Research questions and corresponding statistical tests

The most popular research questions include:

1. whether **two variables** (n = 2) are **correlated** (i.e., associated)
2. whether **multiple variables** (n > 2) are **correlated**
3. whether **two groups** (n = 2) of samples **differ** from each other
4. whether **multiple groups** (n >= 2) of samples **differ** from each other
5. whether the **variability** of two samples differ

Each of these questions can be answered using the following statistical tests:

1. **Correlation test** between two variables
2. **Correlation matrix** between multiple variables
3. **Comparing the means of two groups**:
   - **Student's t-test** (parametric)
   - **Wilcoxon rank test** (non-parametric)
4. **Comparing the means of more than two groups**
   - **ANOVA test** (analysis of variance, parametric): extension of t-test to compare more than two groups.
   - **Kruskal-Wallis rank sum test** (non-parametric): extension of Wilcoxon rank test to compare more than two groups
5. **Comparing the variances**:
   - Comparing the variances of two groups: **F-test** (parametric)
   - Comparison of the variances of more than two groups: **Bartlett's test** (parametric), **Levene's test** (parametric) and **Fligner-Killeen test** (non-parametric)

Statistical test requirements (assumptions)

Many of the statistical procedures including correlation, regression, t-test, and analysis of variance assume some certain characteristic about the data. Generally they assume that:

- the data are **normally distributed**
- and the **variances** of the groups to be compared are **homogeneous** (equal).

These assumptions should be taken seriously to draw reliable interpretation and conclusions of the research.

These tests - correlation, t-test and ANOVA - are called **parametric tests**, because their validity depends on the distribution of the data.

Before using parametric test, we should perform some **preleminary tests** to make sure that the test assumptions are met. In the situations where the assumptions are violated, **non-paramatric** tests are recommended.

How to assess the normality of the data?

1. With **large enough sample sizes** (n > 30) the violation of the normality assumption should not cause major problems (central limit theorem). This implies that we can ignore the distribution of the data and use parametric tests.
2. However, to be consistent, we can use **Shapiro-Wilk's significance test** comparing the sample distribution to a normal one in order to ascertain whether data show or not a serious deviation from normality.

How to assess the equality of variances?

The standard **Student's t-test** (comparing two independent samples) and the ANOVA test (comparing multiple samples) assume also that the samples to be compared have equal variances.

If the samples, being compared, follow normal distribution, then it's possible to use:

- **F-test** to compare the variances of two samples
- **Bartlett's Test** or **Levene's Test** to compare the variances of multiple samples.

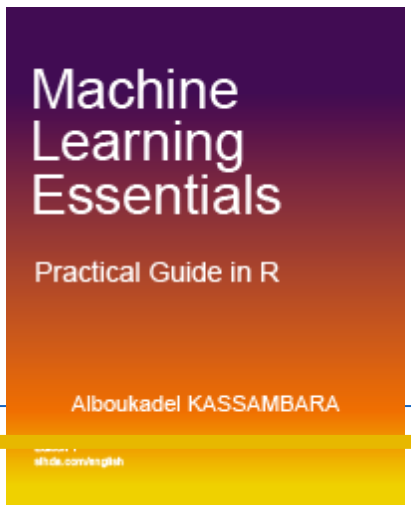
Infos

This analysis has been performed using **R software** (ver. 3.2.4).

Enjoyed this article? I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

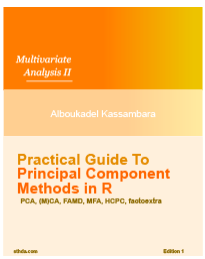
Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!



Machine Learning Essentials: Practical Guide in R

Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R

Normality and the other assumptions made by these tests should be taken seriously to draw reliable interpretation and conclusions of the research.

Before using a parametric test, we should perform some **preleminary tests** to make sure that the test assumptions are met. In the situations where the assumptions are violated, **non-paramatric** tests are recommended.

Here, we'll describe how to check the normality of the data by visual inspection and by significance tests.

Install required R packages

1. **dplyr** for data manipulation

```
install.packages("dplyr")
```

2. ggpubr for an easy ggplot2-based data visualization

- Install the latest version from GitHub as follow:

```
# Install
if(!require(devtools)) install.packages("devtools")
devtools::install_github("kassambara/ggpubr")
```

- Or, install from CRAN as follow:

```
install.packages("ggpubr")
```

Load required R packages

```
library("dplyr")
library("ggpubr")
```

Import your data into R

1. **Prepare your data** as specified here: Best practices for preparing your data set for R
2. **Save your data** in an external .txt tab or .csv files
3. **Import your data into R** as follow:

```
# If .txt tab file, use this
my_data <- read.delim(file.choose())
# Or, if .csv file, use this
my_data <- read.csv(file.choose())
```

Here, we'll use the built-in R data set named ToothGrowth.

```
# Store the data in the variable my_data
my_data <- ToothGrowth
```

Check your data

We start by displaying a random sample of 10 rows using the function **sample_n**()[in **dplyr** package].

Show 10 random rows:

```
set.seed(1234)
dplyr::sample_n(my_data, 10)
```

```
    len supp dose
7  11.2   VC  0.5
37  8.2   OJ  0.5
36 10.0   OJ  0.5
58 27.3   OJ  2.0
49 14.5   OJ  1.0
57 26.4   OJ  2.0
1   4.2   VC  0.5
13 15.2   VC  1.0
35 14.5   OJ  0.5
27 26.7   VC  2.0
```

Assess the normality of the data in R

We want to test if the variable *len* (tooth length) is normally distributed.

Case of large sample sizes

If the sample size is large enough (n > 30), we can ignore the distribution of the data and use parametric tests.

The central limit theorem tells us that no matter what distribution things have, the sampling distribution tends to be normal if the sample is large enough (n > 30).

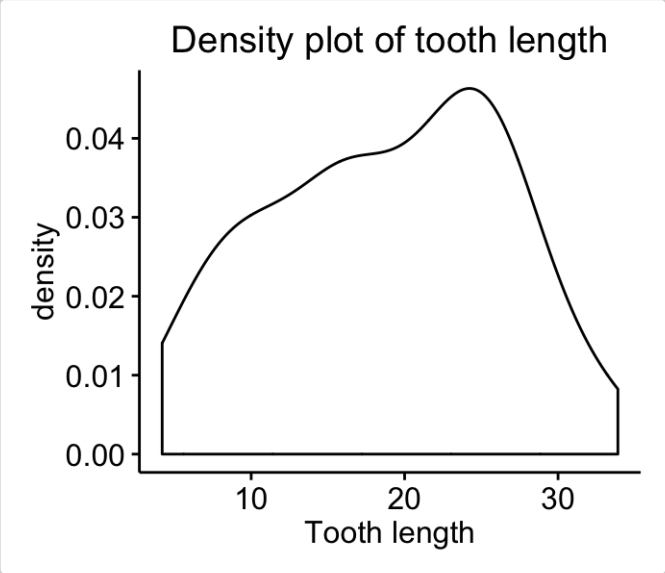
However, to be consistent, normality can be checked by visual inspection [**normal plots (histogram)**, **Q-Q plot** (quantile-quantile plot)] or by **significance tests**].

Visual methods

Density plot and **Q-Q plot** can be used to check normality visually.

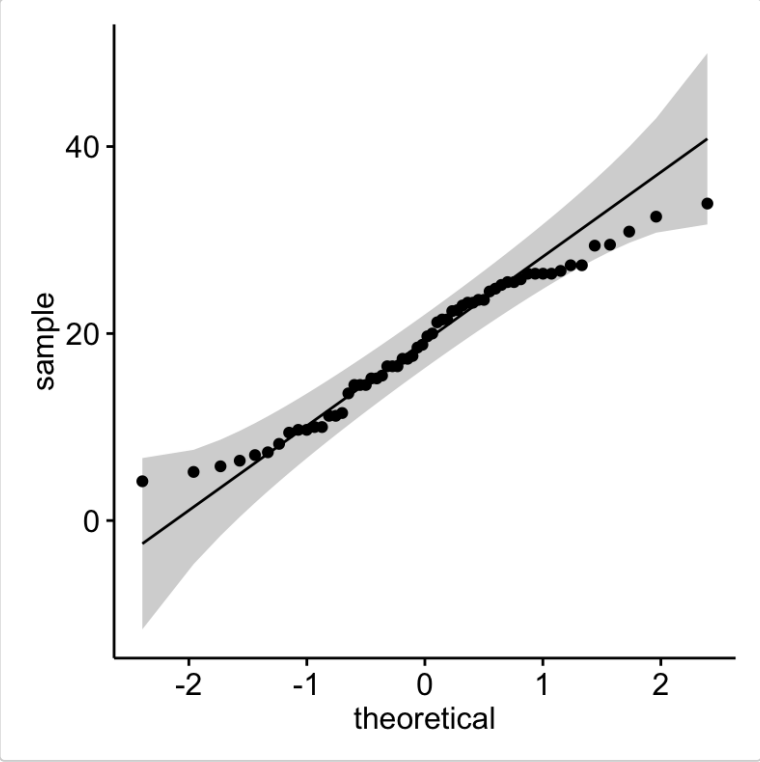
1. **Density plot**: the **density** plot provides a visual judgment about whether the distribution is bell shaped.

```
library("ggpubr")
ggdensity(my_data$len,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```



2. **Q-Q plot**: **Q-Q plot** (or quantile-quantile plot) draws the correlation between a given sample and the normal distribution. A 45-degree reference line is also plotted.

```
library(ggpubr)
ggqqplot(my_data$len)
```



It's also possible to use the function **qqPlot**() [in **car** package]:

```
library("car")
qqPlot(my_data$len)
```

As all the points fall approximately along this reference line, we can assume normality.

Normality test

Visual inspection, described in the previous section, is usually unreliable. It's possible to use a **significance test** comparing the sample distribution to a normal one in order to ascertain whether data show or not a serious deviation from normality.

There are several methods for **normality test** such as **Kolmogorov-Smirnov (K-S) normality test** and **Shapiro-Wilk's test**.

The null hypothesis of these tests is that "sample distribution is normal". If the test is **significant**, the distribution is non-normal.

Shapiro-Wilk's method is widely recommended for normality test and it provides better power than K-S. It is based on the correlation between the data and the corresponding normal scores.

Note that, normality test is sensitive to sample size. Small samples most often pass normality tests. Therefore, it's important to combine visual inspection and significance test in order to take the right decision.

The R function **shapiro.test**() can be used to perform the Shapiro-Wilk test of normality for one variable (univariate):

