Lecture 25.1
Dimensionality reduction techniques

Nicola Romanò
nicola.romano@ed.ac.uk

浙江大学爱丁堡大学联合学院
ZJU-UoE Institute

Learning objectives

At the end of this lecture and practical this afternoon you should be able to:

- Explain what dimensionality reduction techniques are and why are they used
- Describe the idea (and some simple process) behind principal component analysis
- Perform, visualise and analyse PCA using R

Measuring gene expression!

We are interested in measuring gene expression in three sets of patients

Control group (**Ctrl**) : no liver pathology

Liver fibrosis group (**Fibr**) : patients with liver fibrosis

Liver cirrhosis group (**Cir**) : patients with liver cirrhosis

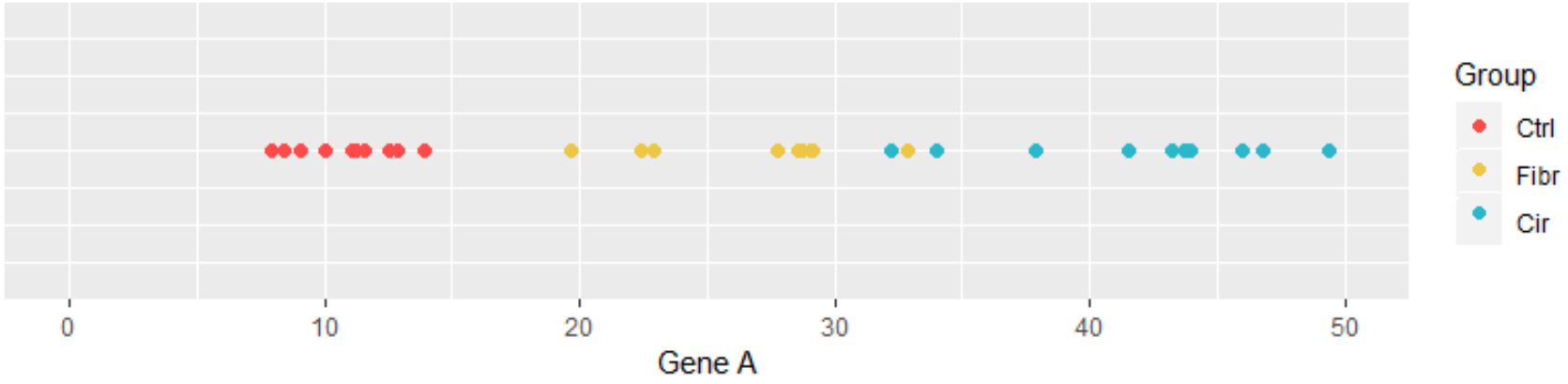
10 patients / group

RNA extracted from liver biopsy, RT-PCR, then expression of target genes assessed by qPCR.

Experiment 1 – 1 gene

Very simple experiment – we only measure gene A

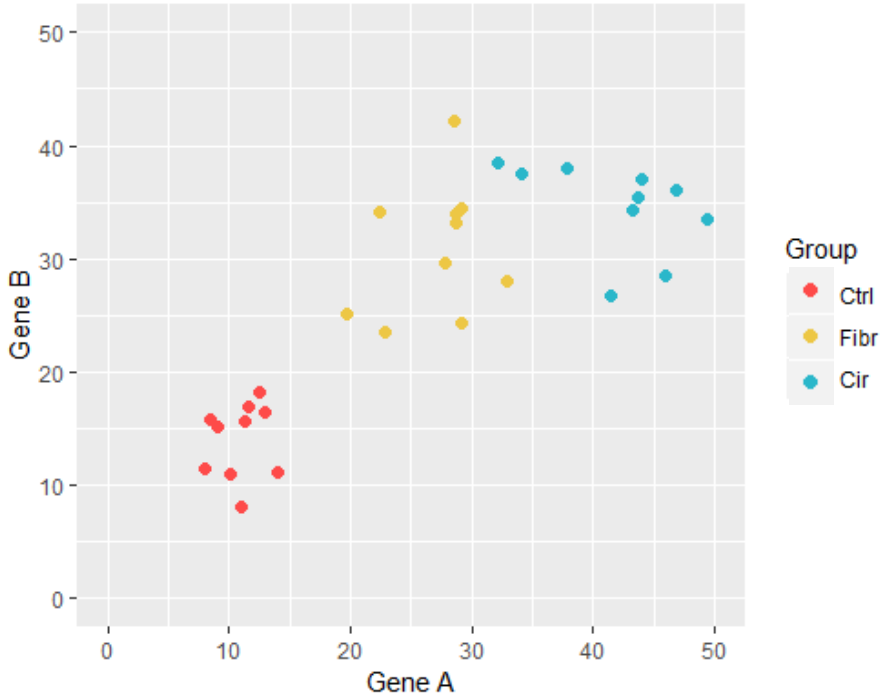
	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	…	Patient 30
A	12.5	13.9	8.4	11.3	12.9	…	22.4



Experiment 2 – 2 genes

Now we measure 2 genes

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	…	Patient 30
A	12.5	13.9	8.4	11.3	12.9	…	22.4
B	18.2	11.1	16.4	15.8	15.6	….	24.3



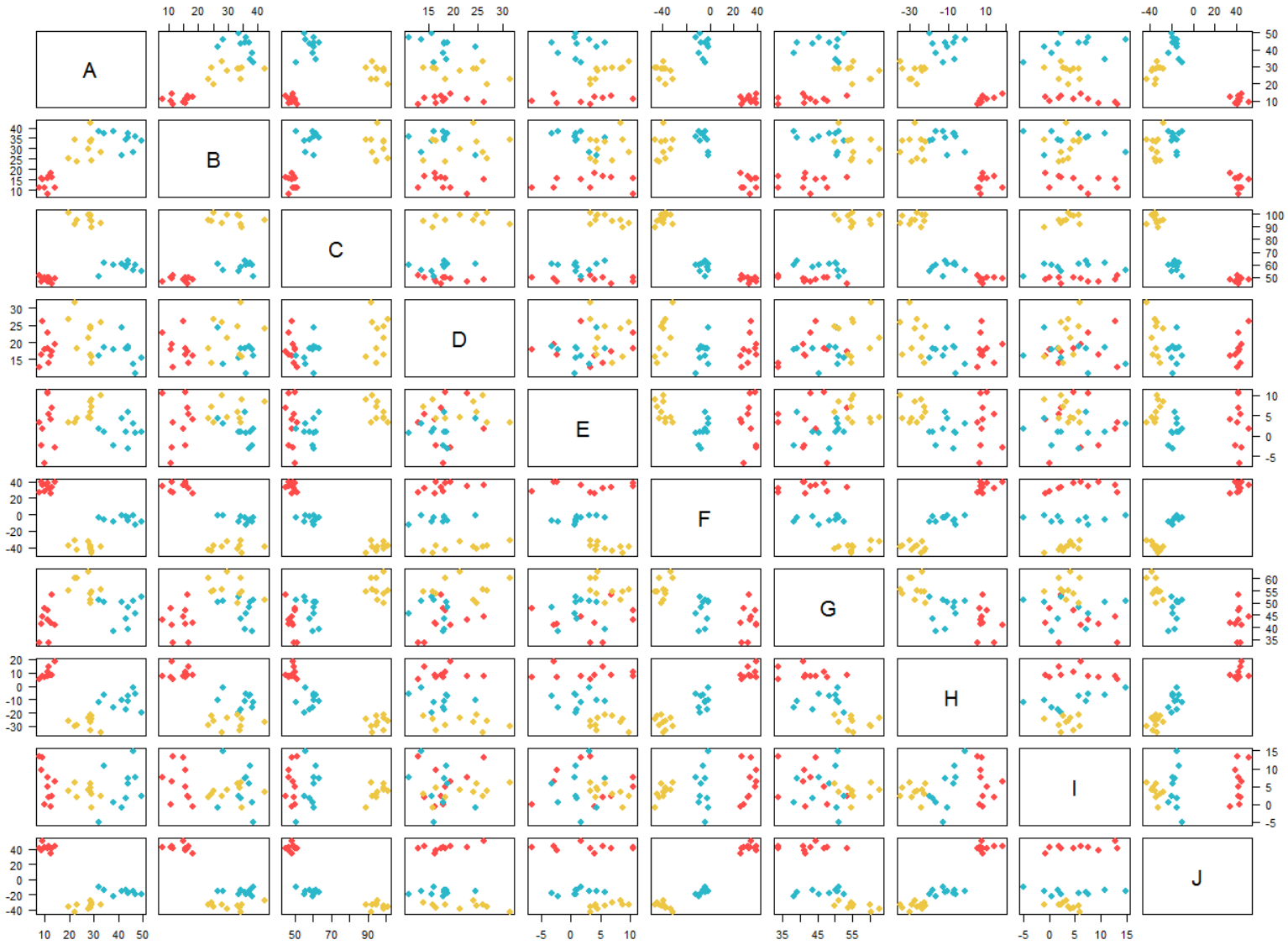
Experiment 3 – 10 genes

Now we measure 10 genes

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	…	Patient 30
A	12.5	13.9	8.4	11.3	12.9	…	22.4
B	18.2	11.1	16.4	15.8	15.6	….	24.3
…	….	….	….	….	….	….	….
J	32.7	28.0	30.9	42.8	31.4		28.5

Would need a 10D plot!!!!

