Multivariate Statistical Analysis

Eliana Ibrahimi

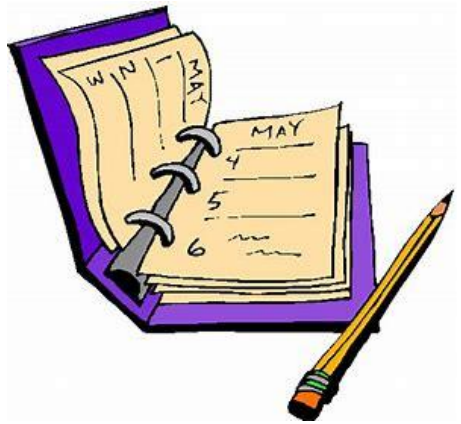
Department of Biology, University of Tirana, Albania

MSB 2023 Training School

October 18-20, 2023, Tirana, Albania

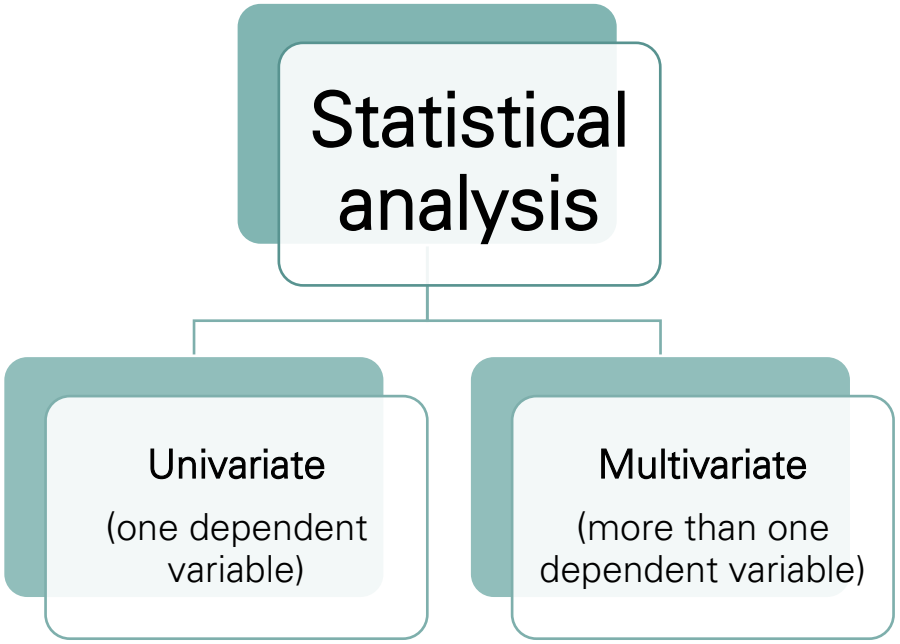
Lecture overview

- Why multivariate analysis?
- Multivariate techniques
    - Multivariate Analysis of Variance (MANOVA)
    - Principal Component Analysis (PCA)



Multivariate analysis

- Statistically speaking, multivariate analysis has to include at least two dependent variables to analyze differences or relationships with the independent variable(s).