```
shapiro.test(my_data$len)
```

```
    Shapiro-Wilk normality test
data:  my_data$len
W = 0.96743, p-value = 0.1091
```

From the output, the p-value > 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

Infos

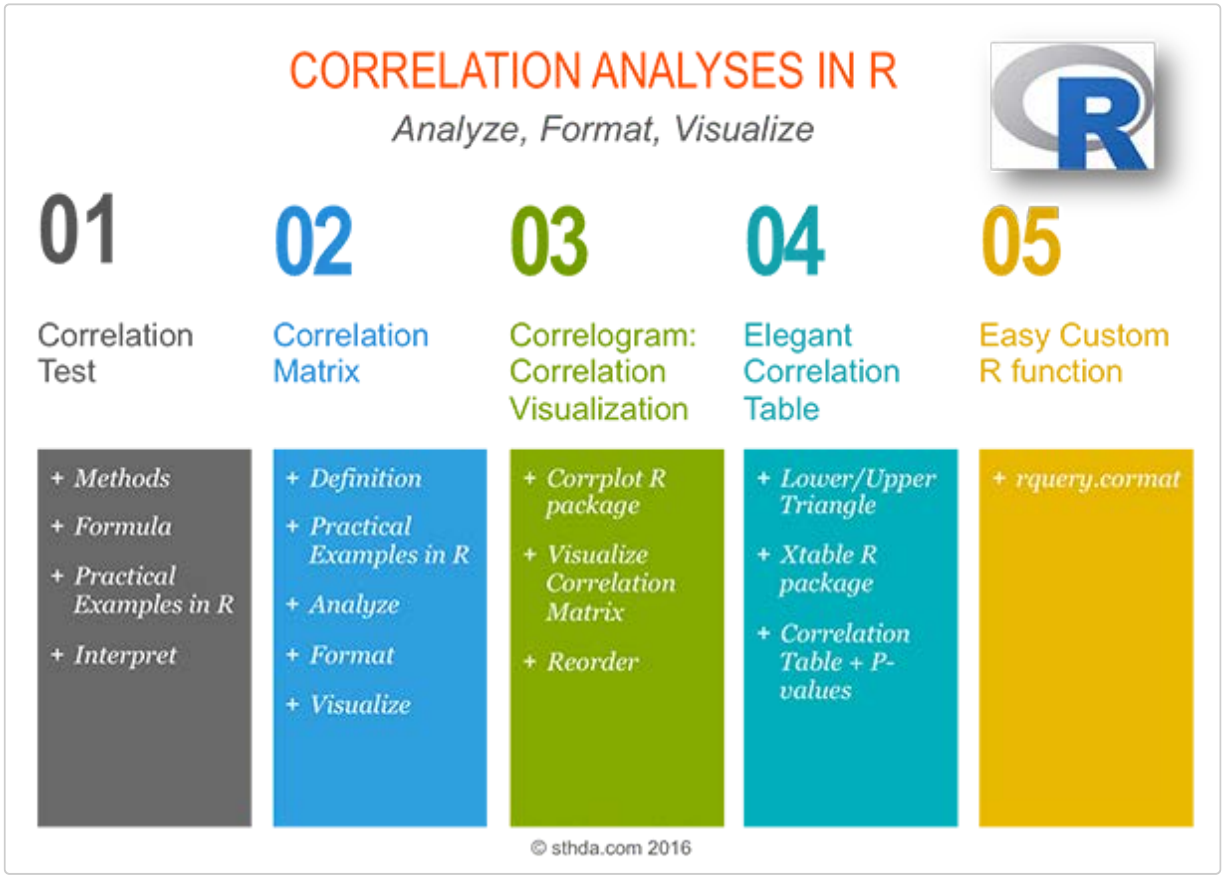
This analysis has been performed using **R software** (ver. 3.2.4).

Enjoyed this article? I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!

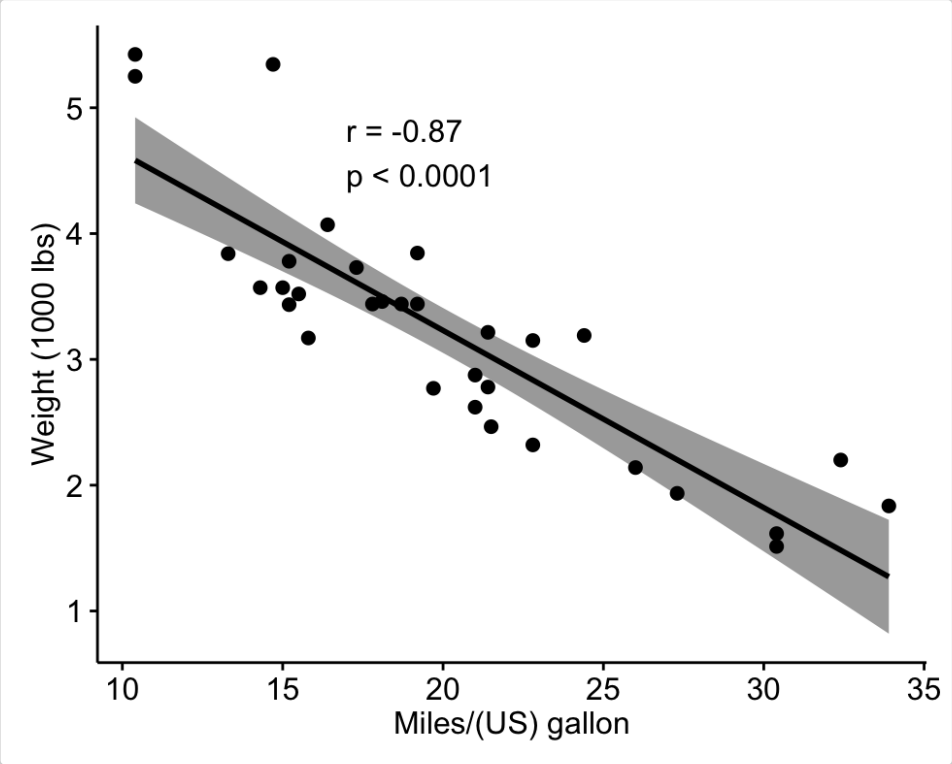
How this chapter is organized?



Correlation Test Between Two Variables in R

Brief outline:

- What is correlation test?
- Methods for correlation analyses
- Correlation formula
  - Pearson correlation formula
  - Spearman correlation formula
  - Kendall correlation formula
- Compute correlation in R
  - R functions
  - Import your data into R
  - Visualize your data using scatter plots
  - Preliminary test to check the test assumptions
  - Pearson correlation test
  - Kendall rank correlation test
  - Spearman rank correlation coefficient
- Interpret correlation coefficient



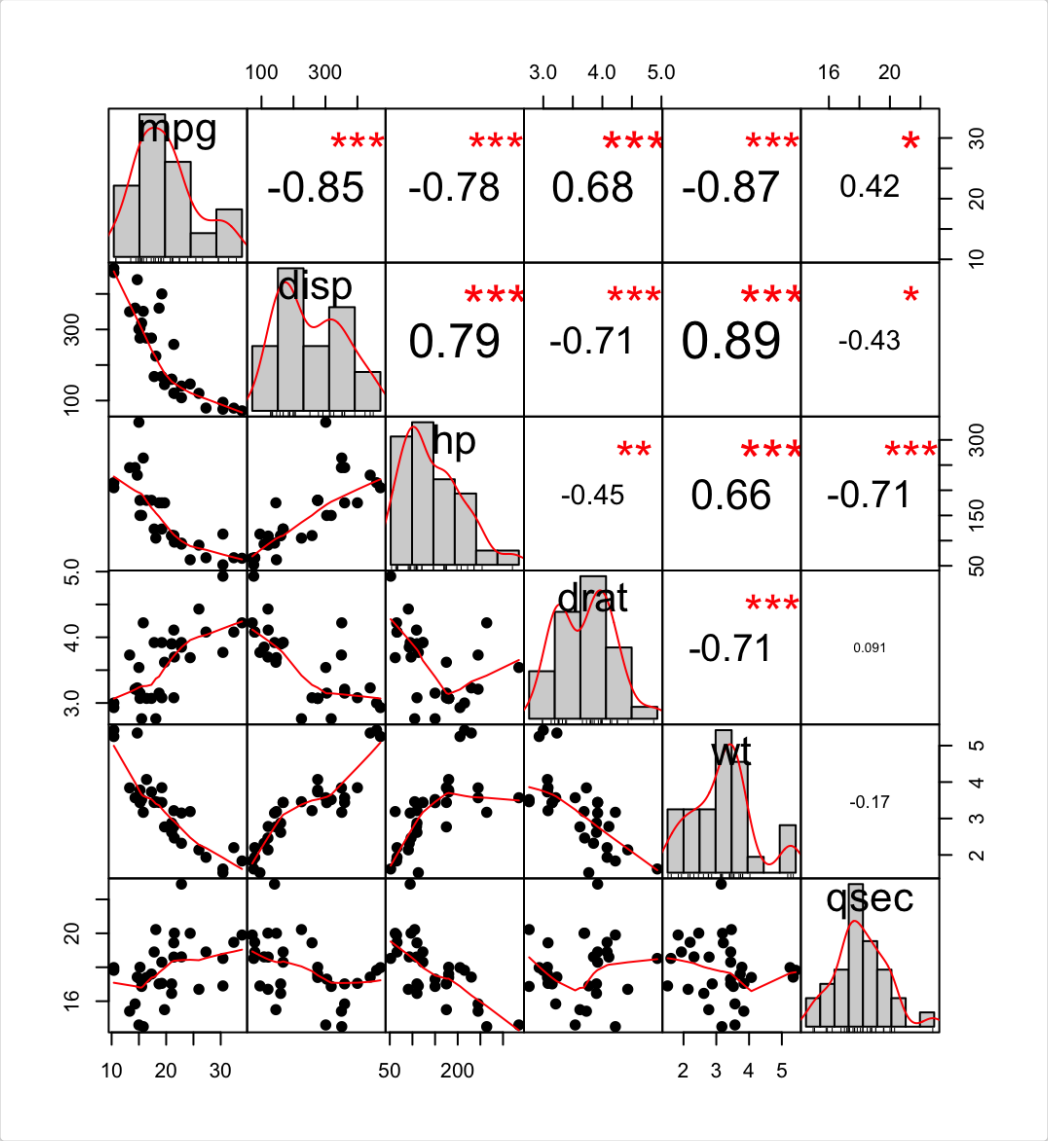
Read more: —> Correlation Test Between Two Variables in R.

Correlation Matrix: Analyze, Format and Visualize

Correlation matrix is used to analyze the correlation between multiple variables at the same time.

Brief outline:

- What is correlation matrix?
- Compute correlation matrix in R
  - R functions
  - Compute correlation matrix
  - Correlation matrix with significance levels (p-value)
  - A simple function to format the correlation matrix
  - Visualize correlation matrix
    - Use symnum() function: Symbolic number coding
    - Use corrplot() function: Draw a correlogram
    - Use chart.Correlation(): Draw scatter plots
    - Use heatmap()



Read more: —> Correlation Matrix: Analyze, Format and Visualize.

Visualize Correlation Matrix using Correlogram

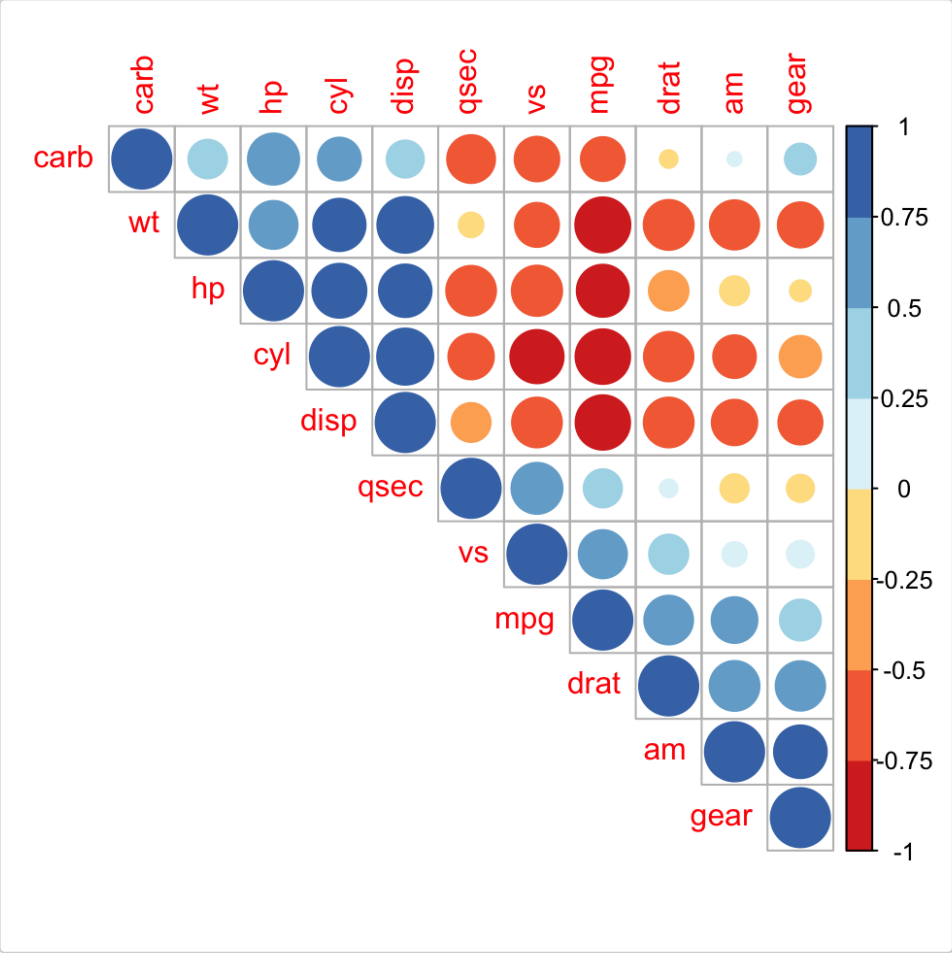
Correlogram is a **graph of correlation matrix**. Useful to highlight the most correlated variables in a data table. In this plot, **correlation coefficients** are colored according to the value. **Correlation matrix** can be also reordered according to the degree of association between variables.

Brief outline:

- Install R corrplot package
- Data for correlation analysis
- Computing correlation matrix
- Correlogram : Visualizing the correlation matrix
  - Visualization methods
  - Types of correlogram layout
  - Reordering the correlation matrix
  - Changing the color of the correlogram
  - Changing the color and the rotation of text labels
  - Combining correlogram with the significance test
  - Customize the correlogram

```
library(corrplot)
library(RColorBrewer)
M <-cor(mtcars)
```

```
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))
```



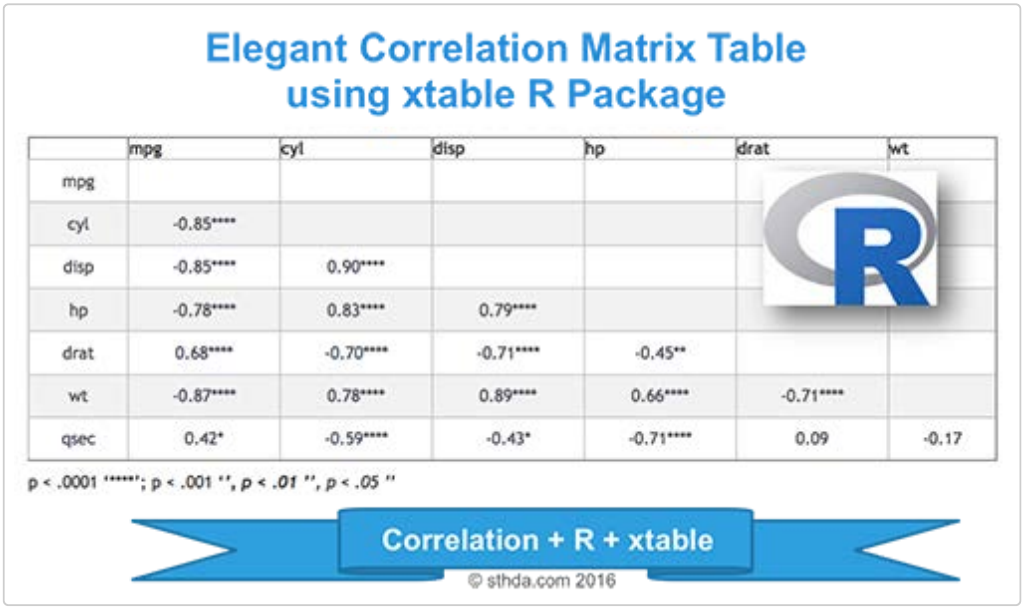
Read more: —> Visualize Correlation Matrix using Correlogram.

Elegant Correlation Table using xtable R Package

The aim of this article is to show you how to get the **lower and the upper triangular part of a correlation matrix**. We will use also **xtable R package** to display a nice **correlation table**.

Brief outline:

- Correlation matrix analysis
- Lower and upper triangular part of a correlation matrix
- Use xtable R package to display nice correlation table in html format
- Combine matrix of correlation coefficients and significance levels



Read more: —> Elegant correlation table using xtable R package.

Correlation Matrix : An R Function to Do All You Need

The goal of this article is to provide you a custom **R function**, named **rquery.cormat**(), for **calculating** and **visualizing** easily a **correlation matrix** in a single line R code.

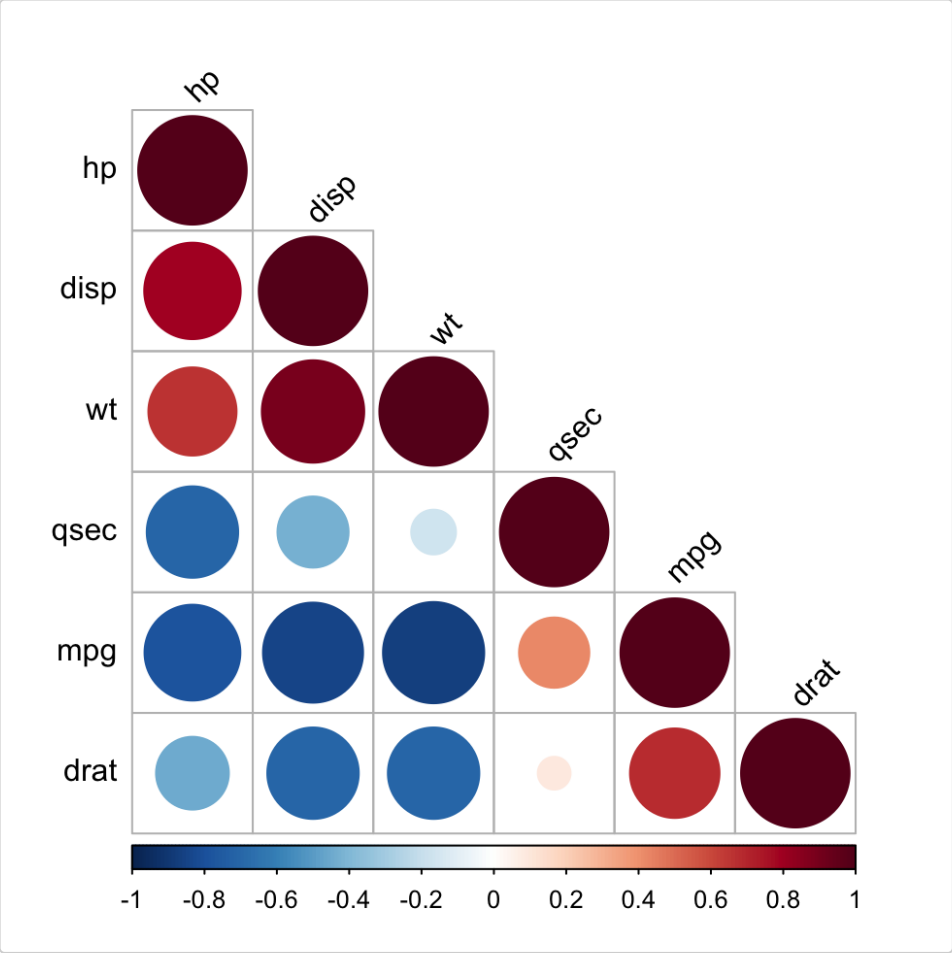
Brief outline:

- Computing the correlation matrix using rquery.cormat()
  - Upper triangle of the correlation matrix
  - Full correlation matrix
  - Change the colors of the correlogram
  - Draw a heatmap
- Format the correlation table
- Description of rquery.cormat() function

```
source("http://www.sthda.com/upload/rquery_cormat.r")
mydata <- mtcars[, c(1,3,4,5,6,7)]
require("corrplot")
rquery.cormat(mydata)
```

```
$r
        hp  disp    wt  qsec  mpg drat
hp       1
disp  0.79     1
wt    0.66  0.89     1
qsec -0.71 -0.43 -0.17     1
mpg  -0.78 -0.85 -0.87  0.42    1
drat -0.45 -0.71 -0.71 0.091 0.68    1
$p
          hp    disp      wt  qsec     mpg drat
hp         0
disp 7.1e-08       0
wt   4.1e-05 1.2e-11       0
qsec 5.8e-06   0.013    0.34     0
mpg  1.8e-07 9.4e-10 1.3e-10 0.017       0
drat    0.01 5.3e-06 4.8e-06  0.62 1.8e-05    0
$sym
     hp disp wt qsec mpg drat
hp   1
disp ,  1
wt   ,  +    1
qsec ,  .       1
mpg  ,  +    +  .    1
```

```
drat .  ,    ,      ,   1
attr(,"legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```



Read more: —> Correlation Matrix : An R Function to Do All You Need.