Experiment 3 – 10 genes



Dimensionality reduction techniques

Allow reducing the number of variables under consideration by obtaining a set of principal variables that contain most of the information (variance) of the original ones. The most commonly used is principal component analysis (**PCA**).

That is, starting from:

$$Y = f(V_1, V_2, V_3, \dots, V_{100})$$

PC = principal component; each PC is a linear combination of the original variables

I get to

$$Y = f(PC_1, PC_2, PC_3, \dots, PC_{100})$$

PCs are created so that most of the variance of my original data is contained in the first few ones, so I may as well ignore the rest and use

$$Y = f(PC_1, PC_2)$$

This is much easier to deal with!

Experiment 3 – 10 genes

I start from

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	…	Patient 30
A	12.5	13.9	8.4	11.3	12.9	…	22.4
B	18.2	11.1	16.4	15.8	15.6	….	24.3
…	….	….	….	….	….	….	….
J	32.7	28.0	30.9	42.8	31.4		28.5

I get to

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	…	Patient 30
PC1	12.5	13.9	8.4	11.3	12.9	…	22.4
PC2	18.2	11.1	16.4	15.8	15.6	….	24.3

How to do PCA?

Following is a simplified view of how PCA is calculated.

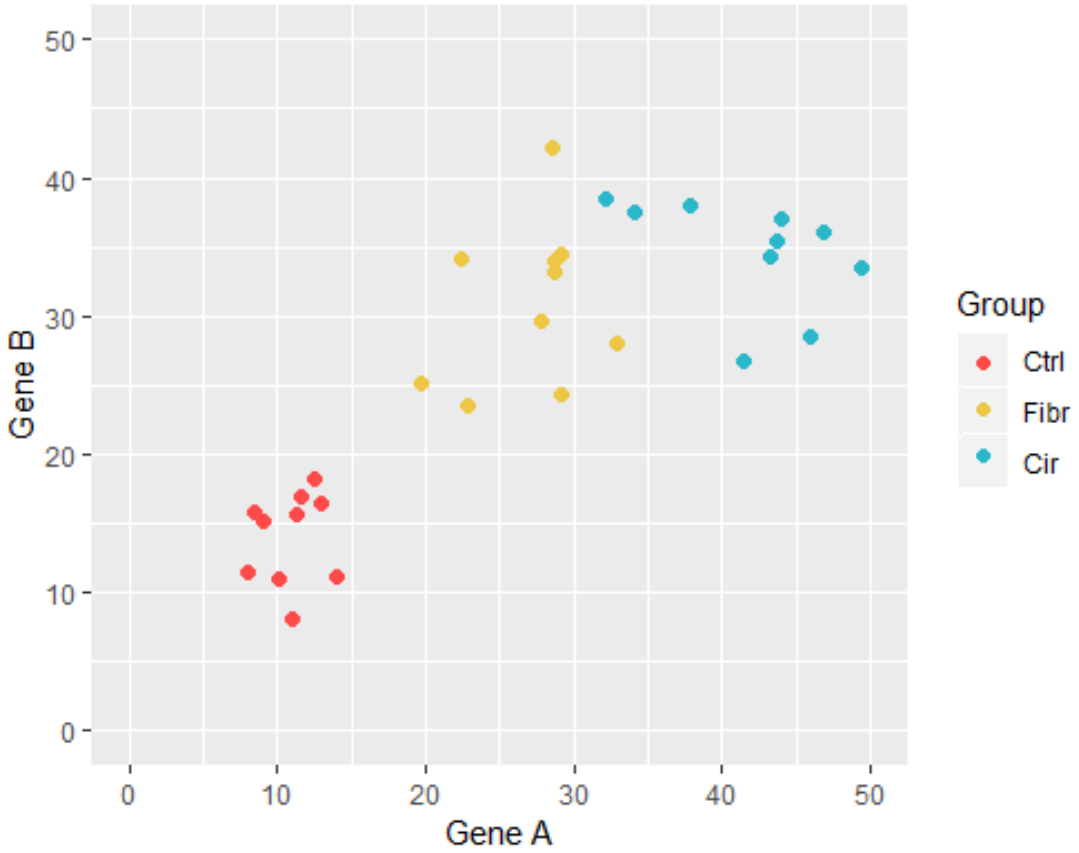
If you really really want to actually learn the maths behind PCA, this is a good start (note, lots of matrix calculus involved!):

A Tutorial on Principal Component Analysis, *Jonathon Shlens, 2005*

The last R workshop will teach you how to do it using R (no fancy math involved!!!)

Back to 2 genes!

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	…	Patient 30
A	12.5	13.9	8.4	11.3	12.9	…	22.4
B	18.2	11.1	16.4	15.8	15.6	….	24.3

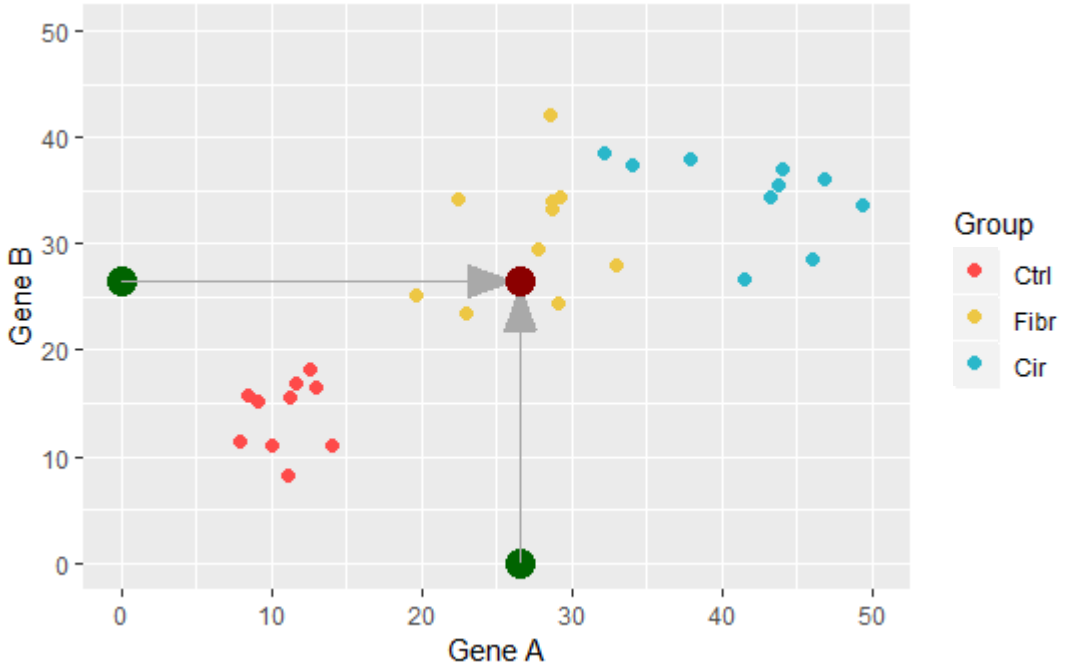


Step 1 – center data

The first thing we want to do is center the data on the origin.

We do this by projecting the data on each dimension (= each axis) and find the mean.

