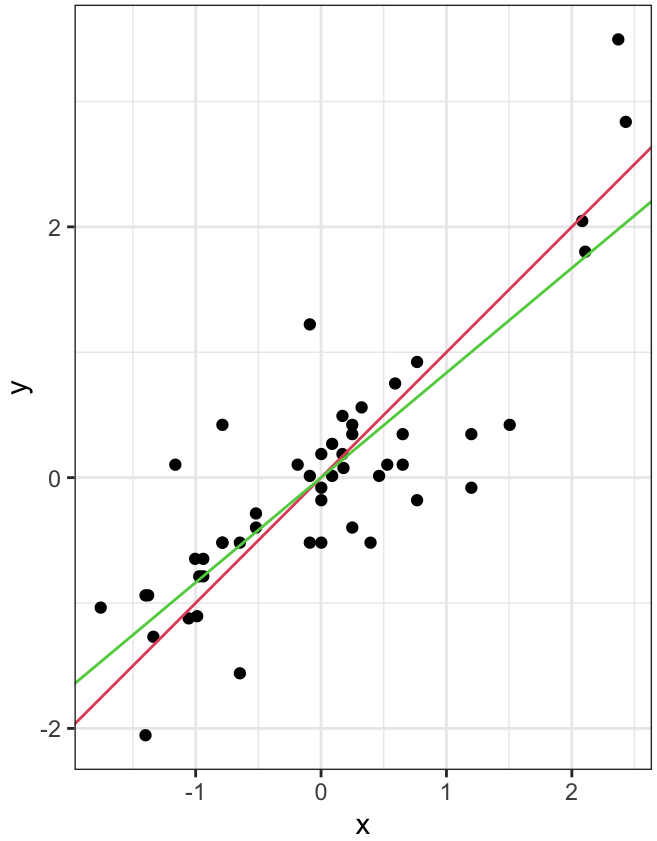


First principal component

- $\text{PC}_1 = a_1 x + a_2 y$
- a_1 and a_2 are the loadings
- PC_1 is the first principal component
- a_1 and a_2 are computed such that PC_1 has maximum variance **AND** $a_1^2 + a_2^2 = 1$



Both on the same plot



Switching to another set of slides…