See also

- R Basics
- Import and Export Data using R
- Preparing and Reshaping Data in R for Easier Analyses
- Data Manipulation in R
- Data visualization
- Descriptive Statistics and Graphics

Infos

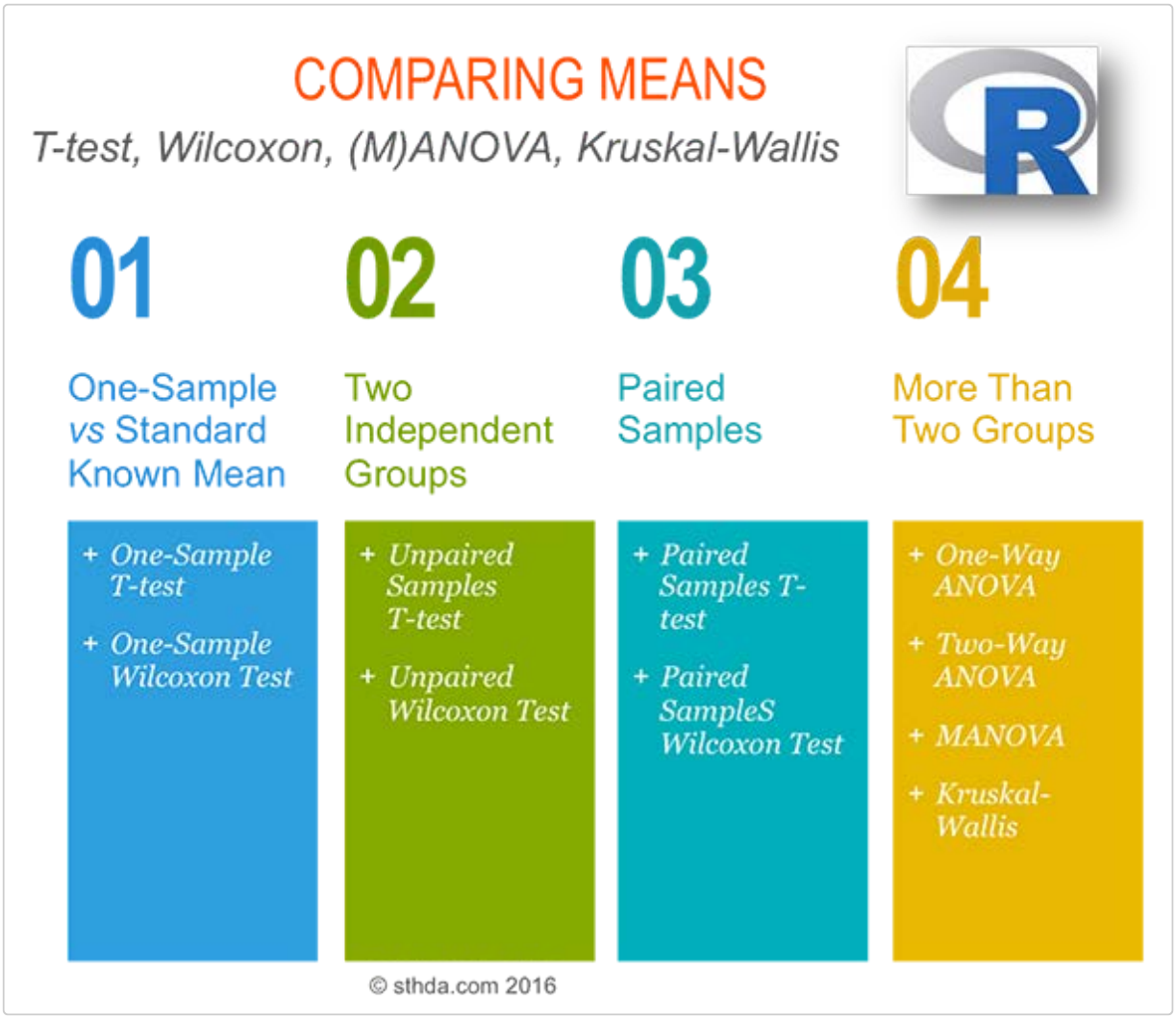
This analysis has been performed using **R statistical software** (ver. 3.2.4).

Enjoyed this article? I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

1 How this chapter is organized?

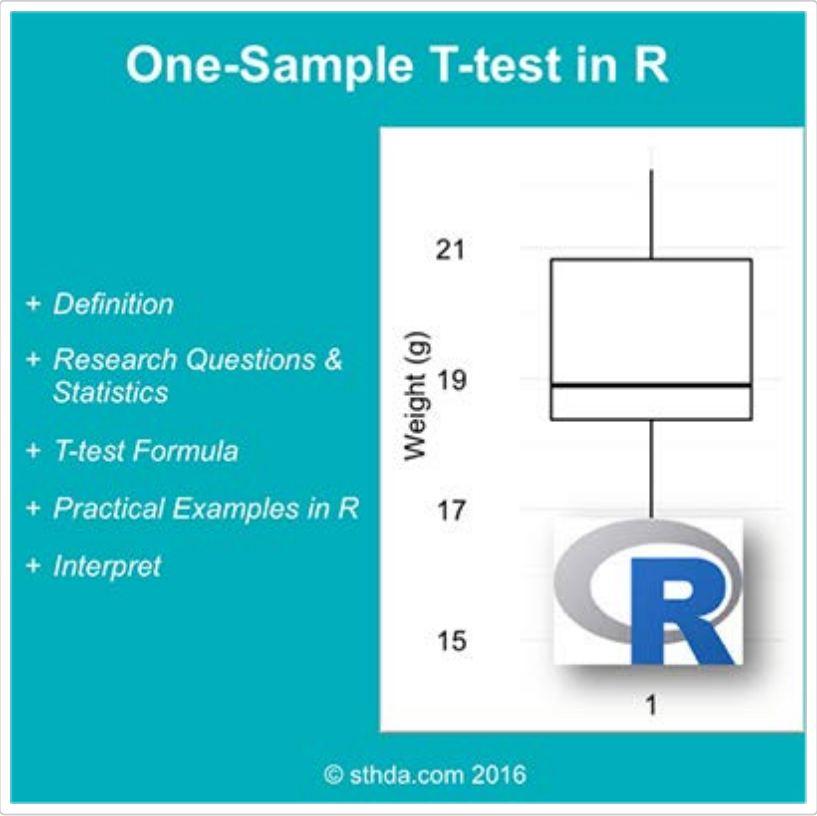
- Comparing one-sample mean to a standard known mean:
  - One-Sample T-test (parametric)
  - One-Sample Wilcoxon Test (non-parametric)
- Comparing the means of two independent groups:
  - Unpaired Two Samples T-test (parametric)
  - Unpaired Two-Samples Wilcoxon Test (non-parametric)
- Comparing the means of paired samples:
  - Paired Samples T-test (parametric)
  - Paired Samples Wilcoxon Test (non-parametric)
- Comparing the means of more than two groups
  - Analysis of variance (ANOVA, parametric):
    - One-Way ANOVA Test in R
    - Two-Way ANOVA Test in R
    - MANOVA Test in R: Multivariate Analysis of Variance
  - Kruskal-Wallis Test in R (non parametric alternative to one-way ANOVA)



2 Comparing one-sample mean to a standard known mean

2.1 One-sample T-test (parametric)

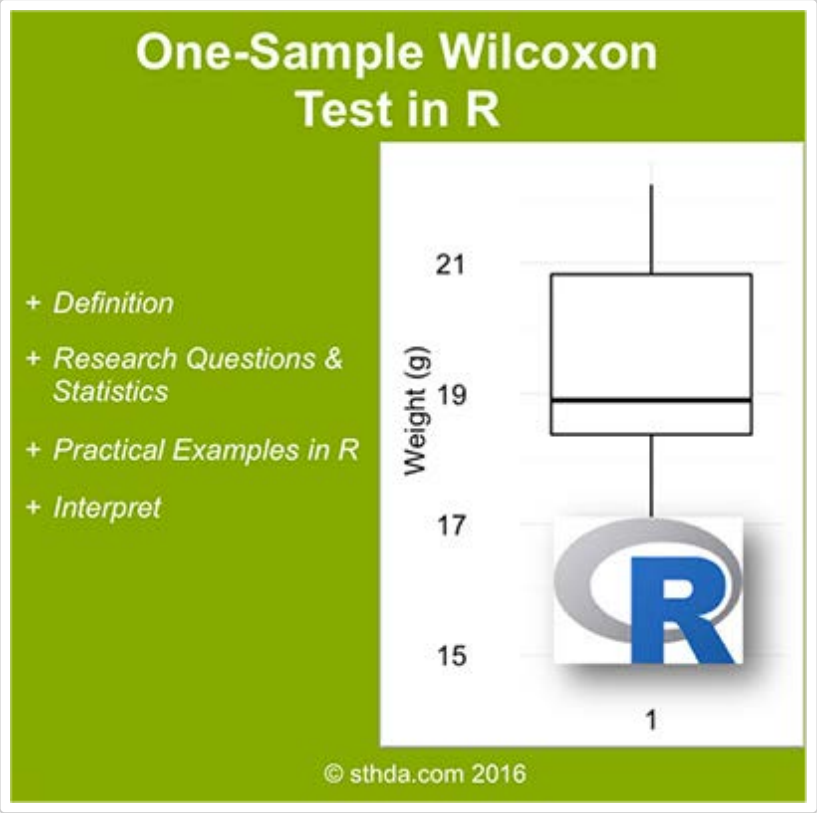
- What is one-sample t-test?
- Research questions and statistical hypotheses
- Formula of one-sample t-test
- Visualize your data and compute one-sample t-test in R
  - R function to compute one-sample t-test
  - Visualize your data using box plots
  - Preliminary test to check one-sample t-test assumptions
  - Compute one-sample t-test
  - Interpretation of the result



Read more: —> One-Sample T-test.

2.2 One-sample Wilcoxon test (non-parametric)

- What's one-sample Wilcoxon signed rank test?
- Research questions and statistical hypotheses
- Visualize your data and compute one-sample Wilcoxon test in R
  - R function to compute one-sample Wilcoxon test
  - Visualize your data using box plots
  - Compute one-sample Wilcoxon test

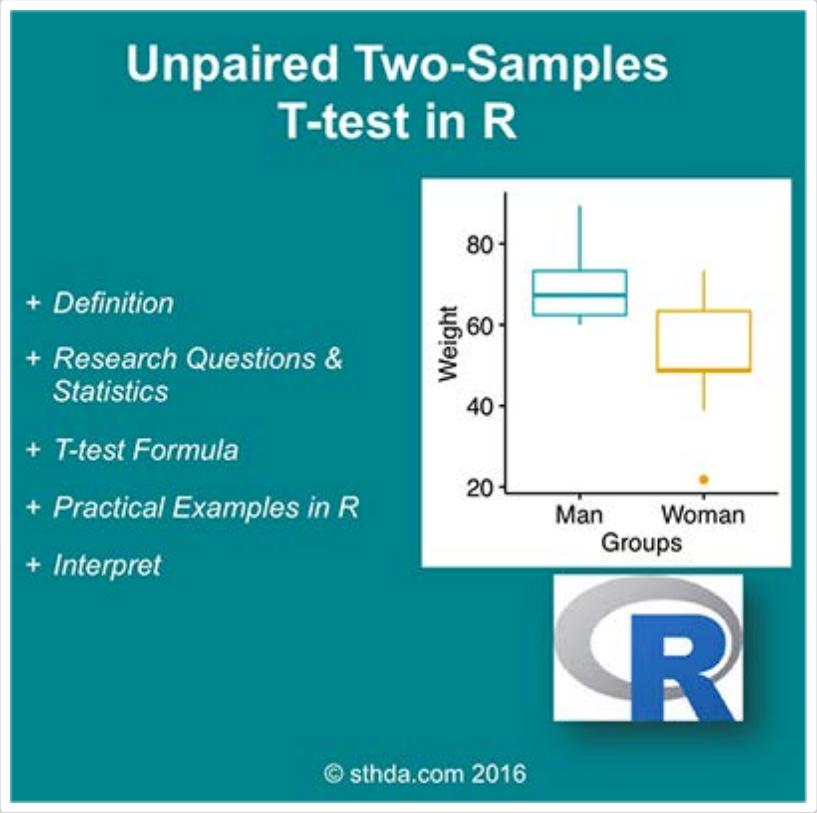


Read more: —> One-Sample Wilcoxon Test (non-parametric).

3 Comparing the means of two independent groups

3.1 Unpaired two samples t-test (parametric)

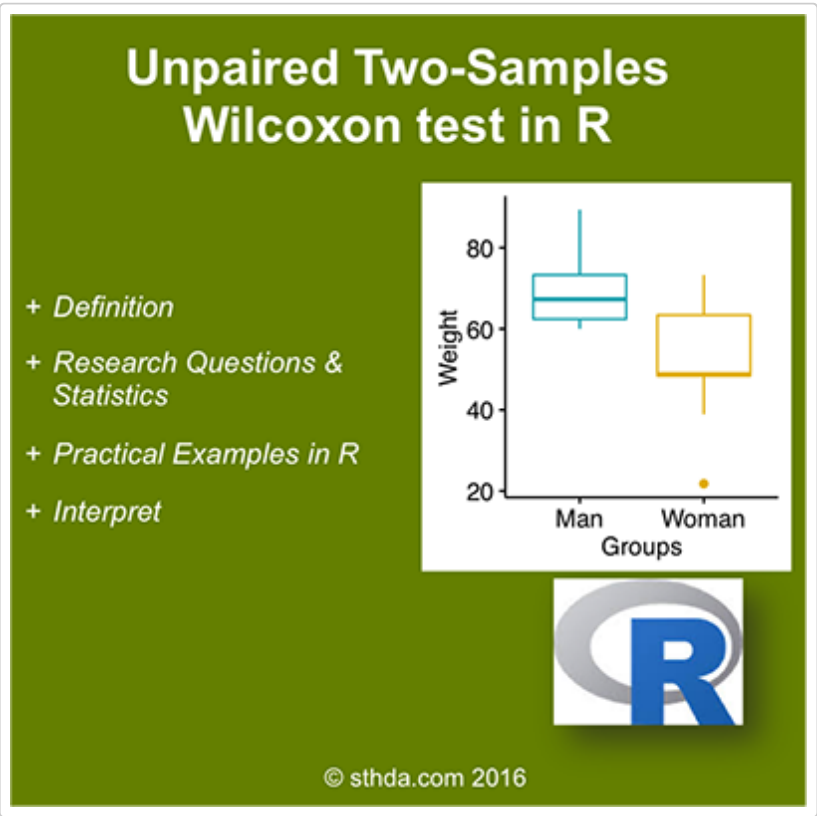
- What is unpaired two-samples t-test?
- Research questions and statistical hypotheses
- Formula of unpaired two-samples t-test
- Visualize your data and compute unpaired two-samples t-test in R
  - R function to compute unpaired two-samples t-test
  - Visualize your data using box plots
  - Preliminary test to check independent t-test assumptions
  - Compute unpaired two-samples t-test
- Interpretation of the result



Read more: —> Unpaired Two Samples T-test (parametric).

3.2 Unpaired two-samples Wilcoxon test (non-parametric)

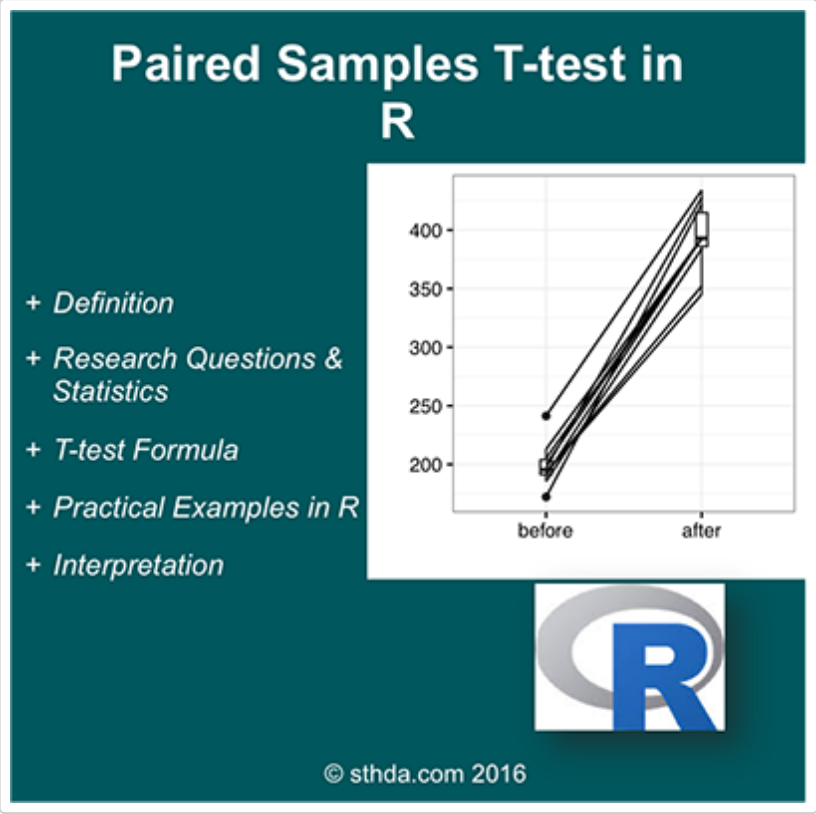
- R function to compute Wilcoxon test
- Visualize your data using box plots
- Compute unpaired two-samples Wilcoxon test



Read more: —> Unpaired Two-Samples Wilcoxon Test (non-parametric).