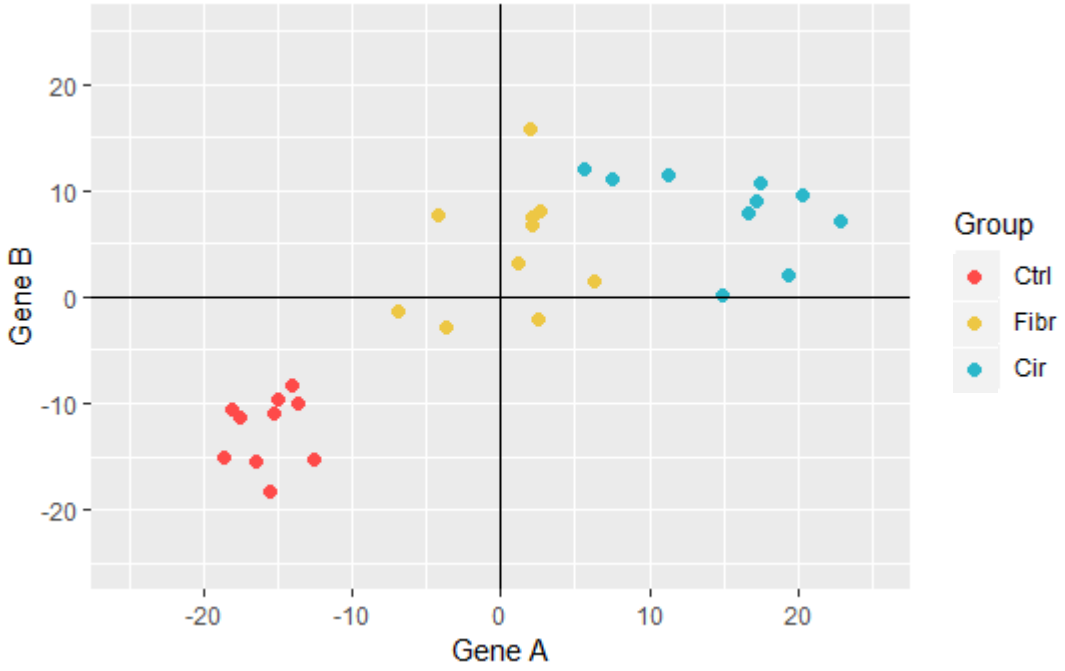
We then subtract the mean from each point



Step 1 – center data

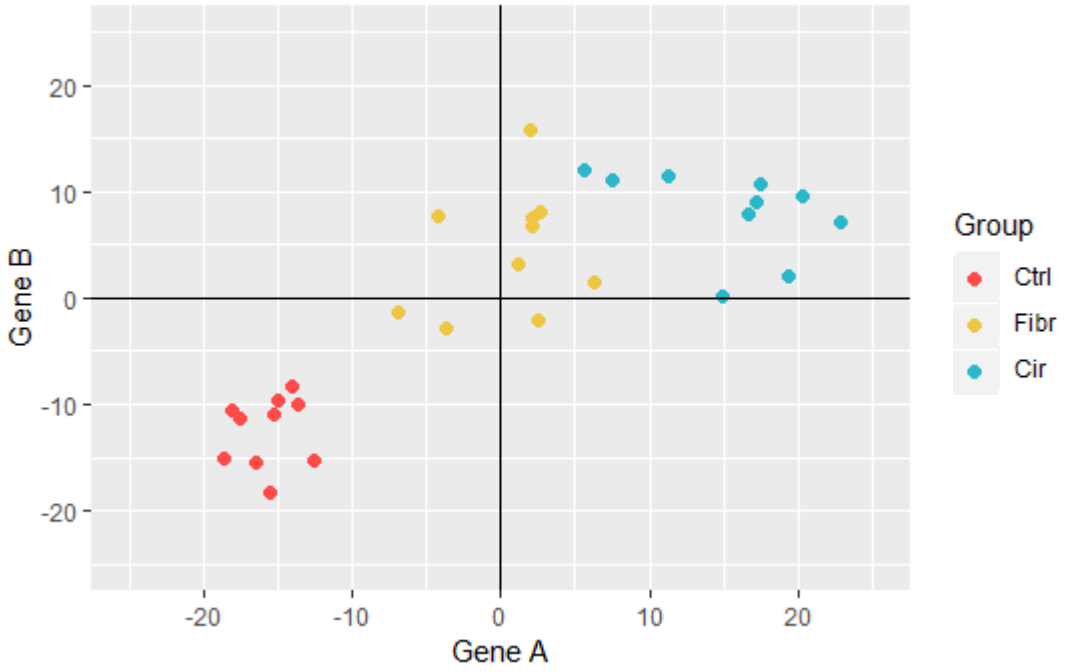
Now our data are centred on (0;0)

Note: if we had more than 2 dimensions, we would shift the data in all of them, and centre them at (0; 0; 0; …; 0)



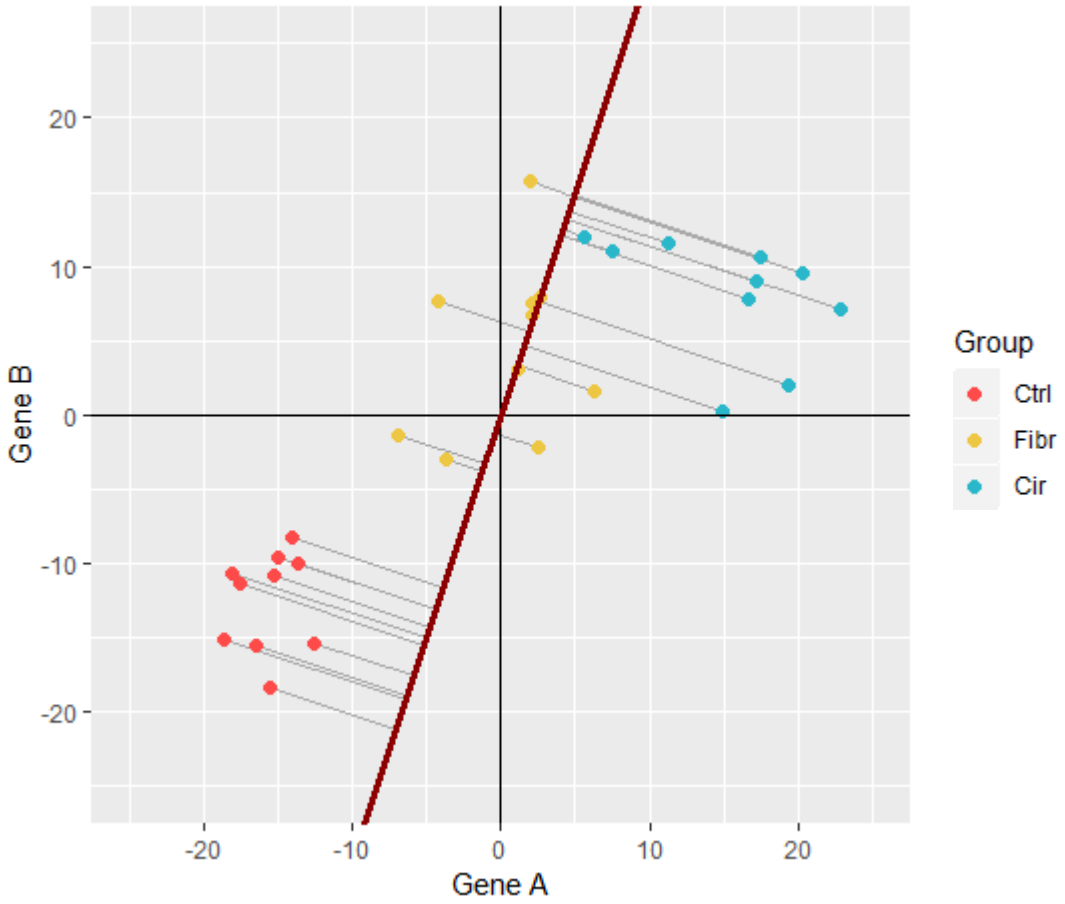
Step 2 – fit a best line

We now want to fit a best fitting line to the data making sure it goes **through the origin**.

