PCA Revealed

Part 2: Introduction

Gaston **S**anchez

August 2014

Readme

License:

You are free to:

Share — copy and redistribute the material

Adapt — rebuild and transform the material

Under the following conditions:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.

NonCommercial — You may not use this work for commercial purposes.

Share Alike — If you remix, transform, or build upon this work, you must distribute your contributions under the same license to this one.

Motivation

Cereals Data Set

```
cereals
```

	Cups	Calories	Carbs	Fat	Fiber	Potassium	Protein	Sodium	Sugars
CapnCrunch	0.75	120	12.0	2	0.0	35	1	220	12
CocoaPuffs	1.00	110	12.0	1	0.0	55	1	180	13
Trix	1.00	110	13.0	1	0.0	25	1	140	12
AppleJacks	1.00	110	11.0	0	1.0	30	2	125	14
CornChex	1.00	110	22.0	0	0.0	25	2	280	3
CornFlakes	1.00	100	21.0	0	1.0	35	2	290	2
Nut&Honey	0.67	120	15.0	1	0.0	40	2	190	9
Smacks	0.75	110	9.0	1	1.0	40	2	70	15
MultiGrain	1.00	100	15.0	1	2.0	90	2	220	6
CracklinOat	0.50	110	10.0	3	4.0	160	3	140	7
GrapeNuts	0.25	110	17.0	0	3.0	90	3	179	3
HoneyNutCheerios	0.75	110	11.5	1	1.5	90	3	250	10
NutriGrain	0.67	140	21.0	2	3.0	130	3	220	7
Product19	1.00	100	20.0	0	1.0	45	3	320	3
TotalRaisinBran	1.00	140	15.0	1	4.0	230	3	190	14
WheatChex	0.67	100	17.0	1	3.0	115	3	230	3
Oatmeal	0.50	130	13.5	2	1.5	120	3	170	10
Life	0.67	100	12.0	2	2.0	95	4	150	6
Maypo	1.00	100	16.0	1	0.0	95	4	0	3
QuakerOats	0.50	100	14.0	1	2.0	110	4	135	6
Muesli	1.00	150	16.0	3	3.0	170	4	150	11
Cheerios	1.25	110	17.0	2	2.0	105	6	290	1
SpecialK	1.00	110	16.0	0	1.0	55	6	230	3

By looking at the data, can you spot ...