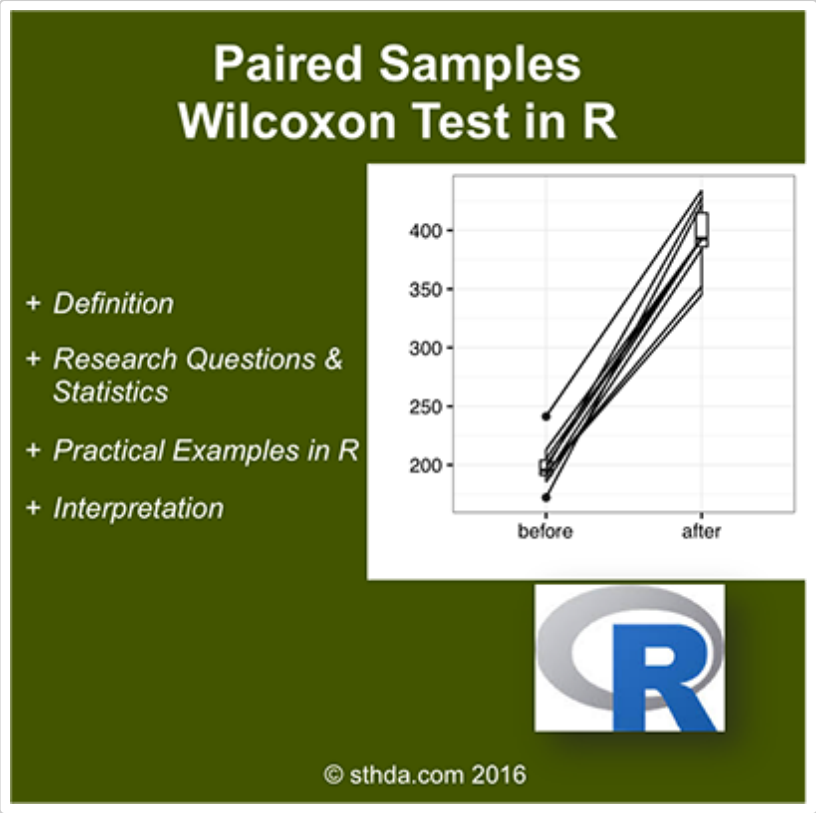
4 Comparing the means of paired samples

4.1 Paired samples t-test (parametric)



Read more: —> Paired Samples T-test (parametric).

4.2 Paired samples Wilcoxon test (non-parametric)



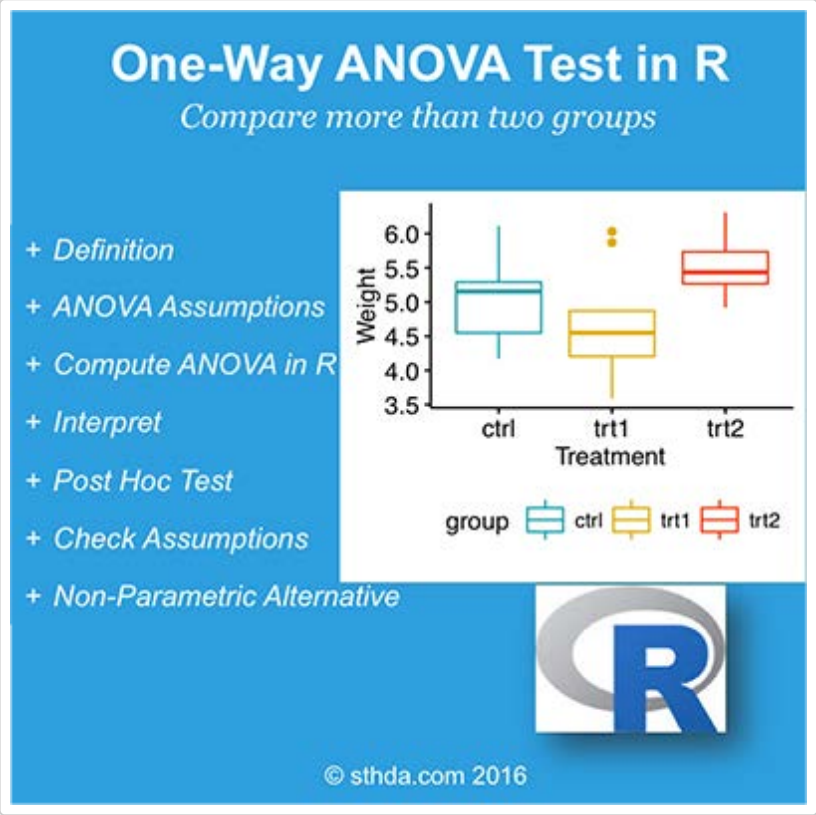
Read more: —> Paired Samples Wilcoxon Test (non-parametric).

5 Comparing the means of more than two groups

5.1 One-way ANOVA test

An extension of independent two-samples **t-test** for comparing means in a situation where there are more than two groups.

- What is one-way ANOVA test?
- Assumptions of ANOVA test
- How one-way ANOVA test works?
- Visualize your data and compute one-way ANOVA in R
  - Visualize your data
  - Compute one-way ANOVA test
  - Interpret the result of one-way ANOVA tests
  - Multiple pairwise-comparison between the means of groups
    - Tukey multiple pairewise-comparisons
    - Multiple comparisons using multcomp package
    - Pairwise t-test
  - Check ANOVA assumptions: test validity?
    - Check the homogeneity of variance assumption
    - Relaxing the homogeneity of variance assumption
    - Check the normality assumption
  - Non-parametric alternative to one-way ANOVA test

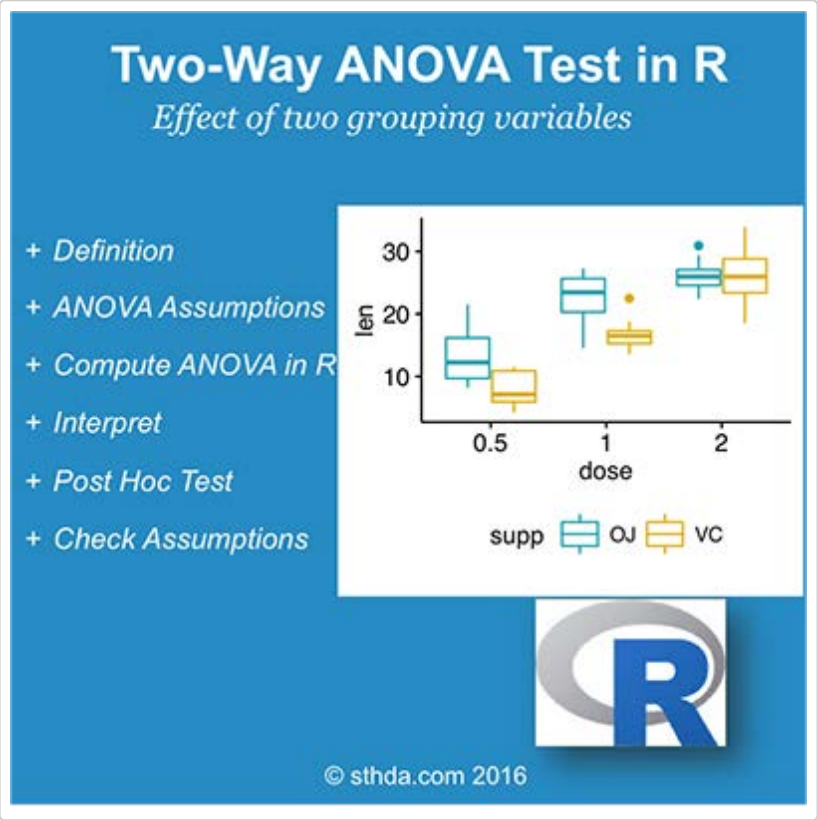


Read more: —> One-Way ANOVA Test in R.

5.2 Two-Way ANOVA test

- What is two-way ANOVA test?
- Two-way ANOVA test hypotheses

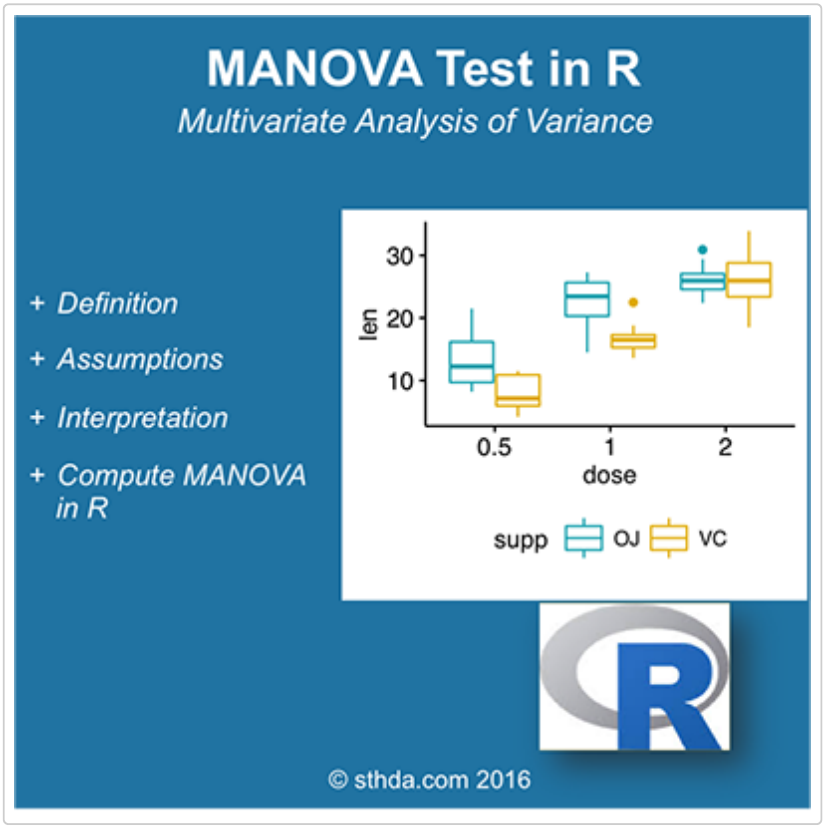
- Assumptions of two-way ANOVA test
- Compute two-way ANOVA test in R: balanced designs
  - Visualize your data
  - Compute two-way ANOVA test
  - Interpret the results
  - Compute some summary statistics
  - Multiple pairwise-comparison between the means of groups
    - Tukey multiple pairewise-comparisons
    - Multiple comparisons using multcomp package
    - Pairwise t-test
  - Check ANOVA assumptions: test validity?
    - Check the homogeneity of variance assumption
  - Check the normality assumption
- Compute two-way ANOVA test in R for unbalanced designs



Read more: —> Two-Way ANOVA Test in R.

6 MANOVA test: Multivariate analysis of variance

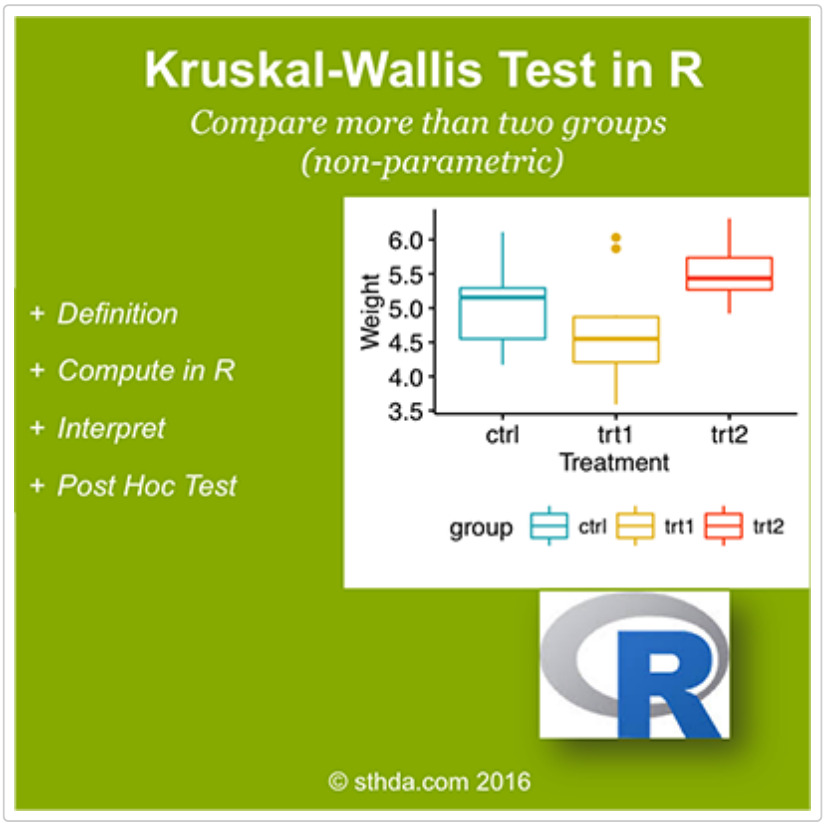
- What is MANOVA test?
- Assumptions of MANOVA
- Interpretation of MANOVA
- Compute MANOVA in R

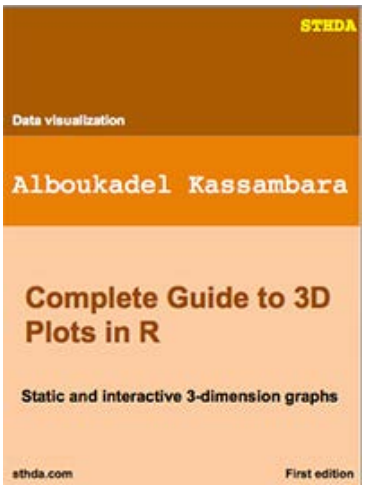


Read more: —> MANOVA Test in R: Multivariate Analysis of Variance.

7 Kruskal-Wallis test

- What is Kruskal-Wallis test?
- Visualize your data and compute Kruskal-Wallis test in R
  - Visualize the data using box plots
  - Compute Kruskal-Wallis test
  - Multiple pairwise-comparison between groups



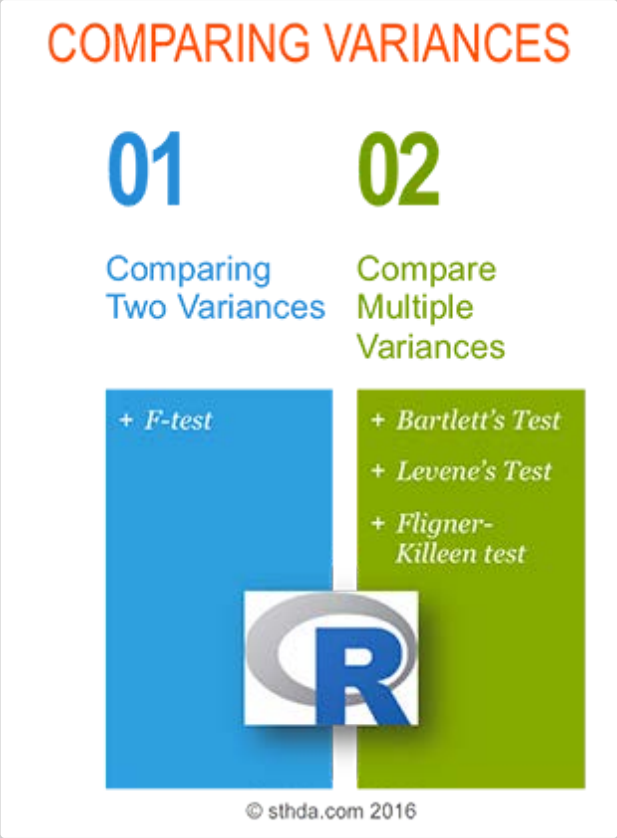


Comparing Variances in R

Previously, we described the essentials of R programming and provided quick start guides for importing data into **R**. Additionally, we described how to compute descriptive or summary statistics, correlation analysis, as well as, how to compare sample means using R software.