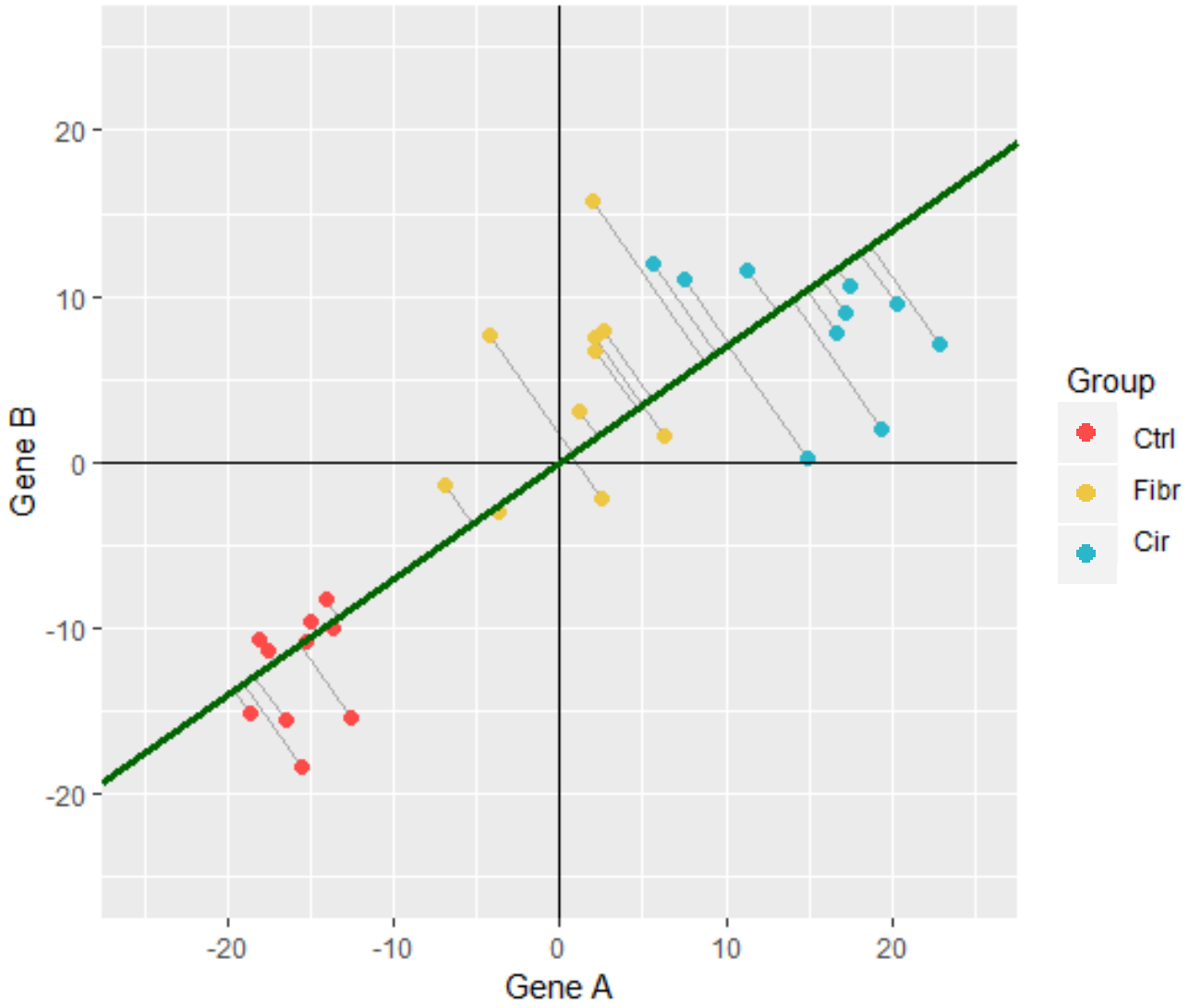


Step 2 – a random line…

PCA tries to **minimise** the distance from each point to the line
This is equal to **maximise** the distance between the projection and the origin.
Note that PCA minimises (or maximises) the sum of squares of the distances

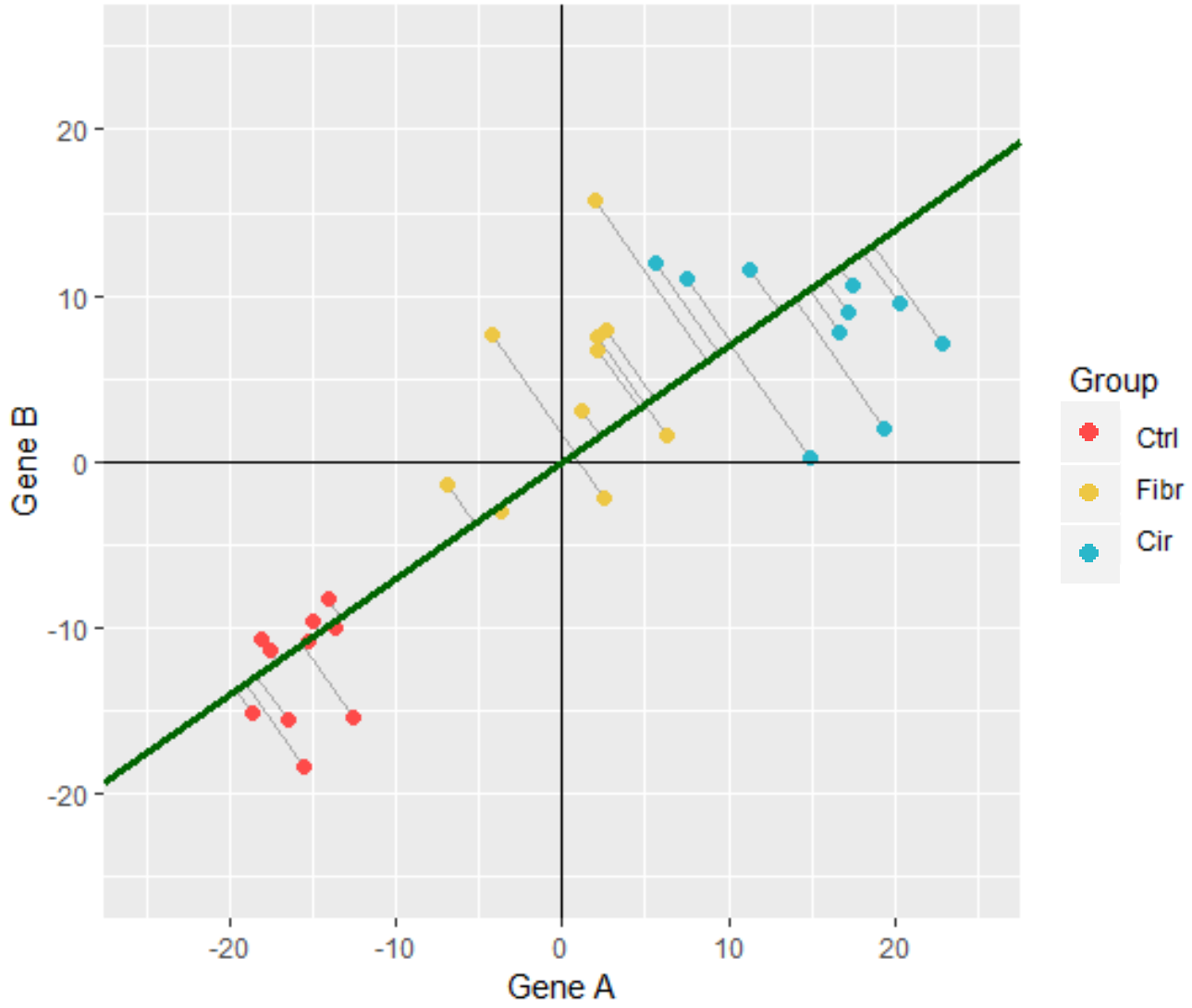


Step 2 – find the best line



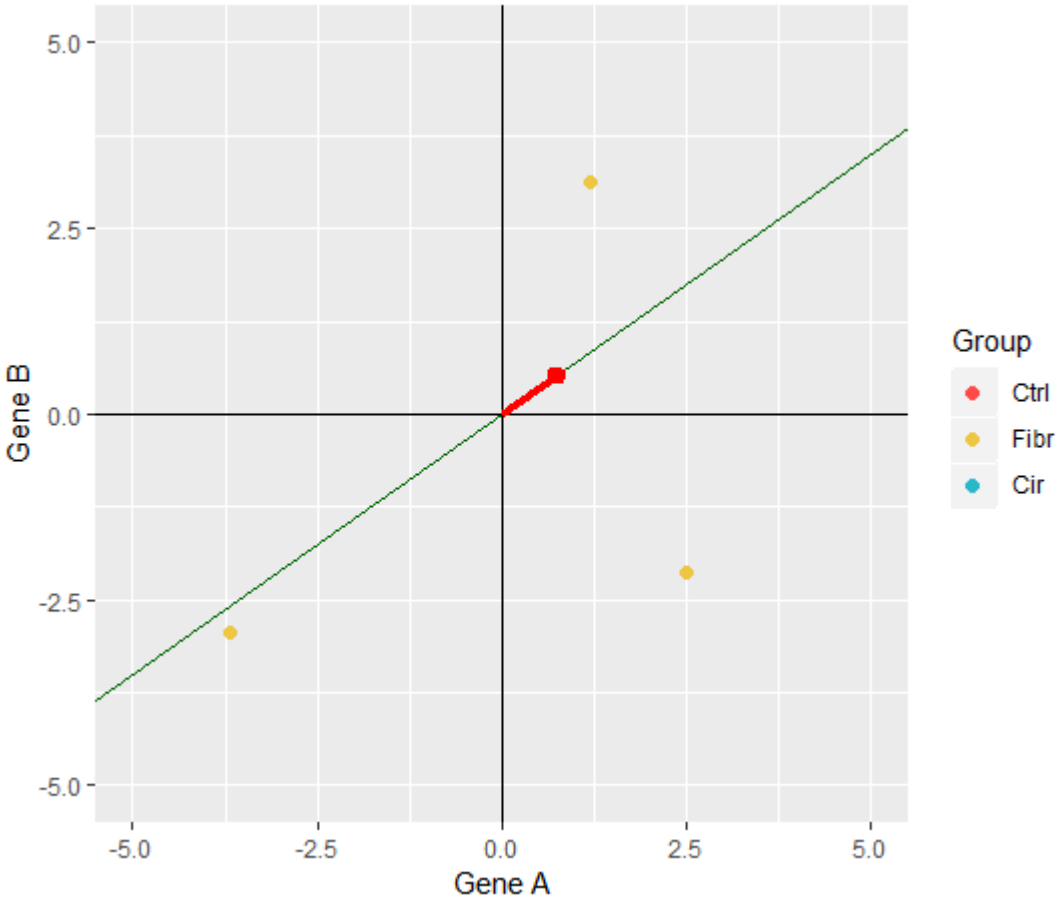
You got PC1 !