Example dataset

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa

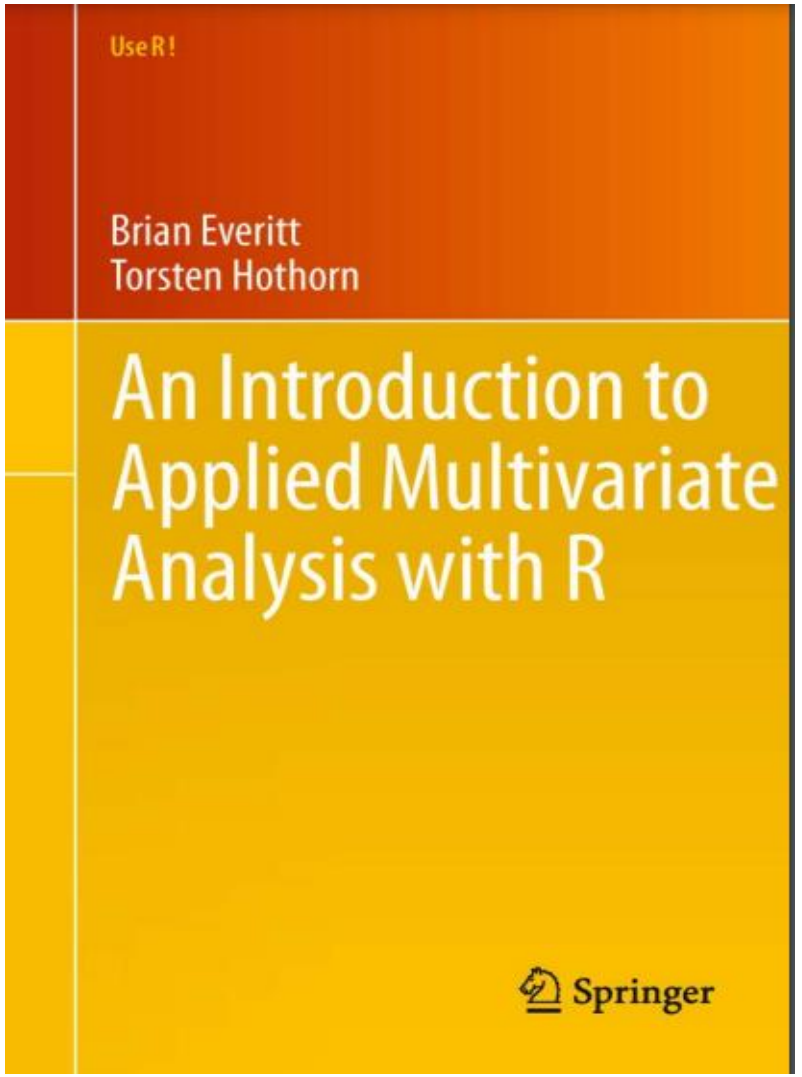
Iris dataset:

1 - Could petal and sepal measurements be summarized in some way by combining the four measurements into a single number?

2- Are there subtypes of measurements amongst the Iris species within which individuals differ?

Multivariate analysis techniques

- Multivariate analysis of variance (MANOVA)
- Principal components analysis (PCA)
- Discriminant analysis and Linear Discriminant analysis
- Factor analysis
- Canonical correlation analysis
- Canonical (or "constrained") correspondence analysis (CCA)
- Multidimensional scaling



One-way MANOVA

MANOVA is a generalization of ANOVA applied with more than one dependent variable. The null and the alternative hypotheses are:

H_0: Group **mean vectors** are the same for all groups or they don't differ significantly.

$$H_0 = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{bmatrix} = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{bmatrix} = \cdots = \begin{bmatrix} \mu_{k1} \\ \mu_{k2} \\ \vdots \\ \mu_{kp} \end{bmatrix}$$

H_1: At least one of the group **mean vectors** is different from the rest.

Test statistics:

1. Pillai's trace
2. Hotelling-Lawley's trace
3. Wilk's lambda

$$\text{Pillai's trace} = \text{trace}[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}] = \sum_{i=1}^{q} \frac{\lambda_i}{1+\lambda_i}.$$

Assumptions of MANOVA

- The dependent variables should be normally distributed within groups. MANOVA, assumes **multivariate normality**.
- Homogeneity of covariances and variances across the range of predictors.
- Linearity between all pairs of dependent variables.

Interpretation of MANOVA

If the global multivariate test is significant, we conclude that the corresponding effect of the independent variable is significant.

