PCA Revealed

Part 1: Presentation

Gaston Sanchez

August 2014

Readme

License:

You are free to:

Under the following conditions:

PCA Revealed?

About

PCA Revealed aims to help you understanding in breadth and depth the legendary data analysis method Principal Components Analysis (PCA).

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 ...

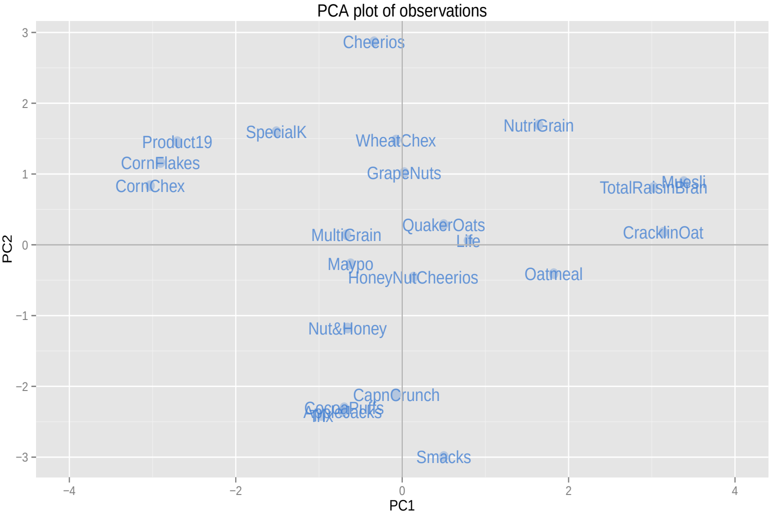
(Dis)similarities among cereals?

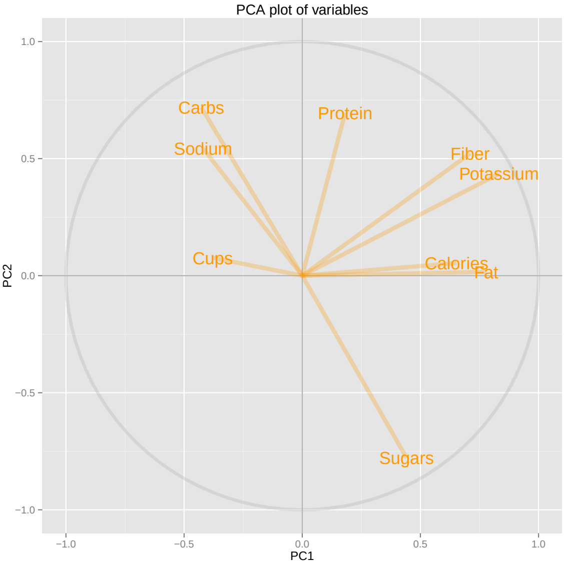
Relationships between variables?

Any patterns of variation?

The global structure of dispersion?

A picture is worth a thousand numbers


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


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

Presentation

Considerations

Scope

Our aim is to study PCA thoroughly, considering its theoretical principles, associated procedures, and application guidelines.

Expectations

- understand the main concepts and notions behind PCA
- know when and how to apply it in practice
- evaluate, interpret and diagnose the provided results

Requirements

Must have:

- Exploratory data analysis attitude
- Keen interest in data visualization
- Knowledge of basic stats concepts (mean, variance, etc)
- Knowledge of basic matrix algebra concepts

Nice to have:

- Previous exposure to PCA
- Solid knowledge about linear (i.e. matrix) algebra
- Some experience working with R
- Some basic programming skills

Keep in mind

Software

We will use the statistical programming language **R** for computations, and its related packages for applying PCA.

Advice

You don't have to memorize all the material, concepts, formulas, commands, etc. Instead, focus on understanding what things mean (and put concepts in your own words).

Resources

Some Books (in English)

- Principal Component Analysis
  by Ian T. Jolliffe
- A User's Guide To Principal Components
  by J. Edward Jackson
- Principles of Multivariate Analysis
  by Wojtek J. Krzanowski
- Exploratory Multivariate Analysis by Example Using R
  by Francois Husson, Sebastien Le, Jerome Pages

French Resources

Some Books (in French)

- Probabilites, Analyse de Donnees et Statistique
  by Gilbert Saporta
- Statistique
  by Michel Tenenhaus
- Analyses Factorielles Simples et Multiples
  by Brigitte Escofier and Jerome Pages
- Statistique Exploratoire Multidimensionnelle
  by Ludovic Lebart, Marie Piron, and Alain Morineau
- Analyse des Donnees
  by Michel Volle

Hard to come by resources (for PCA geeks)

Other Books (not in english, hard to find, but priceless)

- Aprender de los Datos: El Analisis de Componentes Principales
  by Tomas Aluja and Alain Morineau
- Analyse en Composantes Principales (avec illustrations SPAD)
  by Alain Morineau and Tomas Aluja
- L'Analyse des Donnees, Vols. 1 and 2
  by Jean-Paul Benzecri

Outline

Slides

- Preamble
    - Introduction (slides 2)
    - Preliminary Concepts (slides 3)
- Theory
    - PCA from a Summarizing Information Approach (slides 4)
    - PCA from a Geometric Approach (slides 5)
    - PCA from an Minimization Approach (slides 6)
- Practice
    - PCA with R (slides 7)
    - Hacking your own PCA (slides 8)
- Appendix (slides 9)

I've tried to make each slide-deck as much self-contained as possible. If you're in a hurry, you can check them individually without having to go through all of them sequentially.