PC1 corresponds (in this example) to the line: $y = 0.7\,x$



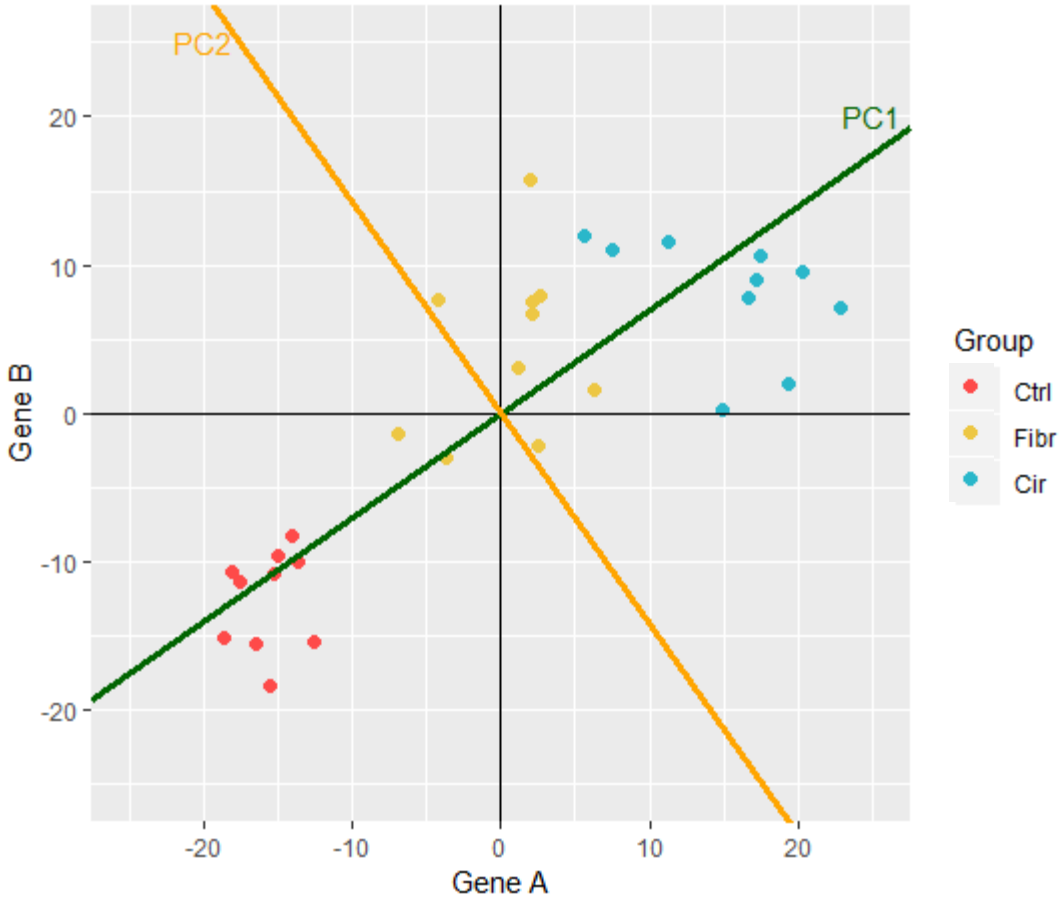
Some PCA terminology

The unit vector (= vector of length 1) on PC1 is called the PC1 **eigenvector.**
In our example it is (0.818; 0.574) <- these are called the PC1 **loading scores.**
The SS of the distances of the projection to the origin is called the PC1 **eigenvalue.**



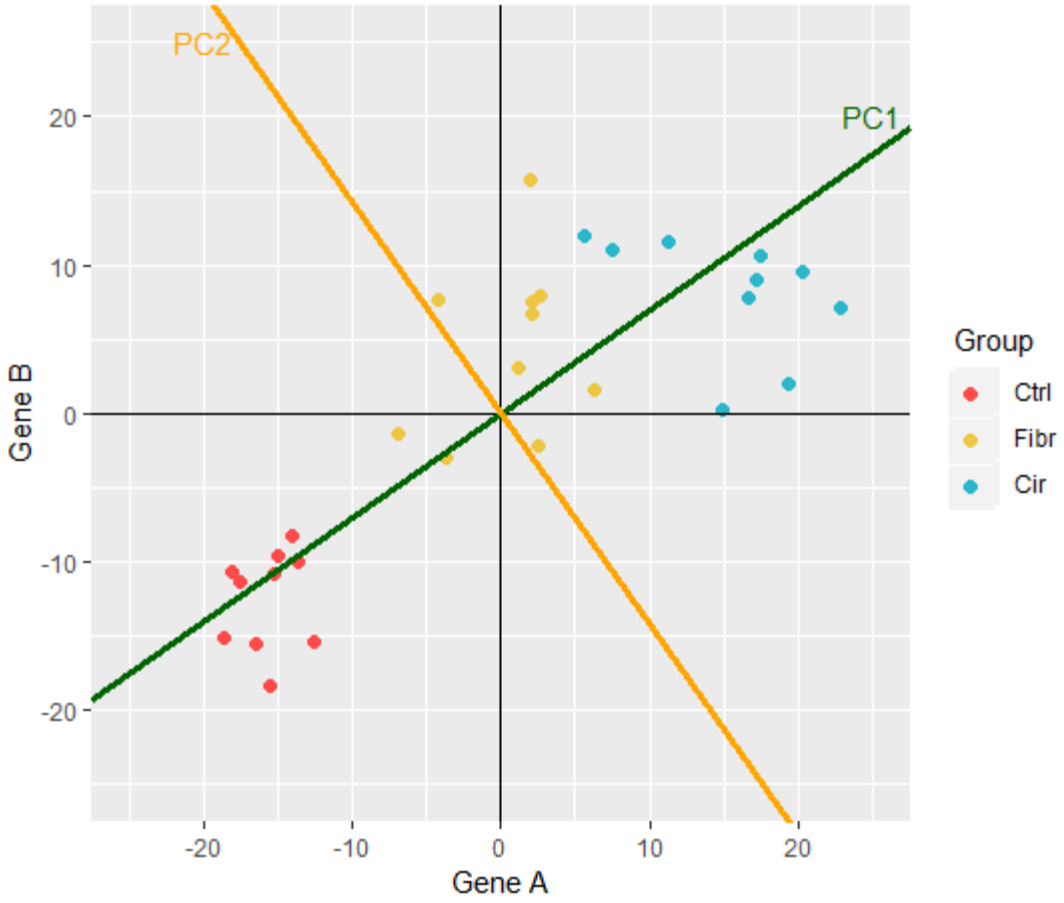
Let's find PC2!

PC1 is the line perpendicular to PC1, passing through the origin and that minimised the SS distances of the points. In our simple 2D example there is only one line to test… but with more variable there would be more!



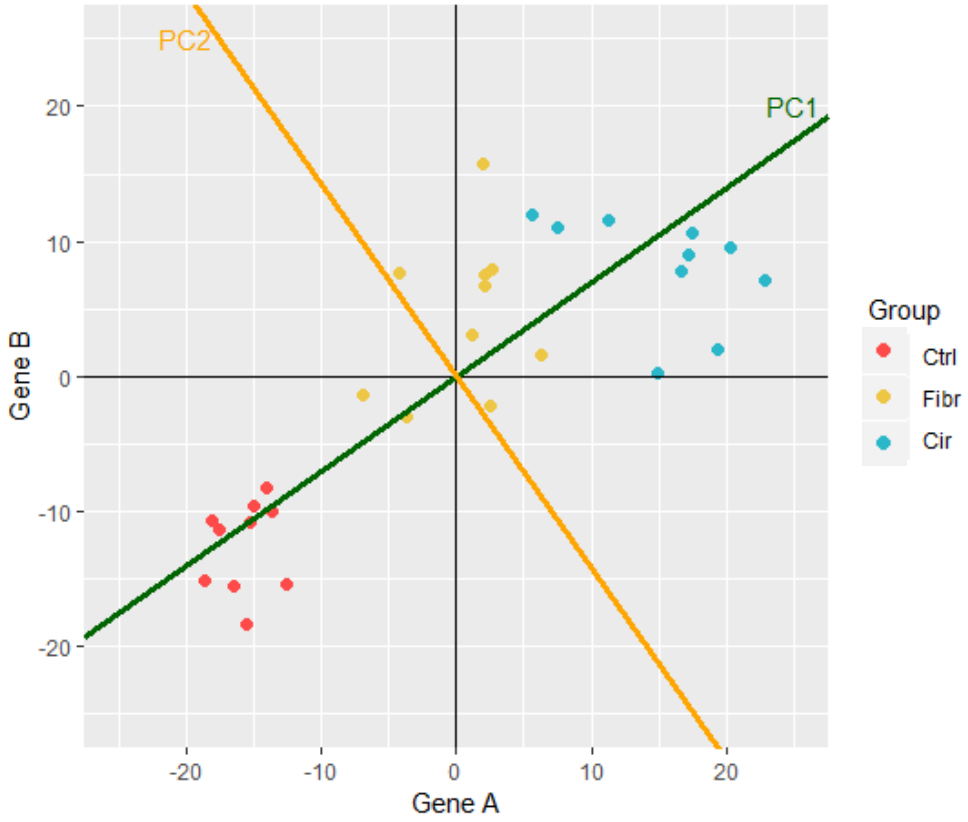
Let's find PC2!