This chapter contains articles describing **statistical tests** to use for **comparing variances**.

1 How this chapter is organized?



2 F-Test: Compare two variances in R

- What is F-test?
- When to you use F-test?
- Research questions and statistical hypotheses
- Formula of F-test
- Compute F-test in R

Read more: —> F-Test: Compare Two Variances in R.

3 Compare multiple sample variances in R

This article describes **statistical tests** for comparing the **variances** of two or more samples.

- Compute **Bartlett's test** in R
- Compute **Levene's test** in R
- Compute **Fligner-Killeen** test in R

Read more: —> Compare Multiple Sample Variances in R.

4 See also

- R Basics
- Import and Export Data using R
- Preparing and Reshaping Data in R for Easier Analyses
- Data Manipulation in R
- Data visualization
- Descriptive Statistics and Graphics
- Correlation Analyses in R
- Comparing Means in R

5 Infos

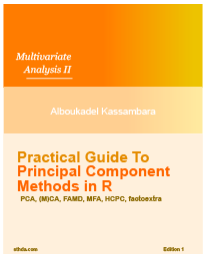
This analysis has been performed using **R statistical software** (ver. 3.2.4).

Enjoyed this article? I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

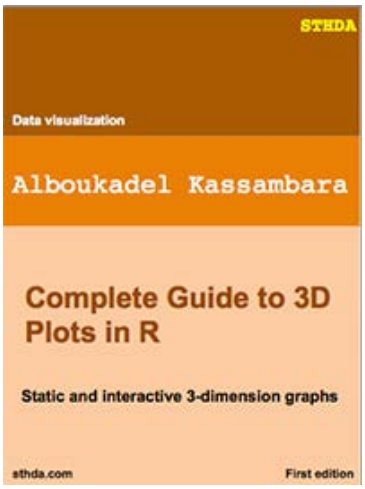
Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!

Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R



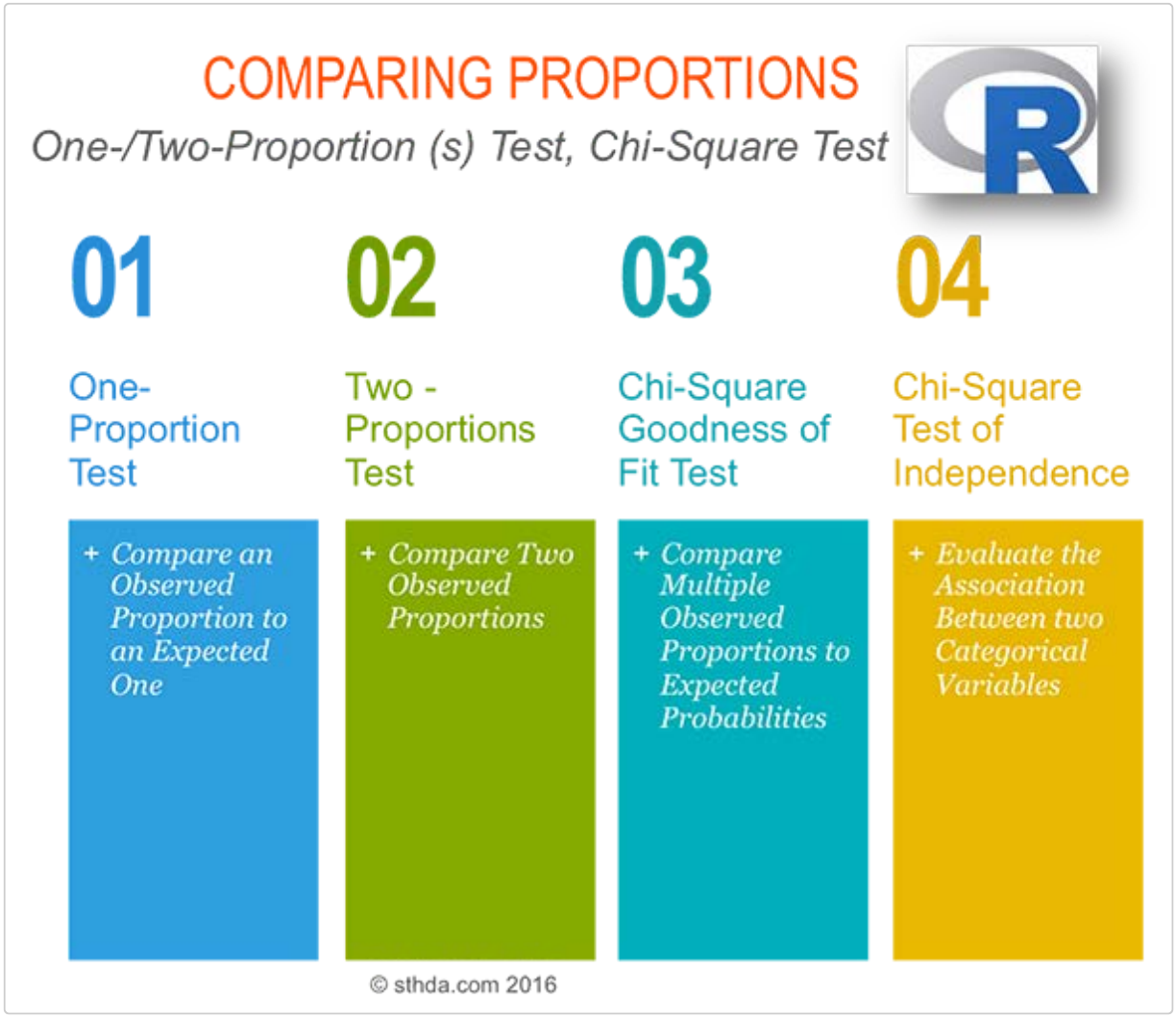
Comparing Proportions in R

Previously, we described the essentials of R programming and provided quick start guides for importing data into **R**. Additionally, we described how to compute descriptive or summary statistics, correlation analysis, as well as, how to compare sample means and variances using R software.

This chapter contains articles describing **statistical tests** to use for **comparing proportions**.

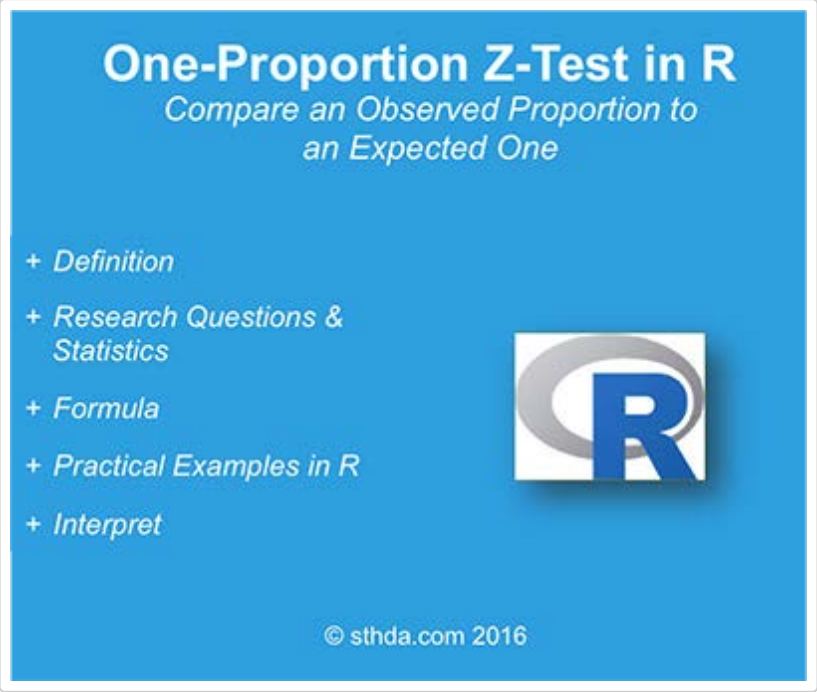
1 How this chapter is organized?

- One-Proportion Z-Test in R: Compare an Observed Proportion to an Expected One
- Two Proportions Z-Test in R: Compare Two Observed Proportions
- Chi-Square Goodness of Fit Test in R: Compare Multiple Observed Proportions to Expected Probabilities
- Chi-Square Test of Independence in R: Evaluate The Association Between Two Categorical Variables



2 One-proportion z-Test

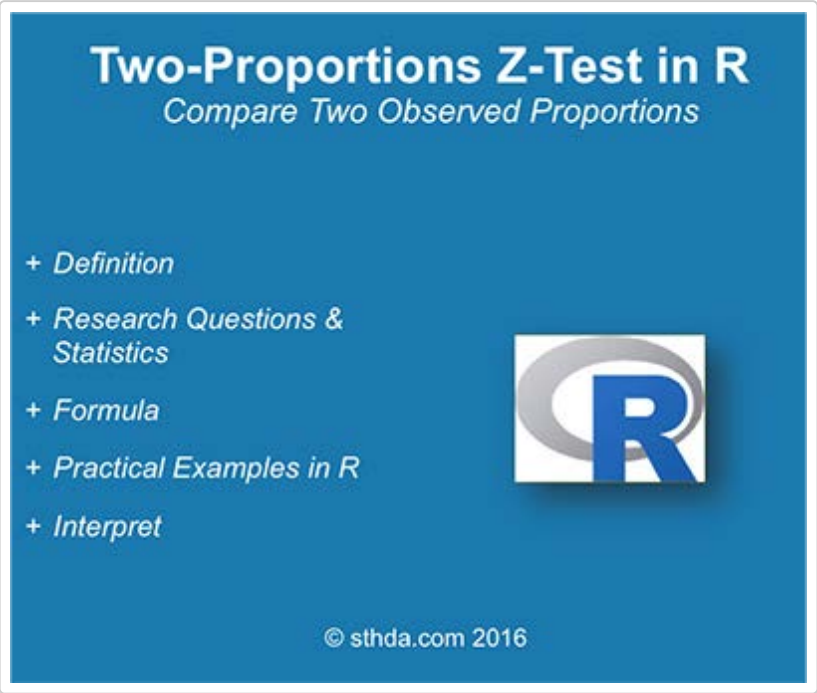
Compare an observed proportion to an expected one.



Read more: —> One-Proportion Z-Test in R.

3 Two-proportions z-Test

Compare two observed proportions.



Read more: —> Two Proportions Z-Test.

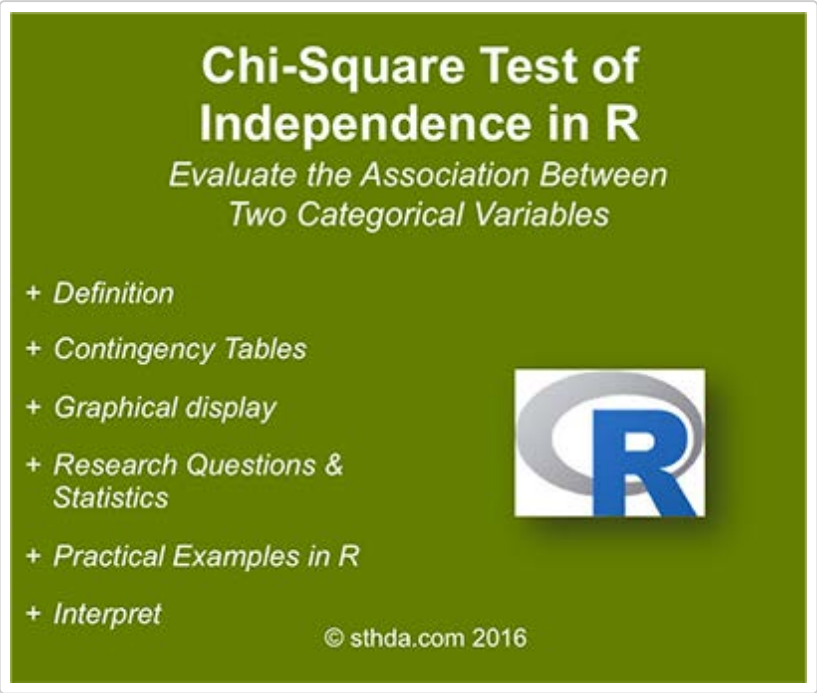
4 Chi-square goodness of fit test in R

Compare multiple observed proportions to expected probabilities.

Read more: —> Chi-square goodness of fit test in R.

5 Chi-Square test of independence in R

Evaluate the association between two categorical variables.



Read more: —> Chi-Square Test of Independence in R.

6 See also

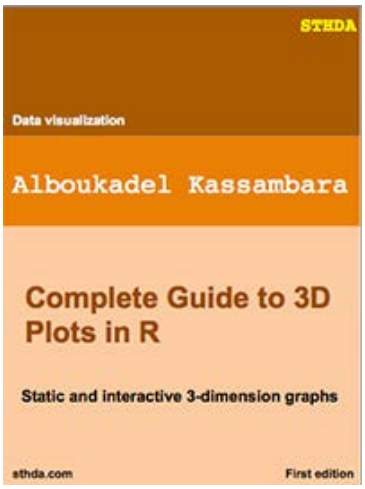
- R Basics
- Import and Export Data using R
- Preparing and Reshaping Data in R for Easier Analyses
- Data Manipulation in R

7 Infos

This analysis has been performed using **R statistical software** (ver. 3.2.4).

Enjoyed this article? I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!



Blogroll
- R-Bloggers

Actions menu for module Wiki
- Home
- Explorer

Cox Proportional-Hazards Model

Tools
- Discussion (9)

The **Cox proportional-hazards model** (Cox, 1972) is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables.

In the previous chapter (survival analysis basics), we described the basic concepts of survival analyses and methods for analyzing and summarizing survival data, including:

- the definition of hazard and survival functions,
- the construction of Kaplan-Meier survival curves for different patient groups
- the logrank test for comparing two or more survival curves

The above mentioned methods - Kaplan-Meier curves and logrank tests - are examples of *univariate analysis*. They describe the survival according to one factor under investigation, but ignore the impact of any others.

Additionally, Kaplan-Meier curves and logrank tests are useful only when the predictor variable is categorical (e.g.: treatment A vs treatment B; males vs females). They don't work easily for quantitative predictors such as gene expression, weight, or age.

An alternative method is the Cox proportional hazards regression analysis, which works for both quantitative predictor variables and for categorical variables. Furthermore, the Cox regression model extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time.

In this article, we'll describe the Cox regression model and provide practical examples using R software.

Contents

The need for multivariate statistical modeling

In clinical investigations, there are many situations, where several known quantities (known as *covariates*), potentially affect patient prognosis.

For instance, suppose two groups of patients are compared: those with and those without a specific genotype. If one of the groups also contains older individuals, any difference in survival may be attributable to genotype or age or indeed both. Hence, when investigating survival in relation to any one factor, it is often desirable to adjust for the impact of others.

Statistical model is a frequently used tool that allows to analyze survival with respect to several factors simultaneously. Additionally, statistical model provides the effect size for each factor.

The cox proportional-hazards model is one of the most important methods used for modelling survival analysis data. The next section introduces the basics of the Cox regression model.

Basics of the Cox proportional hazards model

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time. This rate is commonly referred as the hazard rate. Predictor variables (or factors) are usually termed *covariates* in the survival-analysis literature.