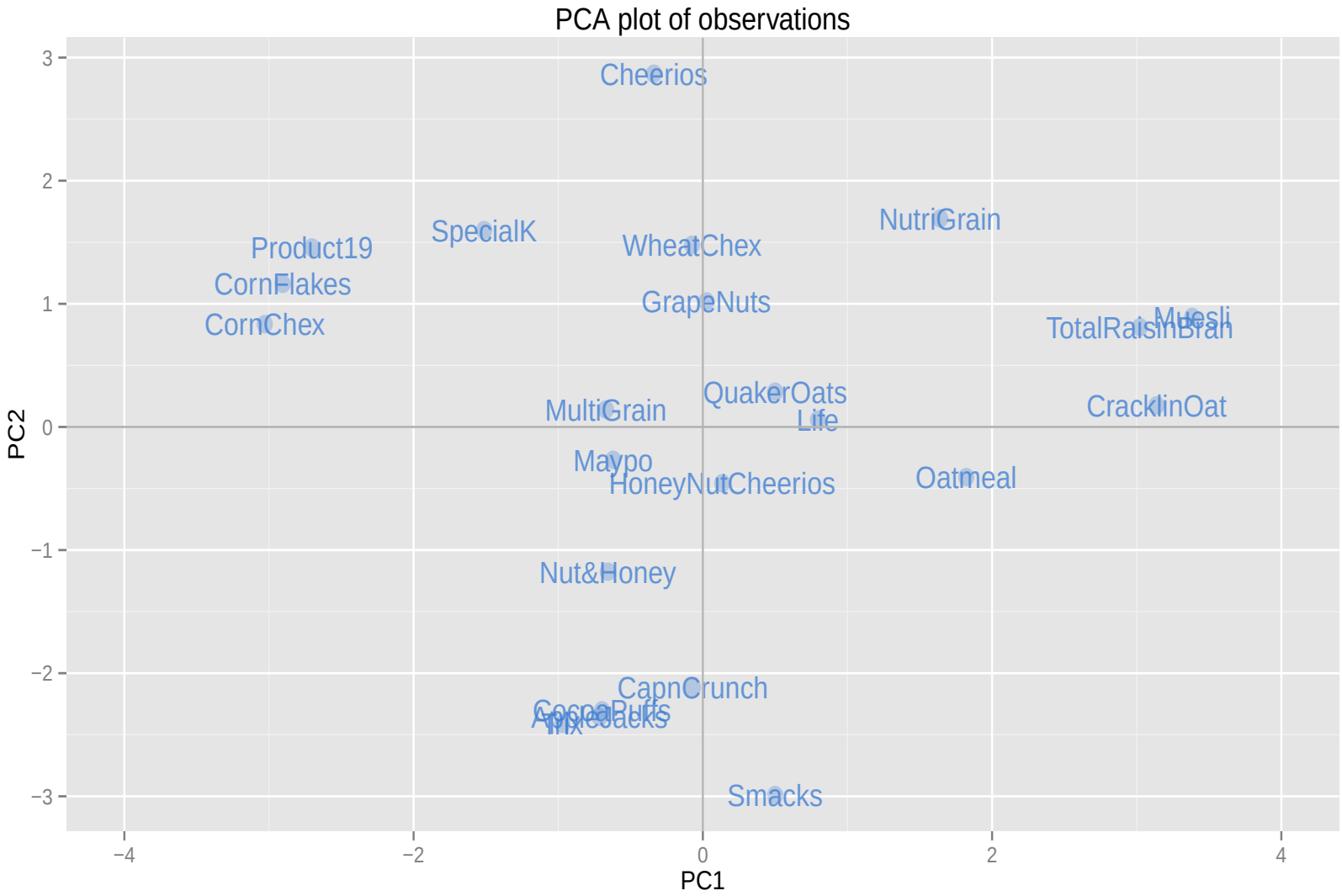
(Dis)similarities among cereals?