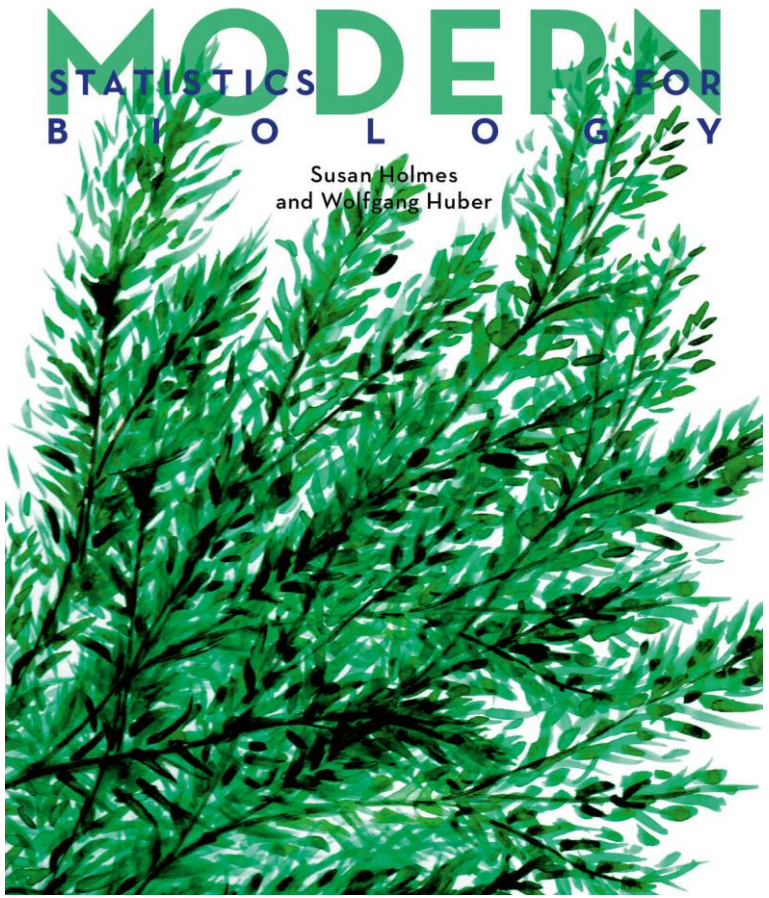
In that case, we want to identify the specific dependent variables that contributed to the significant global effect.

Compute MANOVA in R

Go to ‘Multivariate analysis_MANOVA’ R notebook

Principal components analysis - PCA

Mainly based on …



https://web.stanford.edu/class/bios221/book/00-chap.html

Data, matrices and motivation

Cell Types: Holmes et al. (2005) studied gene expression profiles of sorted T-cell populations from different subjects.

	X3968	X14831	X13492	X5108	X16348	X585
HEA26_EFFE_1	-2.61	-1.19	-0.06	-0.15	0.52	-0.02
HEA26_MEM_1	-2.26	-0.47	0.28	0.54	-0.37	0.11
HEA26_NAI_1	-0.27	0.82	0.81	0.72	-0.90	0.75
MEL36_EFFE_1	-2.24	-1.08	-0.24	-0.18	0.64	0.01
MEL36_MEM_1	-2.68	-0.15	0.25	0.95	-0.20	0.17

Bacterial Species Abundances: Matrices of counts are used in microbial ecology studies

OTU Table: [4 taxa and 8 samples]
taxa are rows

	246140	143239	244960	255340	144887	141782	215972	31759
CL3	0	7	0	153	3	9	0	0
CC1	0	1	0	194	5	35	3	1
SV1	0	0	0	0	0	0	0	0
M31Fcsw	0	0	0	0	0	0	0	0

mRNA reads: RNA-Seq transcriptome data report the number of sequence reads matching each gene[2] in each of several biological samples.