The Cox model is expressed by the *hazard function* denoted by h(t). Briefly, the hazard function can be interpreted as the risk of dying at time t. It can be estimated as follow:

$$h(t) = h_0(t) \times exp(b_1 x_1 + b_2 x_2 + ... + b_p x_p)$$

where,

- t represents the survival time
- $h(t)$ is the hazard function determined by a set of p covariates ($x_1, x_2, ..., x_p$)
- the coefficients ($b_1, b_2, ..., b_p$) measure the impact (i.e., the effect size) of covariates.
- the term h_0 is called the baseline hazard. It corresponds to the value of the hazard if all the x_i are equal to zero (the quantity exp(0) equals 1). The 't' in h(t) reminds us that the hazard may vary over time.

The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables x, with

the baseline hazard being an 'intercept' term that varies with time.

The quantities $exp(b_i)$ are called hazard ratios (HR). A value of b_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the i^{th} covariate increases, the event hazard increases and thus the length of survival decreases.

Put another way, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

In summary,

- HR = 1: No effect
- HR < 1: Reduction in the hazard
- HR > 1: Increase in Hazard

Note that in cancer studies:

- A covariate with hazard ratio > 1 (i.e.: b > 0) is called bad prognostic factor
- A covariate with hazard ratio < 1 (i.e.: b < 0) is called good prognostic factor

A key assumption of the Cox model is that the hazard curves for the groups of observations (or patients) should be proportional and cannot cross.

Consider two patients k and k' that differ in their x-values. The corresponding hazard function can be simply written as follow

- Hazard function for the patient k:

$$h_k(t) = h_0(t)e^{\sum_{i=1}^{n} \beta x}$$

- Hazard function for the patient k':

$$h_{k'}(t) = h_0(t)e^{\sum_{i=1}^{n} \beta x'}$$

- The hazard ratio for these two patients [$\frac{h_k(t)}{h_{k'}(t)} = \frac{h_0(t)e^{\sum_{i=1}^{n} \beta x}}{h_0(t)e^{\sum_{i=1}^{n} \beta x'}} = \frac{e^{\sum_{i=1}^{n} \beta x}}{e^{\sum_{i=1}^{n} \beta x'}}$] is independent of time t.

Consequently, the Cox model is a *proportional-hazards model*: the hazard of the event in any group is a constant multiple of the hazard in any other. This assumption implies that, as mentioned above, the hazard curves for the groups should be proportional and cannot cross.

In other words, if an individual has a risk of death at some initial time point that is twice as high as that of another individual, then at all later times the risk of death remains twice as high.

This assumption of proportional hazards should be tested. We'll discuss methods for assessing proportionality in the next article in this series: Cox Model Assumptions.

Compute the Cox model in R

Install and load required R package

We'll use two R packages:

- **survival** for computing survival analyses
- **survminer** for visualizing survival analysis results

- Install the packages

```
install.packages(c("survival", "survminer"))
```

- Load the packages

```
library("survival")
library("survminer")
```

R function to compute the Cox model: coxph()

The function *coxph*()[in *survival* package] can be used to compute the Cox proportional hazards regression model in R.

The simplified format is as follow:

```
coxph(formula, data, method)
```

- formula: is linear model with a survival object as the response variable. Survival object is created using the function *Surv*() as follow: *Surv(time, event)*.
- data: a data frame containing the variables
- method: is used to specify how to handle ties. The default is 'efron'. Other options are 'breslow' and 'exact'. The default 'efron' is generally preferred to the once-popular "breslow" method. The "exact" method is much more computationally intensive.

Example data sets

We'll use the lung cancer data in the survival R package.

```
data("lung")
head(lung)
```

```
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74   1       1       90       100     1175      NA
2    3  455      2  68   1       0       90        90     1225      15
3    3 1010      1  56   1       0       90        90       NA      15
4    5  210      2  57   1       1       90        60     1150      11
5    1  883      2  60   1       0      100        90       NA       0
6   12 1022      1  74   1       1       50        80      513       0
```

- inst: Institution code
- time: Survival time in days
- status: censoring status 1=censored, 2=dead
- age: Age in years
- sex: Male=1 Female=2
- ph.ecog: ECOG performance score (0=good 5=dead)
- ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
- pat.karno: Karnofsky performance score as rated by patient

- meal.cal: Calories consumed at meals
- wt.loss: Weight loss in last six months

Compute the Cox model

We'll fit the Cox regression using the following covariates: age, sex, ph.ecog and wt.loss.

We start by computing univariate Cox analyses for all these variables; then we'll fit multivariate cox analyses using two variables to describe how the factors jointly impact on survival.

Univariate Cox regression

Univariate Cox analyses can be computed as follow:

```
res.cox <- coxph(Surv(time, status) ~ sex, data = lung)
res.cox
```

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)
      coef exp(coef) se(coef)     z      p
sex -0.531     0.588    0.167 -3.18 0.0015
Likelihood ratio test=10.6  on 1 df, p=0.00111
n= 228, number of events= 165
```

The function *summary*() for Cox models produces a more complete report:

```
summary(res.cox)
```

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)
  n= 228, number of events= 165
       coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    exp(coef) exp(-coef) lower .95 upper .95
sex     0.588      1.701    0.4237     0.816
Concordance= 0.579  (se = 0.022 )
Rsquare= 0.046   (max possible= 0.999 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001111
Wald test            = 10.09  on 1 df,   p=0.001491
Score (logrank) test = 10.33  on 1 df,   p=0.001312
```

The Cox regression results can be interpreted as follow:

1. *Statistical significance*. The column marked "z" gives the Wald statistic value. It corresponds to the ratio of each regression coefficient to its standard error (z = coef/se(coef)). The wald statistic evaluates, whether the beta (β) coefficient of a given variable is statistically significantly different from 0. From the output above, we can conclude that the variable sex have highly statistically significant coefficients.
2. *The regression coefficients*. The second feature to note in the Cox model results is the the sign of the regression coefficients (coef). A positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The variable sex is encoded as a numeric vector. 1: male, 2: female. The R summary for the Cox model gives the hazard ratio (HR) for the second group relative to the first group, that is, female versus male. The beta coefficient for sex = -0.53 indicates that females have lower risk of death (lower survival rates) than males, in these data.
3. *Hazard ratios*. The exponentiated coefficients (exp(coef) = exp(-0.53) = 0.59), also known as *hazard ratios*, give the effect size of covariates. For example, being female (sex=2) reduces the hazard by a factor of 0.59, or 41%. Being female is associated with good prognostic.

4. *Confidence intervals of the hazard ratios*. The summary output also gives upper and lower 95% confidence intervals for the hazard ratio (exp(coef)), lower 95% bound = 0.4237, upper 95% bound = 0.816.
5. *Global statistical significance of the model*. Finally, the output gives p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score logrank statistics. These three methods are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ somewhat. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred.

To apply the univariate coxph function to multiple covariates at once, type this:

```
covariates <- c("age", "sex",  "ph.karno", "ph.ecog", "wt.loss")
univ_formulas <- sapply(covariates,
                        function(x) as.formula(paste('Surv(time, status)~', x)))

univ_models <- lapply( univ_formulas, function(x){coxph(x, data = lung)})
# Extract data
univ_results <- lapply(univ_models,
                       function(x){
                          x <- summary(x)
                          p.value<-signif(x$wald["pvalue"], digits=2)
                          wald.test<-signif(x$wald["test"], digits=2)
                          beta<-signif(x$coef[1], digits=2);#coeficient beta
                          HR <-signif(x$coef[2], digits=2);#exp(beta)
                          HR.confint.lower <- signif(x$conf.int[,"lower .95"], 2)
                          HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)
                          HR <- paste0(HR, " (",
                                       HR.confint.lower, "-", HR.confint.upper, ")")
                          res<-c(beta, HR, wald.test, p.value)
                          names(res)<-c("beta", "HR (95% CI for HR)", "wald.test",
                                        "p.value")
                          return(res)
                          #return(exp(cbind(coef(x),confint(x))))
                         })
res <- t(as.data.frame(univ_results, check.names = FALSE))
as.data.frame(res)
```

```
           beta HR (95% CI for HR) wald.test p.value
age       0.019           1 (1-1)       4.1   0.042
sex       -0.53    0.59 (0.42-0.82)        10  0.0015
ph.karno -0.016      0.98 (0.97-1)       7.9   0.005
ph.ecog    0.48        1.6 (1.3-2)        18 2.7e-05
wt.loss  0.0013         1 (0.99-1)      0.05    0.83
```

The output above shows the regression beta coefficients, the effect sizes (given as hazard ratios) and statistical significance for each of the variables in relation to overall survival. Each factor is assessed through separate univariate Cox regressions.

From the output above,

- The variables sex, age and ph.ecog have highly statistically significant coefficients, while the coefficient for ph.karno is not significant.
- age and ph.ecog have positive beta coefficients, while sex has a negative coefficient. Thus, older age and higher ph.ecog are associated with poorer survival, whereas being female (sex=2) is associated with better survival.

Now, we want to describe how the factors jointly impact on survival. To answer to this question, we'll perform a multivariate Cox regression analysis. As the variable ph.karno is not significant in the univariate Cox analysis, we'll skip it in the multivariate analysis. We'll include the 3 factors (sex, age and ph.ecog) into the multivariate model.

Multivariate Cox regression analysis

A Cox regression of time to death on the time-constant covariates is specified as follow:

```
res.cox <- coxph(Surv(time, status) ~ age + sex + ph.ecog, data =  lung)
summary(res.cox)
```

```
Call:
coxph(formula = Surv(time, status) ~ age + sex + ph.ecog, data = lung)
  n= 227, number of events= 164
   (1 observation deleted due to missingness)
             coef exp(coef)  se(coef)      z Pr(>|z|)
age      0.011067  1.011128  0.009267  1.194 0.232416
sex     -0.552612  0.575445  0.167739 -3.294 0.000986 ***
ph.ecog  0.463728  1.589991  0.113577  4.083 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
        exp(coef) exp(-coef) lower .95 upper .95
age        1.0111     0.9890    0.9929    1.0297
sex        0.5754     1.7378    0.4142    0.7994
ph.ecog    1.5900     0.6289    1.2727    1.9864
Concordance= 0.637  (se = 0.026 )
Rsquare= 0.126   (max possible= 0.999 )
Likelihood ratio test= 30.5  on 3 df,   p=1.083e-06
Wald test            = 29.93  on 3 df,   p=1.428e-06
Score (logrank) test = 30.5  on 3 df,   p=1.083e-06
```

The p-value for all three overall tests (likelihood, Wald, and score) are significant, indicating that the model is significant. These tests evaluate the omnibus null hypothesis that all of the betas (β) are 0. In the above example, the test statistics are in close agreement, and the omnibus null hypothesis is soundly rejected.

In the multivariate Cox analysis, the covariates sex and ph.ecog remain significant ($p < 0.05$). However, the covariate age fails to be significant ($p = 0.23$, which is grater than 0.05).

The p-value for sex is 0.000986, with a hazard ratio HR = exp(coef) = 0.58, indicating a strong relationship between the patients' sex and decreased risk of death. The hazard ratios of covariates are interpretable as multiplicative effects on the hazard. For example, holding the other covariates constant, being female (sex=2) reduces the hazard by a factor of 0.58, or 42%. We conclude that, being female is associated with good prognostic.

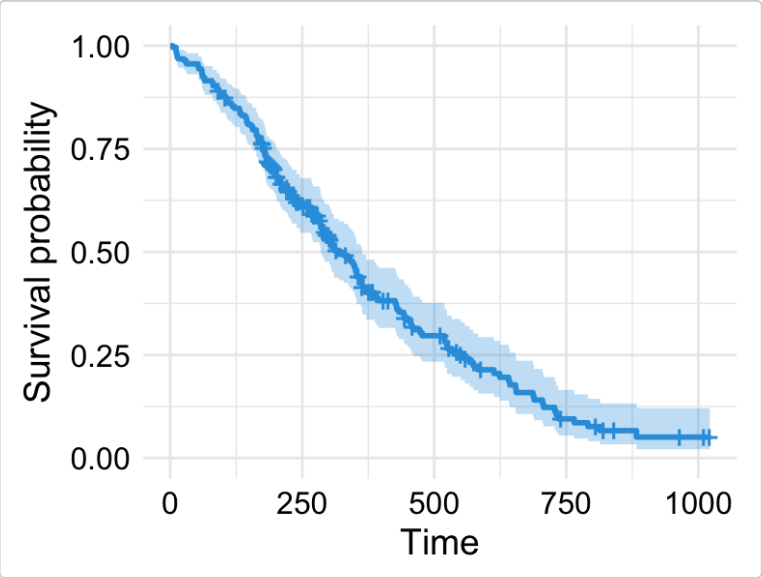
Similarly, the p-value for ph.ecog is 4.45e-05, with a hazard ratio HR = 1.59, indicating a strong relationship between the ph.ecog value and increased risk of death. Holding the other covariates constant, a higher value of ph.ecog is associated with a poor survival.

By contrast, the p-value for age is now p=0.23. The hazard ratio HR = exp(coef) = 1.01, with a 95% confidence interval of 0.99 to 1.03. Because the confidence interval for HR includes 1, these results indicate that age makes a smaller contribution to the difference in the HR after adjusting for the ph.ecog values and patient's sex, and only trend toward significance. For example, holding the other covariates constant, an additional year of age induce daily hazard of death by a factor of exp(beta) = 1.01, or 1%, which is not a significant contribution.

Visualizing the estimated distribution of survival times

Having fit a Cox model to the data, it's possible to visualize the predicted survival proportion at any given point in time for a particular risk group. The function *survfit*() estimates the survival proportion, by default at the mean values of covariates.

```
# Plot the baseline survival function
ggsurvplot(survfit(res.cox), color = "#2E9FDF",
           ggtheme = theme_minimal())
```



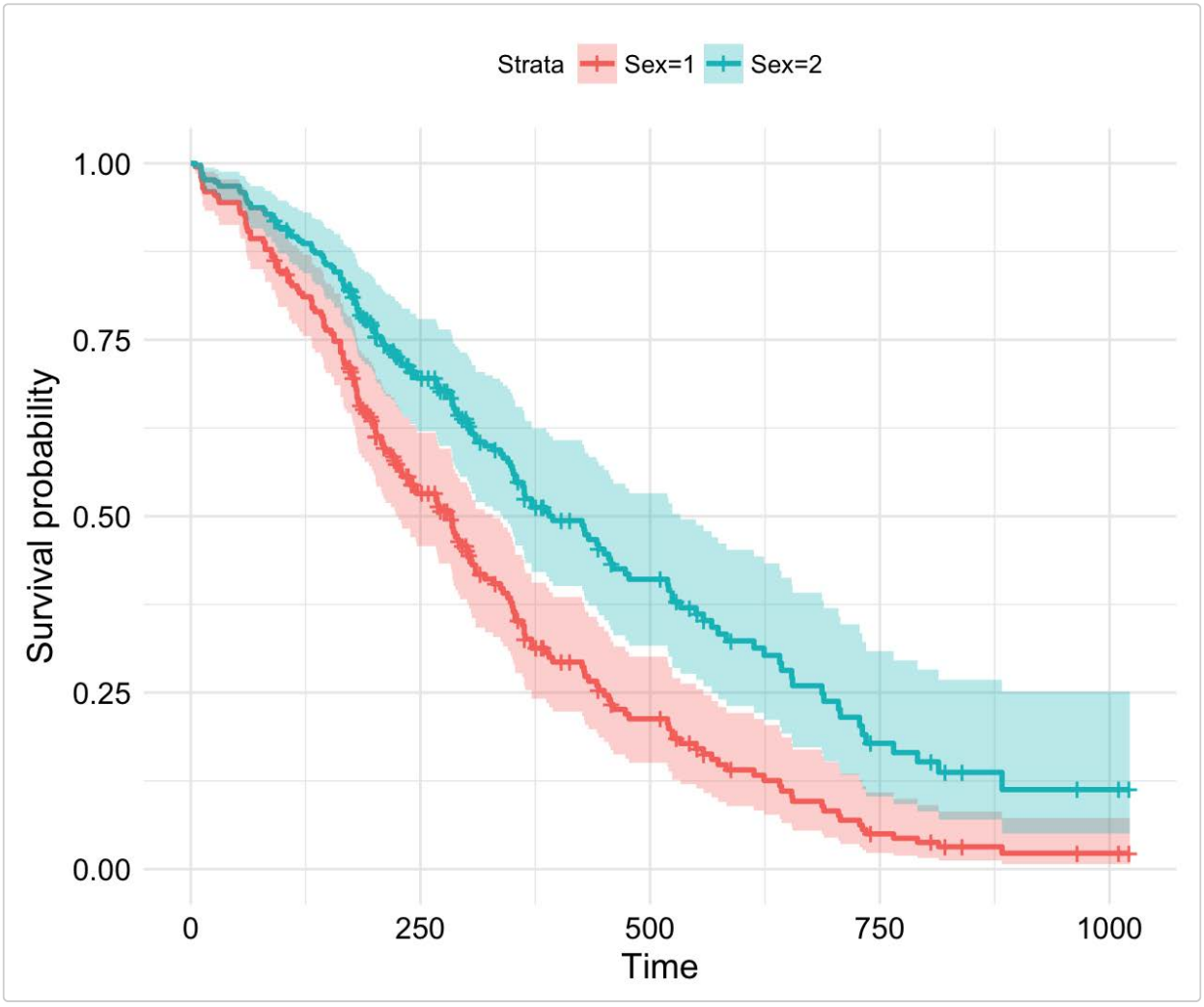
We may wish to display how estimated survival depends upon the value of a covariate of interest.