The eigenvector for PC2 is (0.57; -0.81)
This means Gene B is ~1.4 times more important in determining PC2 then Gene A

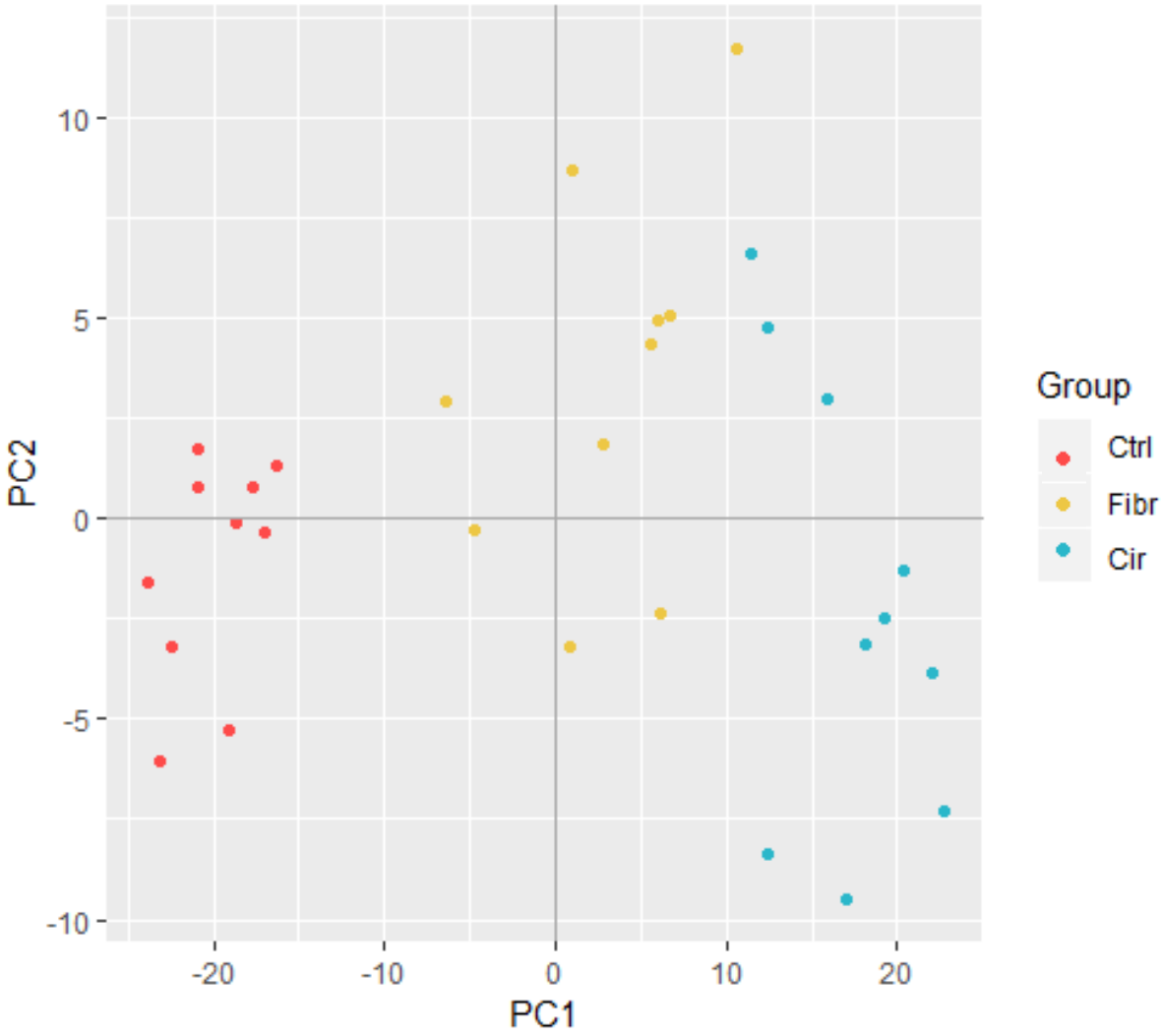


And finally, some rotation

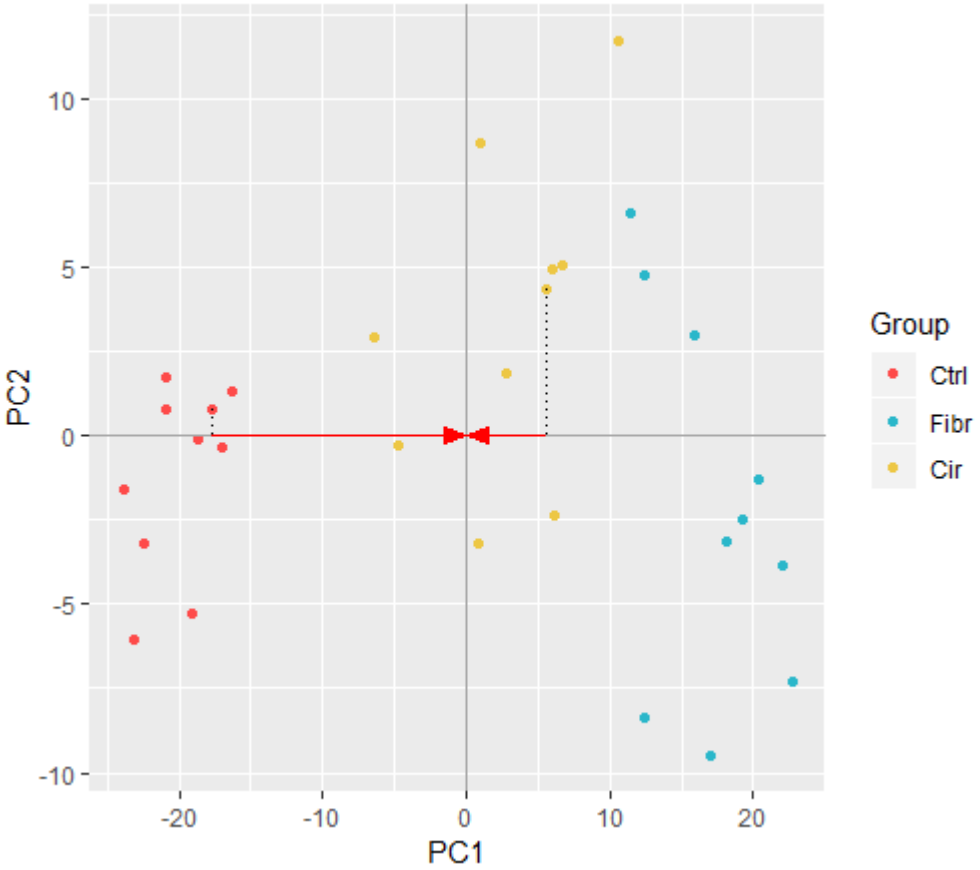
We now rotate our coordinate system, until PC1 is horizontal



Our PCA is done!



Variance explained



To find PC1 we maximised the SS of the distances of the projection of data on PC1 to the origin.

Remember the definition of variance

$$s = \frac{\sum(x - \bar{x})^2}{n - 1}$$

The mean of our data is 0, so if we divide the PC1 SS by n-1 we get the variance contained in PC1.