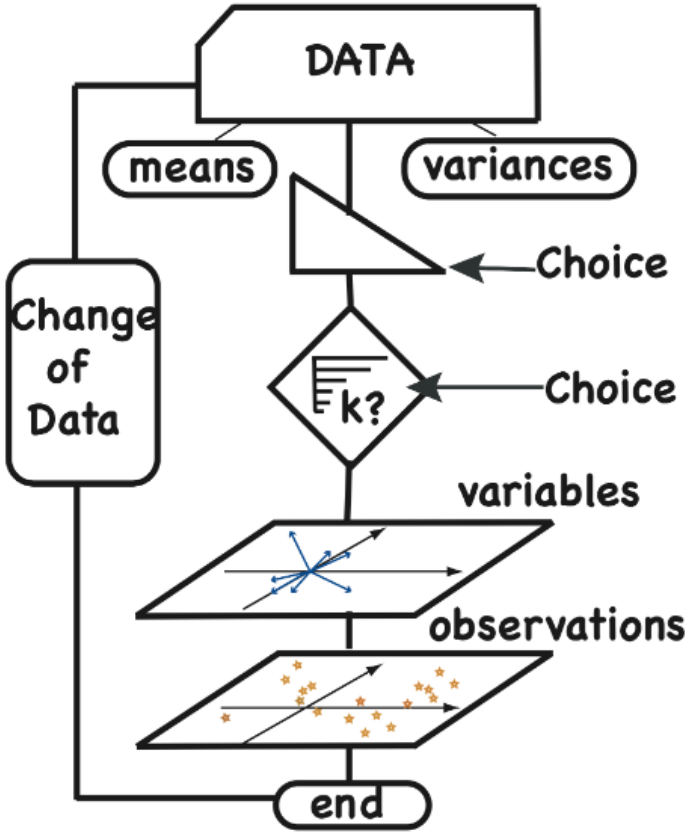
	SRR1039508	SRR1039509	SRR1039512	SRR1039513
ENSG00000000003	679	448	873	408
ENSG00000000005	0	0	0	0
ENSG00000000419	467	515	621	365

Principal components analysis - PCA

- Build new variables, called principal components (PC), that are more useful than the original measurements.

- Visualize what this decomposition achieves and learn how to choose the number of principal components.

- Project factor covariates onto the PCA map to enable a more useful interpretation of the results.



Source: Modern Statistics for Modern Biology, Holmes & Huber, (2019)

Preprocessing the data

Different variables are measured in different units…

For PCA and many other methods, we need to transform the numeric values to make them comparable.

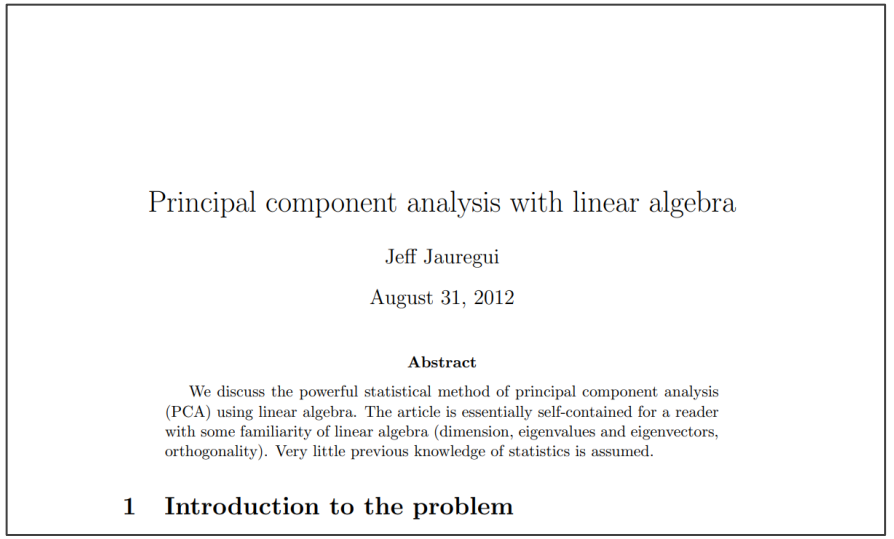
- Centering (subtracting the mean).
- Scaling or standardizing (dividing by the standard deviation)