Consider that, we want to assess the impact of the sex on the estimated survival probability. In this case, we construct a new data frame with two rows, one for each value of sex; the other covariates are fixed to their average values (if they are continuous variables) or to their lowest level (if they are discrete variables). For a dummy covariate, the average value is the proportion coded 1 in the data set. This data frame is passed to *survfit*() via the *newdata* argument:

```
# Create the new data
sex_df <- with(lung,
               data.frame(sex = c(1, 2),
                          age = rep(mean(age, na.rm = TRUE), 2),
                          ph.ecog = c(1, 1)
                          )
               )
sex_df
```

```
  sex      age ph.ecog
1   1 62.44737       1
2   2 62.44737       1
```

```
# Survival curves
fit <- survfit(res.cox, newdata = sex_df)
ggsurvplot(fit, conf.int = TRUE, legend.labs=c("Sex=1", "Sex=2"),
           ggtheme = theme_minimal())
```



Summary

In this article, we described the Cox regression model for assessing simultaneously the relationship between multiple risk factors and patient's survival time. We demonstrated how to compute the Cox model using the *survival* package. Additionally, we described how to visualize the results of the analysis using the *survminer* package.

References

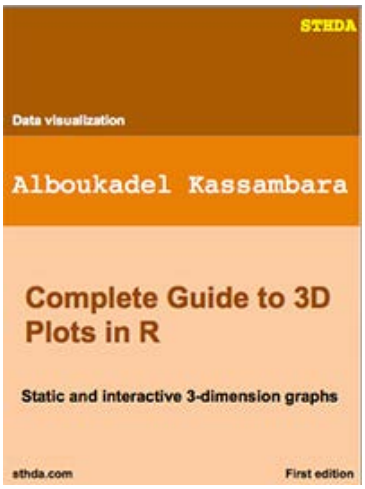
- Cox DR (1972). Regression models and life tables (with discussion). J R Statist Soc B 34: 187–220
- MJ Bradburn, TG Clark, SB Love and DG Altman. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. British Journal of Cancer (2003) 89, 431 – 436

Infos

This analysis has been performed using **R software** (ver. 3.3.2).

Enjoyed this article? I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!



Cox Model Assumptions

Previously, we described the basic methods for analyzing survival data, as well as, the Cox proportional hazards methods to deal with the situation where several factors impact on the survival process.

In the current article, we continue the series by describing methods to evaluate the validity of the **Cox model assumptions**.

Note that, when used inappropriately, statistical models may give rise to misleading conclusions. Therefore, it’s important to check that a given model is an appropriate representation of the data.

Contents

Diagnostics for the Cox model

The Cox proportional hazards model makes sevral assumptions. Thus, it is important to assess whether a fitted Cox regression model adequately describes the data.

Here, we'll disscuss three types of diagonostics for the Cox model:

- Testing the proportional hazards assumption.
- Examining influential observations (or outliers).
- Detecting nonlinearity in relationship between the log hazard and the covariates.

In order to check these model assumptions, *Residuals* method are used. The common residuals for the Cox model include:

- *Schoenfeld residuals* to check the proportional hazards assumption
- *Martingale residual* to assess nonlinearity
- *Deviance residual* (symmetric transformation of the Martinguale residuals), to examine influential observations

Assessing the validy of a Cox model in R

Installing and loading required R packages

We'll use two R packages:

- **survival** for computing survival analyses
- **survminer** for visualizing survival analysis results

- Install the packages

```
install.packages(c("survival", "survminer"))
```

- Load the packages

```
library("survival")
library("survminer")
```

Computing a Cox model

We'll use the lung data sets and the *coxph*() function in the survival package.

To compute a Cox model, type this:

```
library("survival")
res.cox <- coxph(Surv(time, status) ~ age + sex + wt.loss, data =  lung)
res.cox
```

```
Call:
coxph(formula = Surv(time, status) ~ age + sex + wt.loss, data = lung)
             coef exp(coef) se(coef)     z       p
age       0.02009   1.02029  0.00966  2.08 0.0377
sex      -0.52103   0.59391  0.17435 -2.99 0.0028
wt.loss   0.00076   1.00076  0.00619  0.12 0.9024
Likelihood ratio test=14.7  on 3 df, p=0.00212
n= 214, number of events= 152
   (14 observations deleted due to missingness)
```

Testing proportional Hazards assumption

The proportional hazards (PH) assumption can be checked using statistical tests and graphical diagnostics based on the *scaled Schoenfeld residuals*.

In principle, the *Schoenfeld residuals* are independent of time. A plot that shows a non-random pattern against time is evidence of violation of the PH assumption.

The function *cox.zph*() [in the *survival* package] provides a convenient solution to test the proportional hazards assumption for each covariate included in a Cox refression model fit.

For each covariate, the function *cox.zph*() correlates the corresponding set of scaled Schoenfeld residuals with time, to test for independence between residuals and time. Additionally, it performs a global test for the model as a whole.

The proportional hazard assumption is supported by a non-significant relationship between residuals and time, and refuted by a significant relationship.

To illustrate the test, we start by computing a Cox regression model using the lung data set [in survival package]:

```
library("survival")
res.cox <- coxph(Surv(time, status) ~ age + sex + wt.loss, data =  lung)
res.cox
```

```
Call:
coxph(formula = Surv(time, status) ~ age + sex + wt.loss, data = lung)
             coef exp(coef) se(coef)     z       p
age       0.02009   1.02029  0.00966  2.08 0.0377
sex      -0.52103   0.59391  0.17435 -2.99 0.0028
wt.loss   0.00076   1.00076  0.00619  0.12 0.9024
Likelihood ratio test=14.7  on 3 df, p=0.00212
n= 214, number of events= 152
   (14 observations deleted due to missingness)
```

To test for the proportional-hazards (PH) assumption, type this:

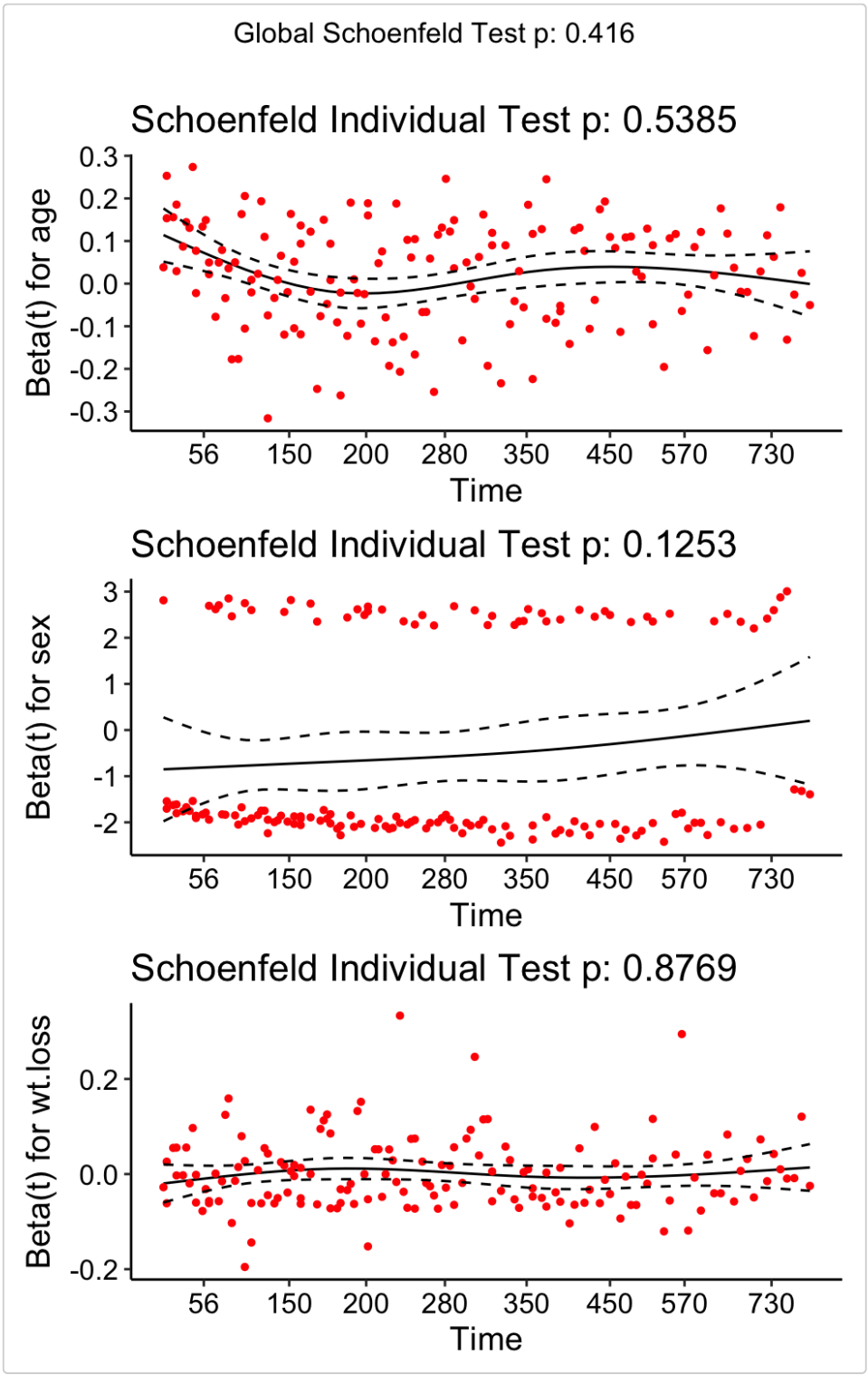
```
test.ph <- cox.zph(res.cox)
test.ph
```

```
            rho chisq     p
age     -0.0483 0.378 0.538
sex      0.1265 2.349 0.125
wt.loss  0.0126 0.024 0.877
GLOBAL       NA 2.846 0.416
```

From the output above, the test is not statistically significant for each of the covariates, and the global test is also not statistically significant. Therefore, we can assume the proportional hazards.

It’s possible to do a graphical diagnostic using the function *ggcoxzph*() [in the *survminer* package], which produces, for each covariate, graphs of the scaled Schoenfeld residuals against the transformed time.

```
ggcoxzph(test.ph)
```



In the figure above, the solid line is a smoothing spline fit to the plot, with the dashed lines representing a +/- 2-standard-error band around the fit.

Note that, systematic departures from a horizontal line are indicative of non-proportional hazards, since proportional hazards assumes that estimates $\beta_1, \beta_2, \beta_3$ do not vary much over time.

From the graphical inspection, there is no pattern with time. The assumption of proportional hazards appears to be supported for the covariates sex (which is, recall, a two-level factor, accounting for the two bands in the graph), wt.loss and age.

Another graphical methods for checking proportional hazards is to plot log(-log(S(t))) vs. t or log(t) and look for parallelism. This can be done only for categorical covariates.

A violations of proportional hazards assumption can be resolved by:

- Adding covariate*time interaction
- Stratification

Stratification is usefull for “nuisance” confounders, where you do not care to estimate the effect. You cannot examine the effects of the stratification variable (John Fox & Sanford Weisberg).

To read more about how to accomodate with non-proportional hazards, read the following articles:

- Jadwiga Borucka, PAREXEL, Warsaw, Poland. Extensions of cox model for non-proportional hazards purpose. 2013.
- John Fox & Sanford Weisberg. Cox Proportional-Hazards Regression for Survival Data in R.
- Max Gordon. Dealing with non-proportional hazards in R. March 29, 2016.

Testing influential observations

To test influential observations or outliers, we can visualize either:

- the *deviance residuals* or
- the *dfbeta* values

The function *ggcoxdiagnostics*()[in *survminer* package] provides a convenient solution for checkind influential observations. The simplified format is as follow:

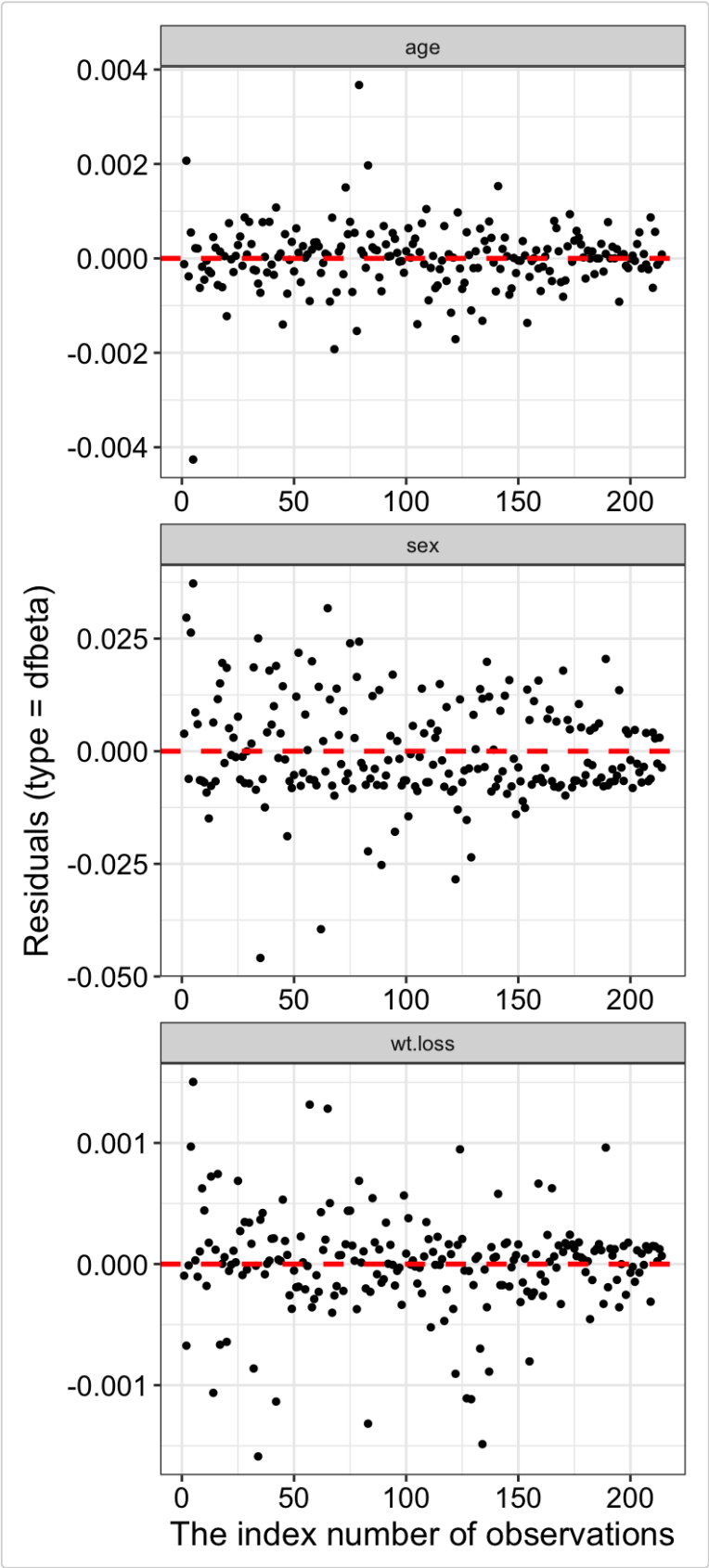
```
ggcoxdiagnostics(fit, type = , linear.predictions = TRUE)
```

- fit: an object of class coxph.object
- type: the type of residuals to present on Y axis. Allowed values include one of c(“martingale”, “deviance”, “score”, “schoenfeld”, “dfbeta”, “dfbetas”, “scaledsch”, “partial”).
- linear.predictions: a logical value indicating whether to show linear predictions for observations (TRUE) or just indexed of observations (FALSE) on X axis.