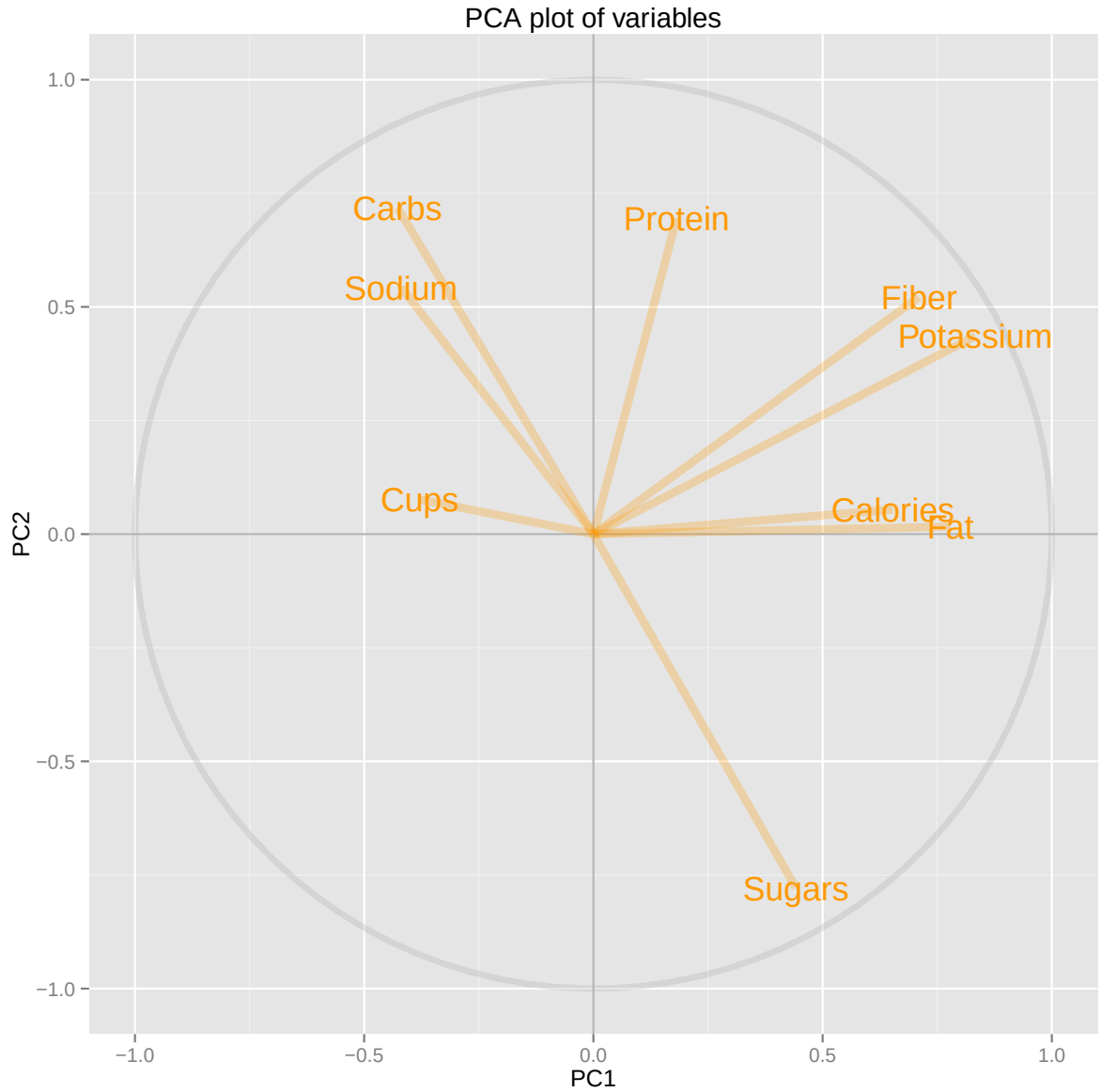
Relationships between variables?

Any patterns of variation?

The global structure of dispersion?

A picture is worth a thousand numbers


PCA plot of observations
Cheerios
SpecialK
NutriGrain
Product19
WheatChex
CornFlakes
GrapeNuts
CornChex
Muesli
TotalRaisinBran
QuakerOats
MultiGrain
CracklinOat
Life
Maypo
HoneyNutCheerios
Oatmeal
Nut&Honey
CapnCrunch
CocoaPuffs
Smacks
PC1
PC2
3
2
1
0
−1
−2
−3
−4
−2
0
2
4


PCA plot of variables
Carbs
Protein
Sodium
Fiber
Potassium
Cups
Calories
Fat
Sugars
PC1
PC2
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

Introduction

PCA

Principal Components Analysis (PCA) is a multivariate method that allows us to study and explore a data set of quantitative variables measured on a set of objects

Importance

PCA

Principal Components Analysis (PCA) is the building block for many multivariate techniques

Core Idea

PCA is perhaps the most popular dimension reduction technique, and it is applied in almost all scientific disciplines

Introduction (con't)

Core Idea

With PCA we seek to **reduce the dimensionality** (reduce the number of variables) of a data set while retaining as much as possible of the variation present in the data

Global Concepts

Landmarks

Good to know

- PCA was first introduced by Karl Pearson (1904)
  On lines and planes of closest fit to systems of points in space
- Further developed by Harold Hotelling (1936)
  Analysis of a complex of statistical variables into principal components
- Singular Value Decomposition (SVD) theorem by Eckart-Young (1936)
  The approximation of a matrix by another of a lower rank
- Computationally implemented with computers (1960s)

Common Usage

Typically, we perform a PCA to get a graphical display of a multivariate data set so we can better understand its main structural features

Geometric Idea

One of the goals behind PCA is to graphically represent the essential information contained in a (quantitative) data table

PCA Concept

Summarizing Information

PCA allows us to obtain a descriptive model that we can use for tasks that would benefit from the insight gained from summarizing the data in new and interesting ways.