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
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
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	-0.89767388	1.01560199	-1.33575163	-1.3110521482
2	-1.13920048	-0.13153881	-1.33575163	-1.3110521482
3	-1.38072709	0.32731751	-1.39239929	-1.3110521482
4	-1.50149039	0.09788935	-1.27910398	-1.3110521482
5	-1.01843718	1.24503015	-1.33575163	-1.3110521482
6	-0.53538397	1.93331463	-1.16580868	-1.0486667950
7	-1.50149039	0.78617383	-1.33575163	-1.1798594716
8	-1.01843718	0.78617383	-1.27910398	-1.3110521482
9	-1.74301699	-0.36096697	-1.33575163	-1.3110521482
10	-1.13920048	0.09788935	-1.27910398	-1.4422448248
11	-0.53538397	1.47445831	-1.27910398	-1.3110521482

Dimensionality reduction

- PCA is called an **unsupervised learning** technique because, as in clustering, it treats all variables as having the same **status**.
- PCA does not predict or explain one particular variable's value from the others but gives a mathematical model for an underlying structure for all the variables.
- Mathematical formulation of these methods is done through linear algebra …



Principal component analysis with linear algebra

Jeff Jauregui

August 31, 2012

Abstract

We discuss the powerful statistical method of principal component analysis (PCA) using linear algebra. The article is essentially self-contained for a reader with some familiarity of linear algebra (dimension, eigenvalues and eigenvectors, orthogonality). Very little previous knowledge of statistics is assumed.

1 Introduction to the problem

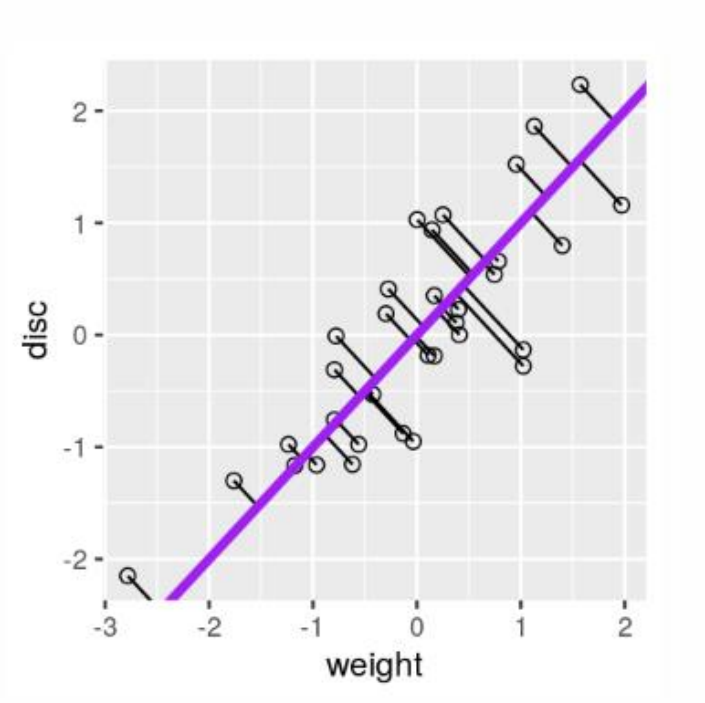
https://www.math.union.edu/~jaureguj/PCA.pdf

Linear combination

PCA is a **linear** technique, which uses new variables that are linear functions of the original ones.

Principal components are linear combinations of the variables that were originally measured, they provide a new coordinate system.