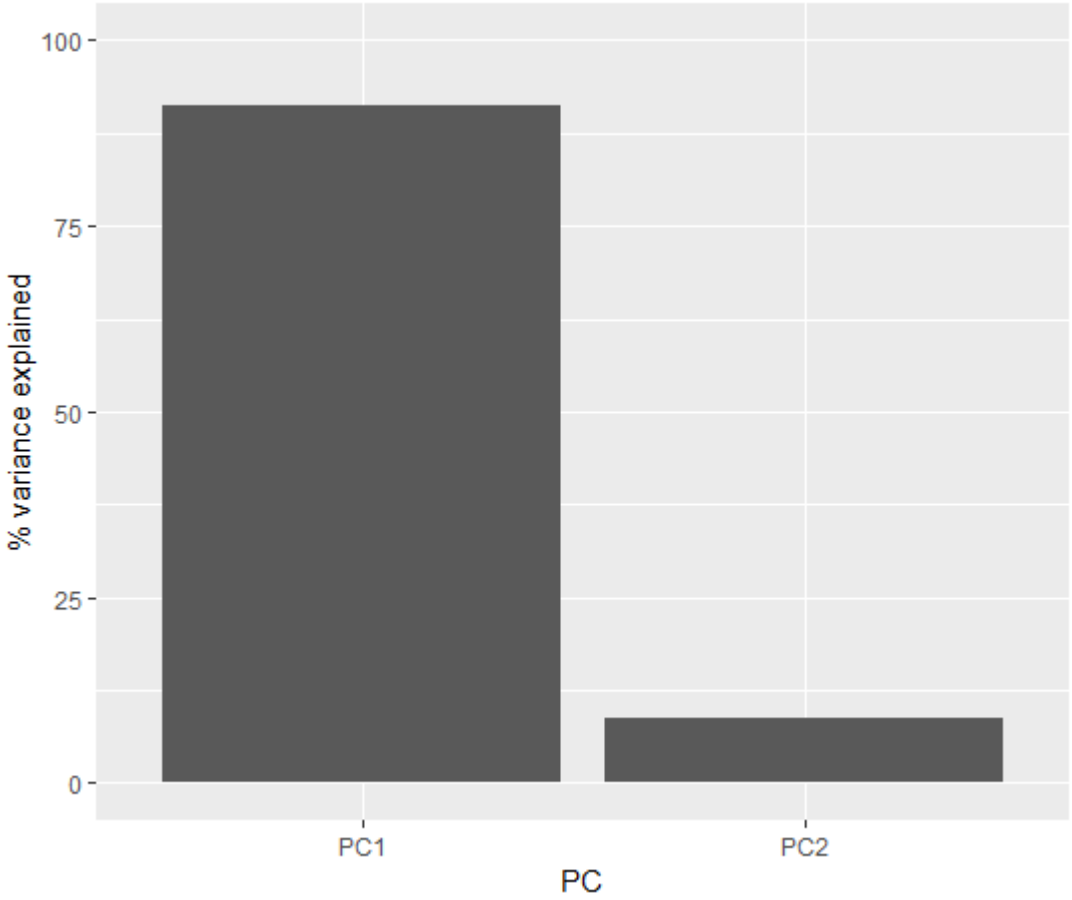
In our example:

var_{PC1} = 8.9239 var_{PC2} = 0.8477

So, PC1 explains 8.9/(8.9+0.8) = 91.3% of variance in the data and PC2 8.7%.

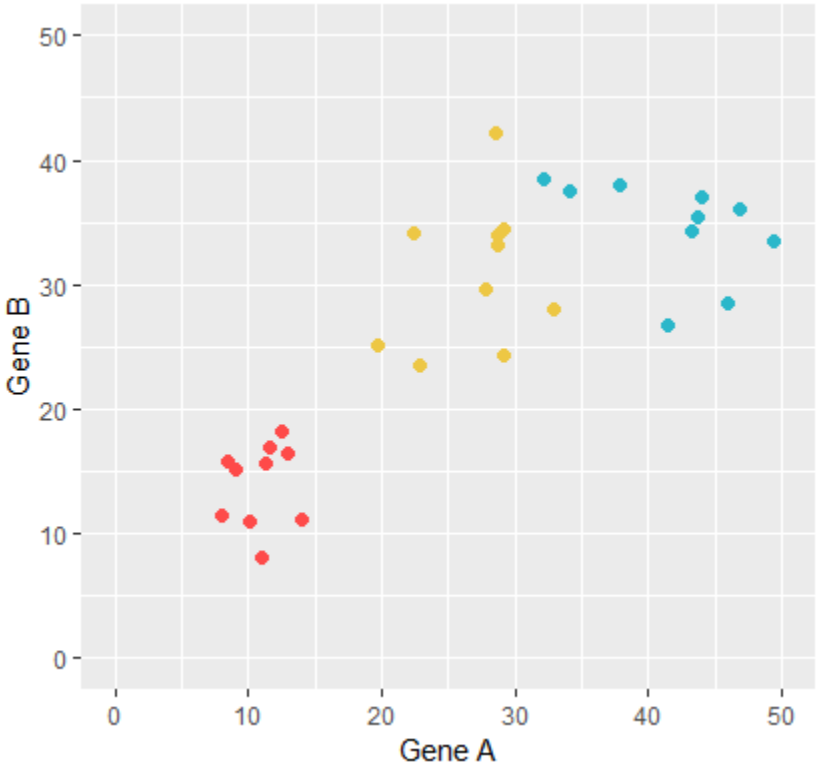
Scree plots

A common way to visualise the contribute of each component is through a scree plot, a bar plot of the variance explained.

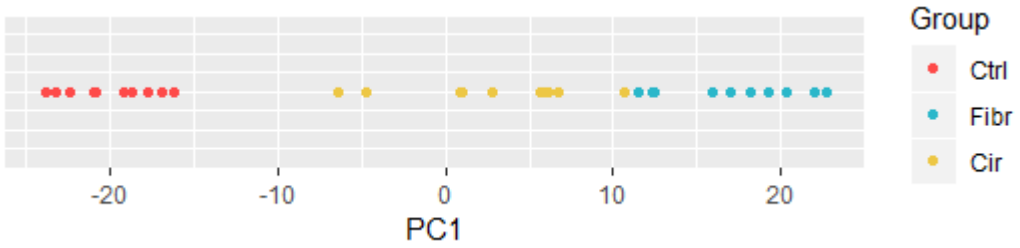


Dimension reduction

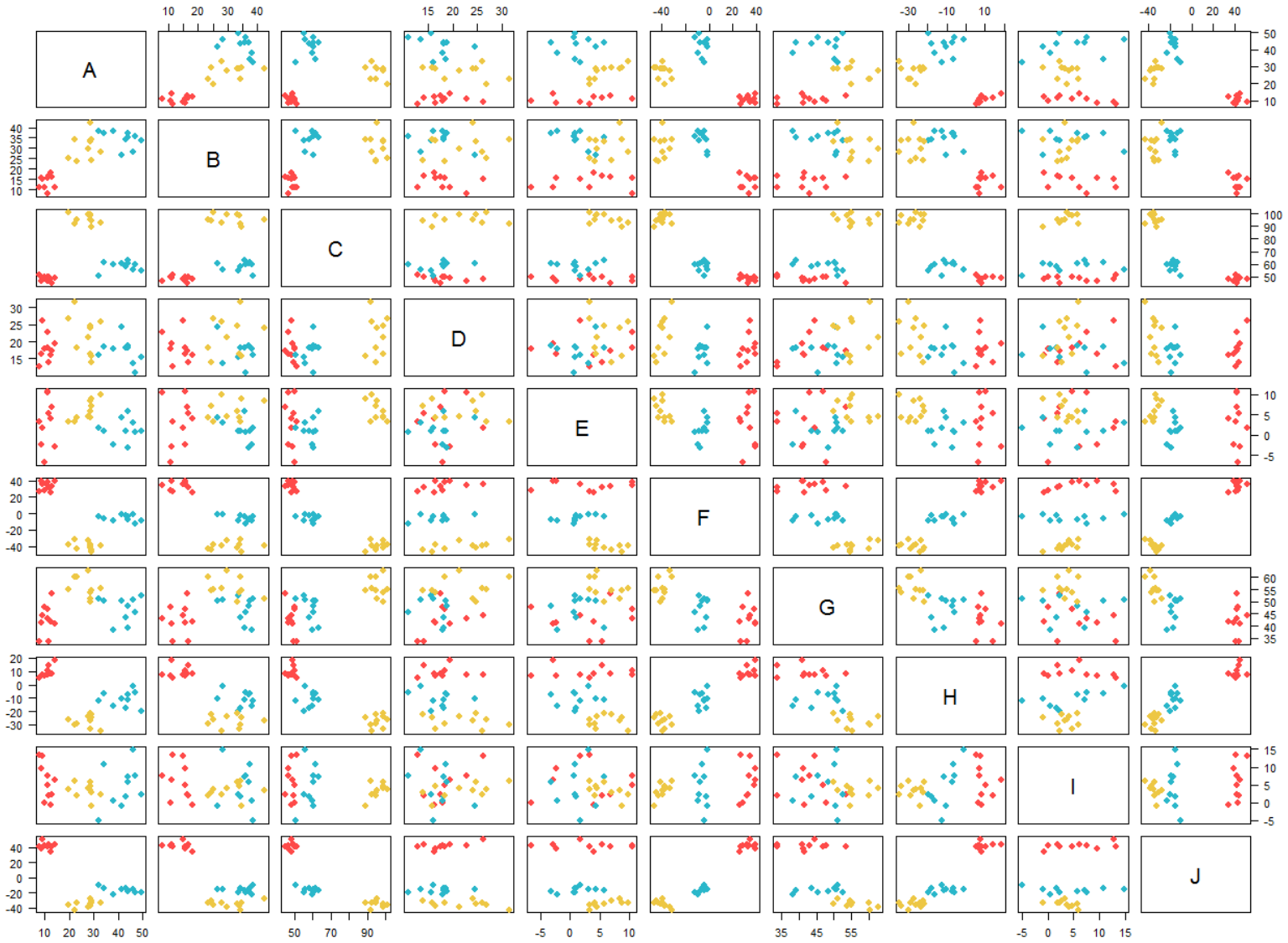
From 2D…



…to 1D!