Pattern Discovery?

PCA helps us in pattern discovery tasks: used to identify frequent associations within data.

PCA Overall Goals

Summarizing Information

- Extract the most important information from a data table
- Compress the size of the data set by keeping only this important information
- Simplify the description of the data set
- Analyze the structure of the observations and the variables

Applications

PCA can be used for

1. Dimension Reduction
2. Visualization
3. Feature Extraction
4. Data Compression
5. Smoothing of Data
6. Detection of Outliers
7. Preliminary process for further analyses

Principal Components?

Meaning of *Principal*

The term **Principal**, as used in PCA, has to do with the notion of **principal axis** from geometry and linear algebra

Principal Axis

A *principal axis* is a certain line in a Euclidean space associated to an ellipsoid or hyperboloid, generalizing the major and minor axes of an ellipse

PCA Approaches

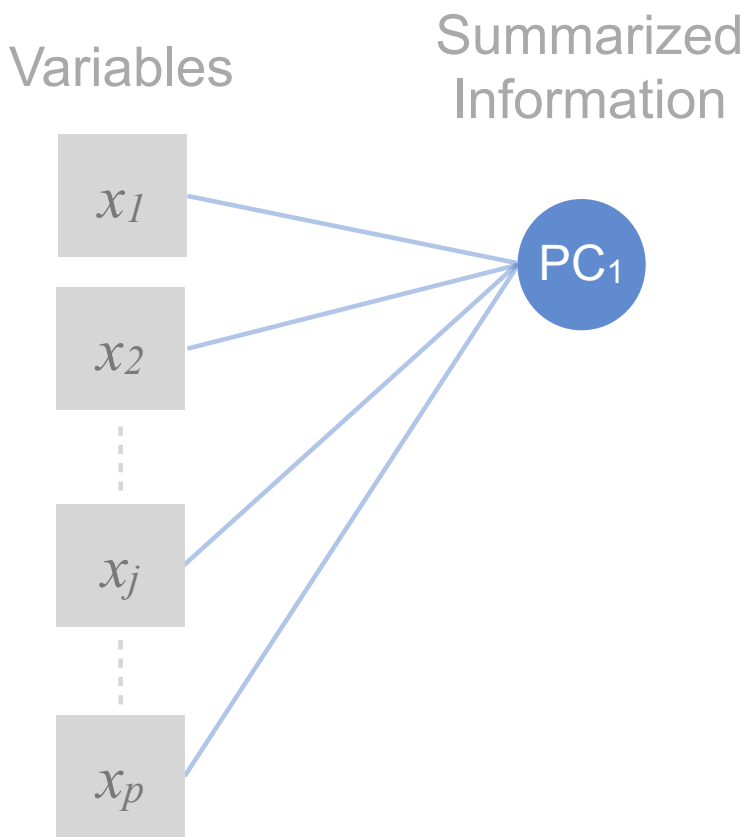
Looking at PCA from different perspectives