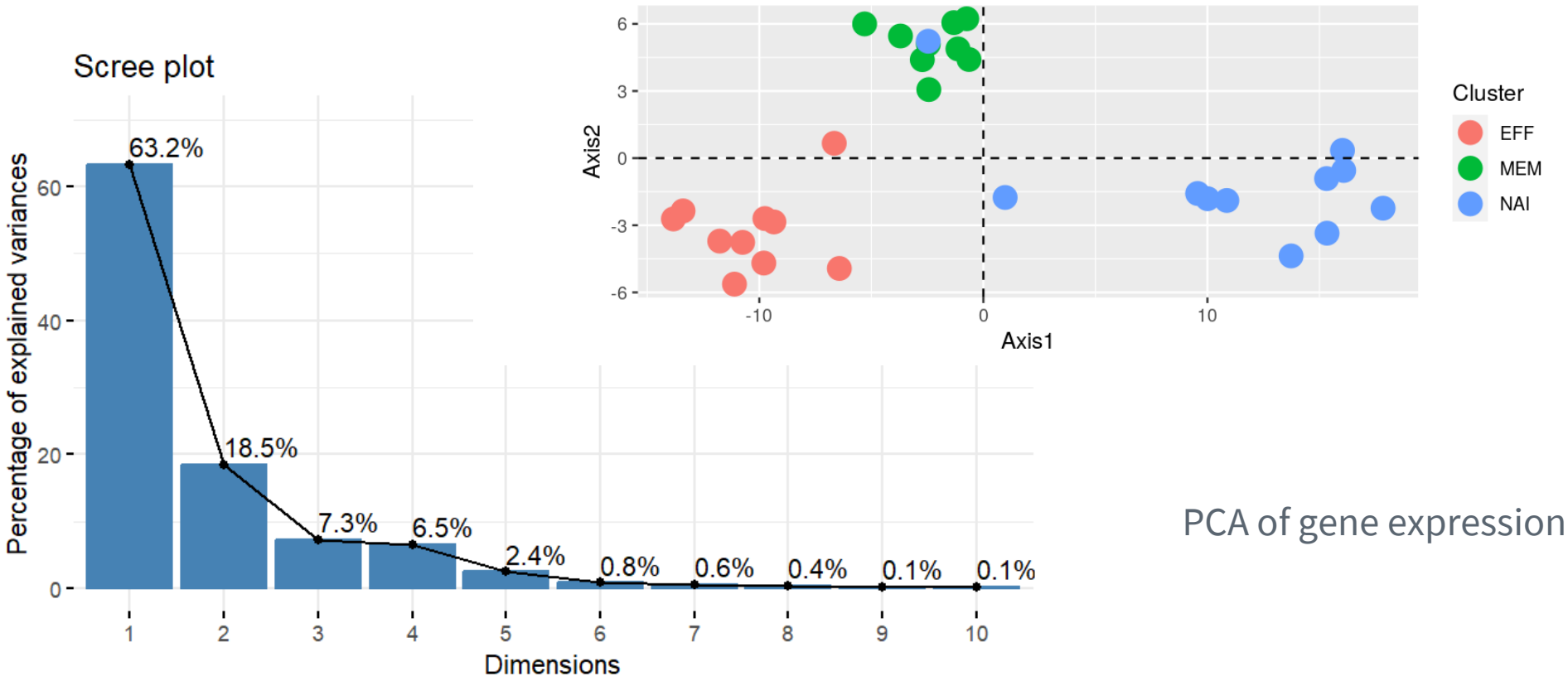
$$PC = \frac{1}{2}\text{disc} + \frac{1}{2}\text{weight}.$$



The purple **principal component** line minimizes the sums of squares of the orthogonal projections.

Source: Modern Statistics for Modern Biology, Holmes & Huber, (2019)

How to choose the number of dimensions



PCA of gene expression

Compute PCA in R

Let’s start the practical...

Go to the ‘Multivariate Analysis_PCA’ R notebook

References

Modern Statistics for Modern Biology, by Susan Holmes and Wolfgang Huber, Cambridge University Press (2019)

Tabachnick, Barbara, and Linda.S. Fidell. 2012. Using Multivarite Statistics. 6th ed. Pearson.

https://www.statology.org/univariate-vs-multivariate-analysis/

https://www.r-bloggers.com/2022/01/manova-in-r-how-to-implement-and-interpret-one-way-manova/