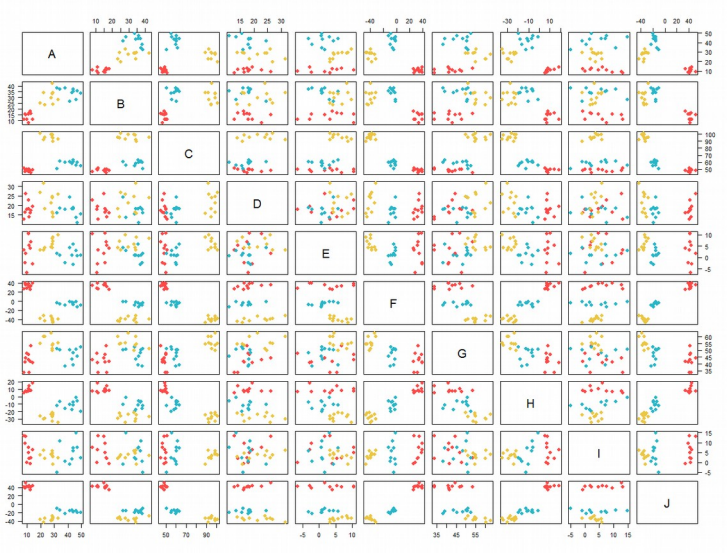


A larger dataset!

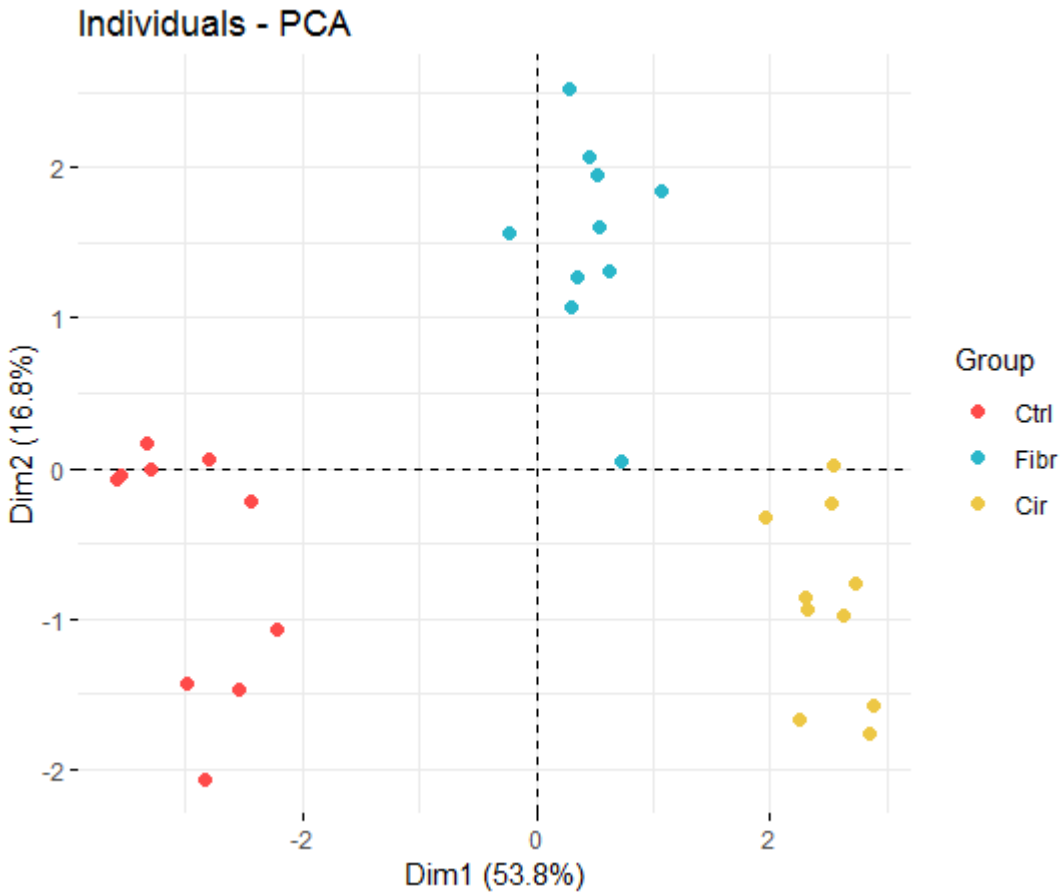


PCA

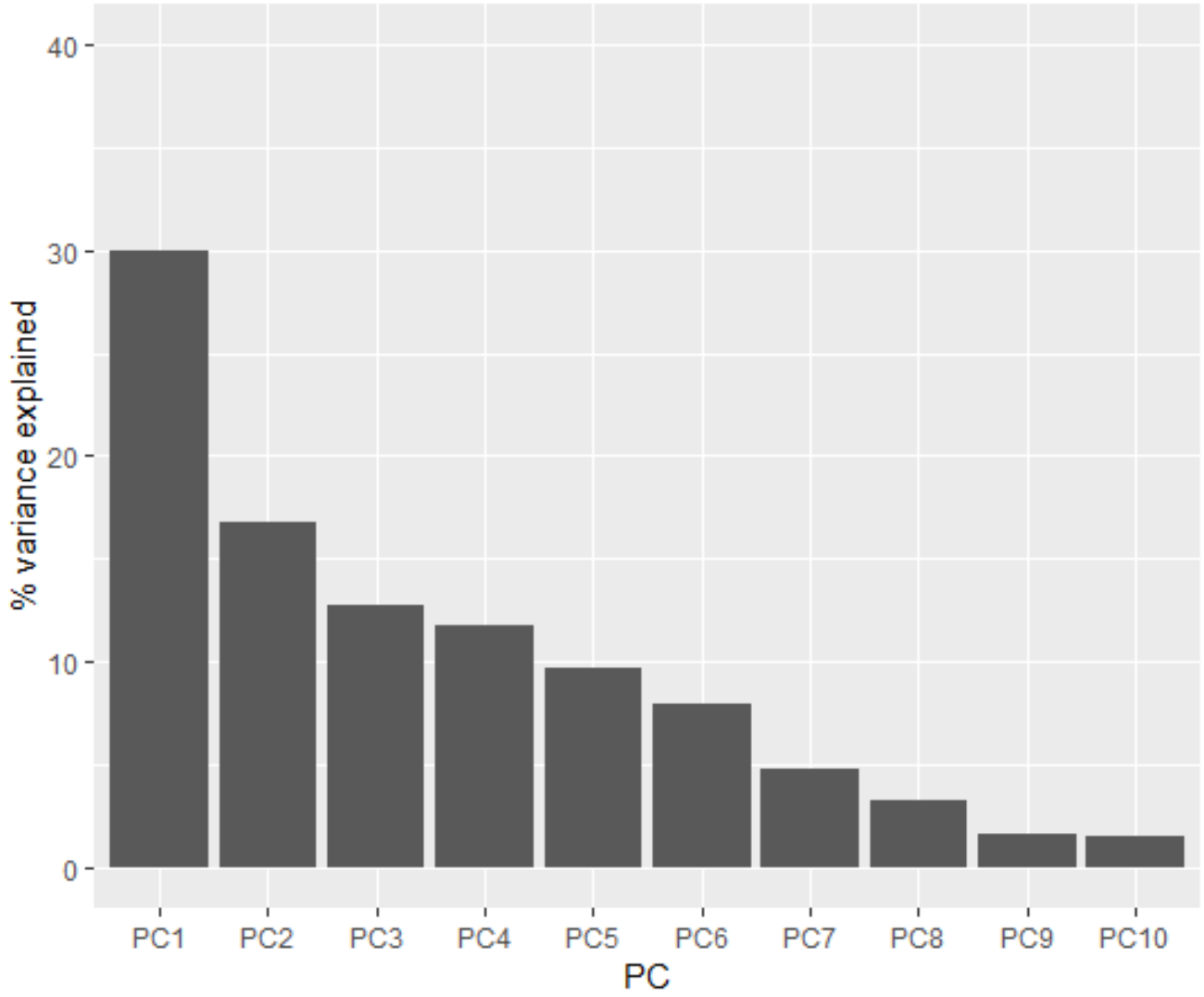
From 10D…



…to 2D!



Scree plot


40
30
20
10
0
% variance explained
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
PC

Summary

- PCA is one of the dimension reduction techniques, that allow reducing the number of variables in a dataset
- Useful for simplify dataset with many variables
- Useful to find groups of observations that differ because of complex relations between variables
- It relies on rotating the data so that we express them relative to the direction in which they have biggest variance