PCA allows us

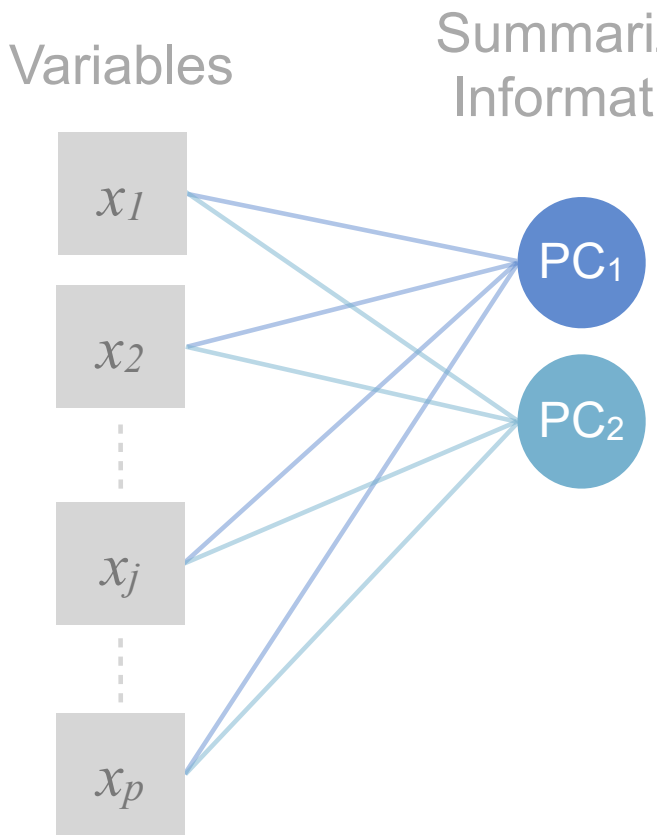
- to **"best summarize"** the important information contained in a data table
- to find a **"graphical representation"** of the essential information contained within a data set
- to find an **"optimal approximation"** of a data set with a minimal loss of information

PCA as Best Summary of Information

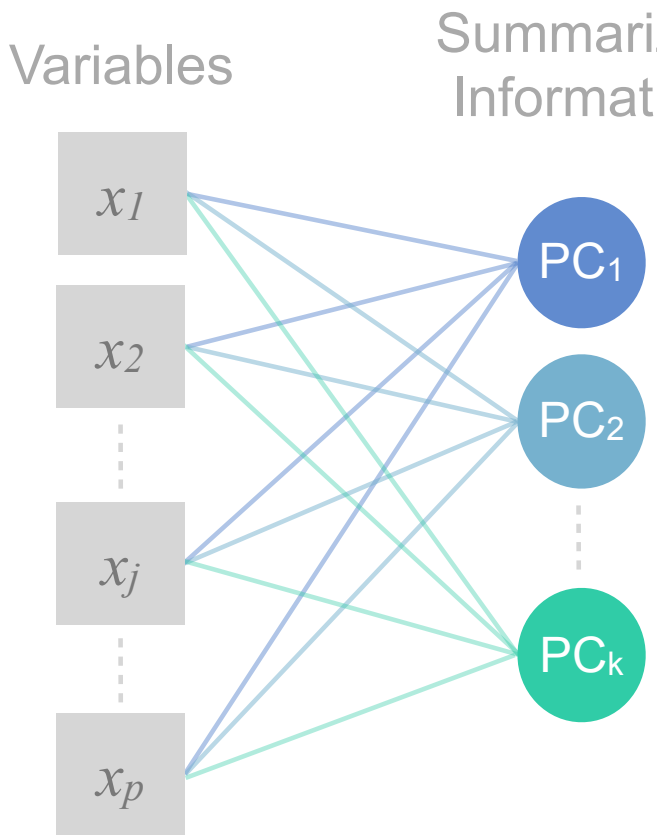
PCA as Best Summary of Information



PCA as Best Summary of Information

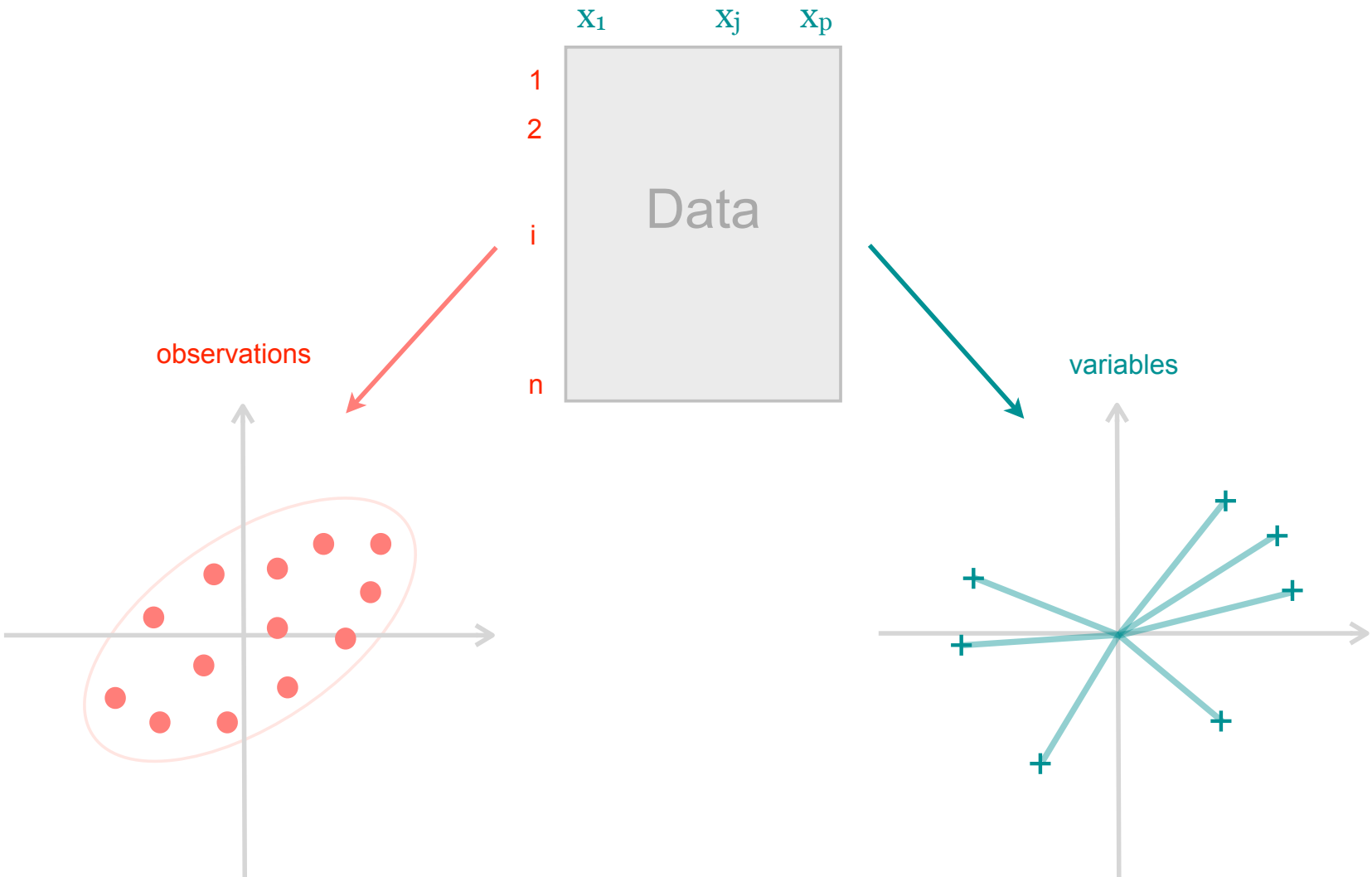


PCA as Best Summary of Information

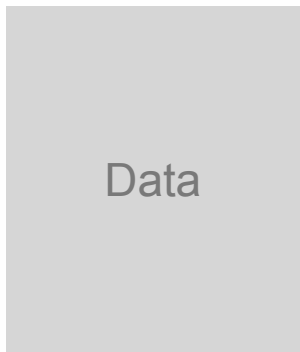


PCA for Graphical Representation

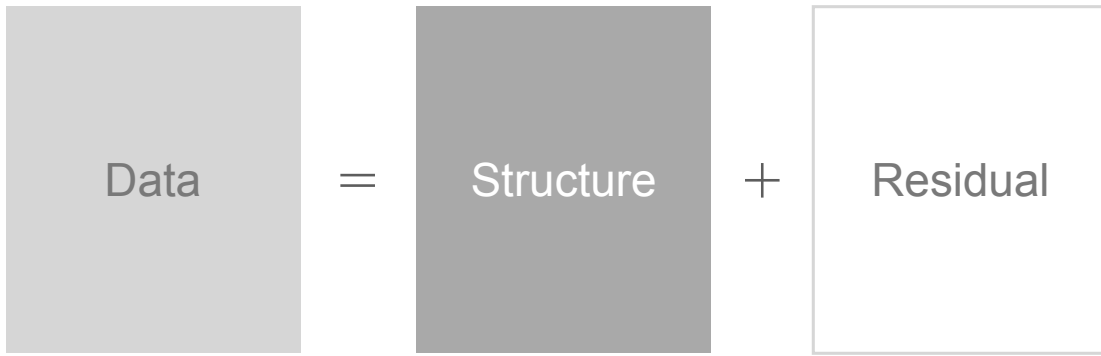
PCA for Graphical Representation



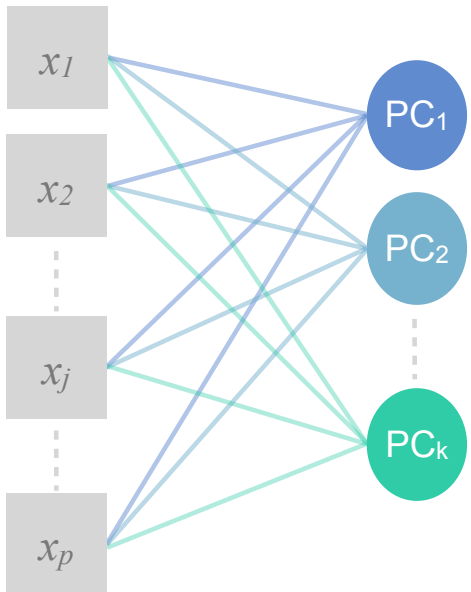
PCA as Optimal Approximation of Data

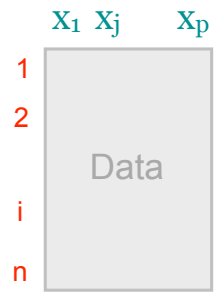


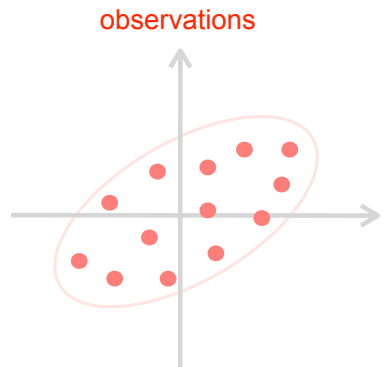
PCA as Optimal Approximation of Data

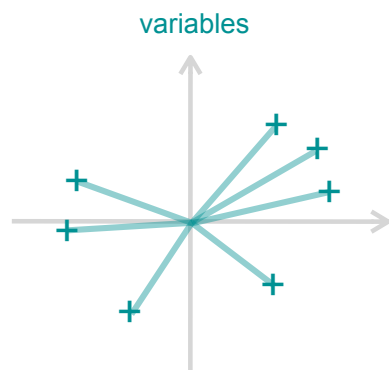


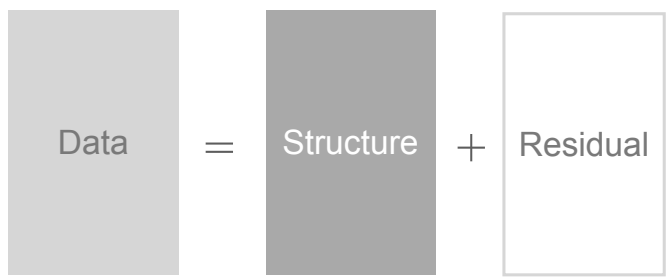
Three Perspectives for PCA











PCA Concept

Perspectives

PCA can be presented using various —different but equivalent— approaches. Each approach corresponds to a unique perspective and a way of thinking about data.

The most common approaches:

- Data in terms of variation
- Data as points (i.e. vectors) in a multidimensional space
- Data that follows a model

PCA Criteria

Approaches

The most common approaches are:

- Algebraic: Eigenvalue Decomposition
- Geometric: Projected Inertia
- Minimization: Least Squares and Low-Rank Approximation