Specifying the argument *type = “dfbeta”*, plots the estimated changes in the regression coefficients upon deleting each observation in turn; likewise, *type=“dfbetas”* produces the estimated changes in the coefficients divided by their standard errors.

For example:

```
ggcoxdiagnostics(res.cox, type = "dfbeta",
                 linear.predictions = FALSE, ggtheme = theme_bw())
```



(Index plots of dfbeta for the Cox regression of time to death on age, sex and wt.loss)

The above index plots show that comparing the magnitudes of the largest dfbeta values to the regression coefficients suggests that none of the observations is terribly influential individually, even though some of the dfbeta values for age and wt.loss are large compared with the others.

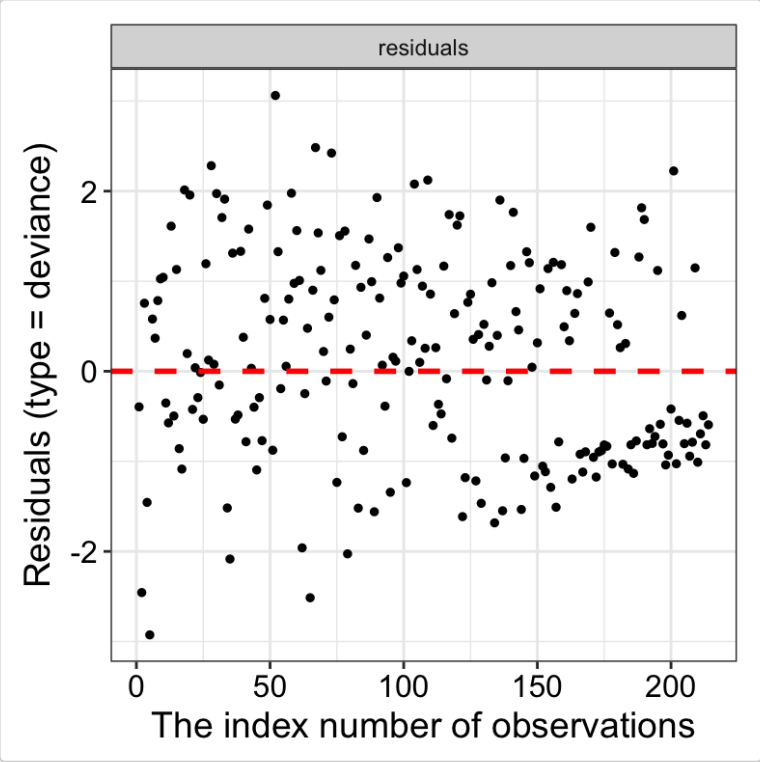
It’s also possible to check outliers by visualizing the deviance residuals. The deviance residual is a normalized transform of the martingale residual. These residuals should be roughtly symmetrically distributed about zero with a standard deviation of 1.

- Positive values correspond to individuals that “died too soon” compared to expected survival times.

Negative values correspond to individual that “lived too long”.
- Very large or small values are outliers, which are poorly predicted by the model.

Example of deviance residuals:

```
ggcoxdiagnostics(res.cox, type = "deviance",
                 linear.predictions = FALSE, ggtheme = theme_bw())
```



The pattern looks fairly symmetric around 0.

Testing non linearity

Often, we assume that continuous covariates have a linear form. However, this assumption should be checked.

Plotting the *Martingale residuals* against continuous covariates is a common approach used to detect *nonlinearity* or, in other words, to assess the functional form of a covariate. For a given continuous covariate, patterns in the plot may suggest that the variable is not properly fit.

Nonlinearity is not an issue for categorical variables, so we only examine plots of martingale residuals and partial residuals against a continuous variable.

Martingale residuals may present any value in the range (-INF, +1):

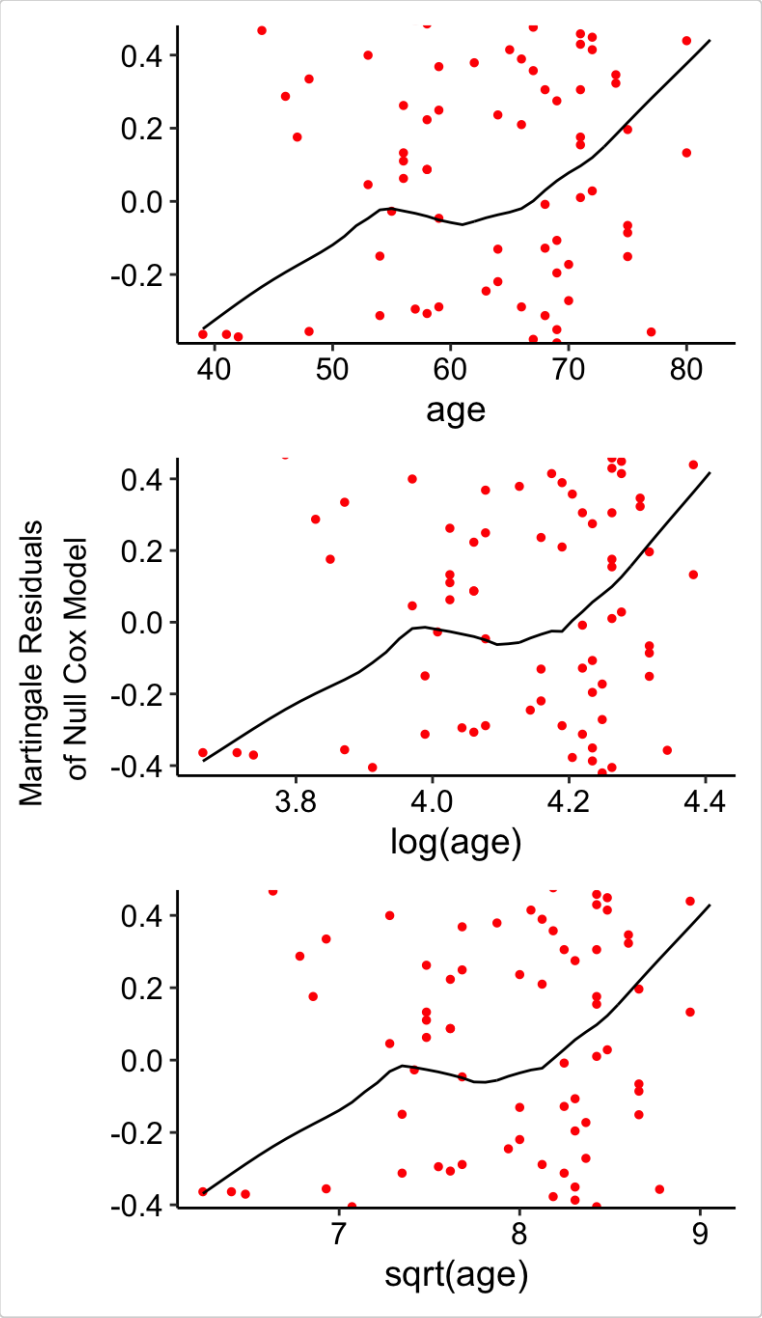
- A value of martinguale residuals near 1 represents individuals that “died too soon”,
- and large negative values correspond to individuals that “lived too long”.

To assess the functional form of a continuous variable in a Cox proportional hazards model, we’ll use the function *ggcoxfunctional*() [in the *survminer* R package].

The function *ggcoxfunctional*() displays graphs of continuous covariates against martingale residuals of null cox proportional hazards model. This might help to properly choose the functional form of continuous variable in the Cox model. Fitted lines with lowess function should be linear to satisfy the Cox proportional hazards model assumptions.

For example, to assess the functional forme of age, type this:

```
ggcoxfunctional(Surv(time, status) ~ age + log(age) + sqrt(age), data = lung)
```



It appears that, nonlinearity is slightly here.

Summary

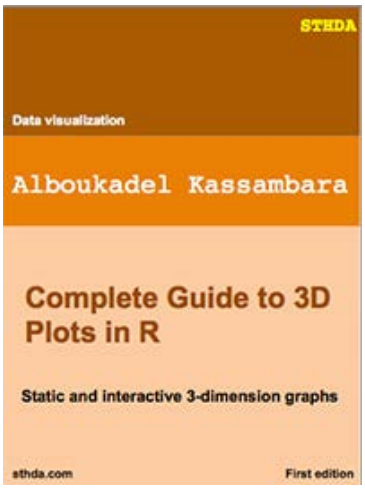
We described how to assess the valididy of the Cox model assumptions using the survival and survminer packages.

Infos

This analysis has been performed using **R software** (ver. 3.3.2).

Enjoyed this article? I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!



Guest Book

This website is excellent, it's extremely useful. It boasts very good techniques, elegant code, and is well-written and organised. It's one of the best of its kind.

By *Zahra H*

Guest Book

Blogroll

- R-Bloggers

1. Home
2. Articles
3. Cluster Analysis in R: Practical Guide
4. Clustering Basics
5. Data Preparation and Essential R Packages for Cluster Analysis

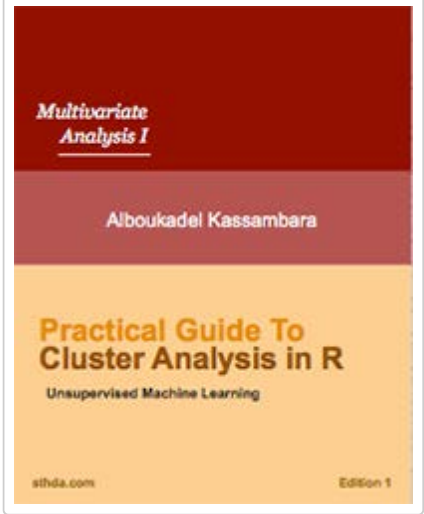
Articles - Clustering Basics

Data Preparation and Essential R Packages for Cluster Analysis

kassambara | 04/09/2017 | 4223 | Comments (6) | Clustering Basics

In this chapter, we start by presenting the data format and preparation for **cluster** analysis. Next, we introduce two main R packages - **cluster** and **factoextra** - for computing and visualizing clusters.

Related Books:



Practical Guide to Cluster Analysis in R

Data preparation

To perform a cluster analysis in R, generally, the data should be prepared as follow:

1. Rows are observations (individuals) and columns are variables
2. Any missing value in the data must be removed or estimated.
3. The data must be standardized (i.e., scaled) to make variables comparable. Recall that, standardization consists of transforming the variables such that they have mean zero and standard deviation one. Read more about data standardization in chapter @ref(clustering-distance-measures).

Here, we'll use the built-in R data set "USArrests", which contains statistics in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. It includes also the percent of the population living in urban areas.

```
data("USArrests")  # Load the data set
df <- USArrests    # Use df as shorter name
```

1. To remove any missing value that might be present in the data, type this:

```
df <- na.omit(df)
```

2. As we don't want the clustering algorithm to depend to an arbitrary variable unit, we start by scaling/standardizing the data using the R function *scale*():

```
df <- scale(df)
head(df, n = 3)
```

```
##          Murder Assault UrbanPop     Rape
## Alabama 1.2426   0.783   -0.521 -0.00342
## Alaska  0.5079   1.107   -1.212  2.48420
## Arizona 0.0716   1.479    0.999  1.04288
```

Required R Packages

In this book, we'll use mainly the following R packages:

- **cluster** for computing clustering algorithms, and
- **factoextra** for ggplot2-based elegant visualization of clustering results. The official online documentation is available at: http://www.sthda.com/english/rpkgs/factoextra.

factoextra contains many functions for cluster analysis and visualization, including:

Functions	Description
dist(fviz_dist, get_dist)	Distance Matrix Computation and Visualization
get_clust_tendency	Assessing Clustering Tendency
fviz_nbclust(fviz_gap_stat)	Determining the Optimal Number of Clusters
fviz_dend	Enhanced Visualization of Dendrogram
fviz_cluster	Visualize Clustering Results
fviz_mclust	Visualize Model-based Clustering Results
fviz_silhouette	Visualize Silhouette Information from Clustering
hcut	Computes Hierarchical Clustering and Cut the Tree
hkmeans	Hierarchical k-means clustering
eclust	Visual enhancement of clustering analysis

To install the two packages, type this:

```
install.packages(c("cluster", "factoextra"))
```

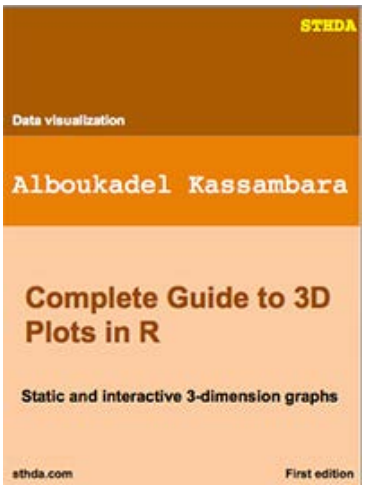
Last update : 24/09/2017
□ □ □ □ □ 0 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Comments

The fields marked with a * are required !



Articles - Clustering Basics

Clustering Distance Measures Essentials

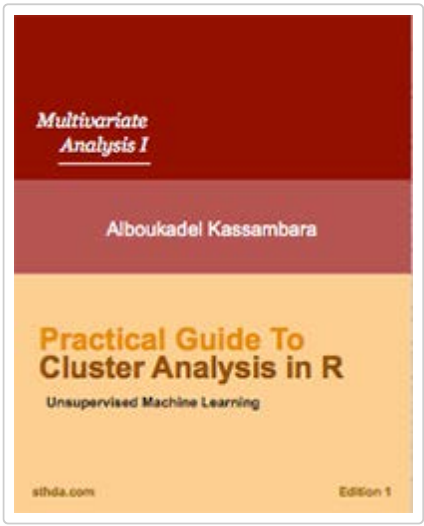
kassambara | 04/09/2017 | 12289 | Comments (2) | Clustering Basics | Unsupervised machine learning, Multivariate Analysis

The classification of observations into groups requires some methods for computing the **distance** or the (dis)**similarity** between each pair of observations. The result of this computation is known as a dissimilarity or **distance matrix**.

There are many methods to calculate this distance information. In this article, we describe the common **distance measures** and provide R codes for computing and visualizing distances.

Contents:

Related Books:



Practical Guide to Cluster Analysis in R

Methods for measuring distances

The choice of distance measures is a critical step in clustering. It defines how the similarity of two elements (x, y) is calculated and it will influence the shape of the clusters.

The classical methods for distance measures are *Euclidean* and *Manhattan distances*, which are defined as follow:

1. *Euclidean distance*:

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2. *Manhattan distance*:

$$d_{man}(x,y) = \sum_{i=1}^{n} |(x_i - y_i)|$$

Where, *x* and *y* are two vectors of length *n*.

Other dissimilarity measures exist such as **correlation-based distances**, which is widely used for gene expression data analyses. Correlation-based distance is defined by subtracting the correlation coefficient from 1. Different types of correlation methods can be used such as:

1. **Pearson correlation distance**:

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

Pearson correlation measures the degree of a linear relationship between two profiles.

2. **Eisen cosine correlation distance** (Eisen et al., 1998):

It's a special case of Pearson's correlation with $\bar{x}$ and $\bar{y}$ both replaced by zero:

$$| \quad n \quad |$$

$$d_{eisen}(x,y) = 1 - \frac{\left|\sum_{i=1} x_i y_i\right|}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

3. **Spearman correlation distance**:

The spearman correlation method computes the correlation between the rank of x and the rank of y variables.

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

Where $x_i' = rank(x_i)$ and $y_i' = rank(y)$.

4. **Kendall correlation distance**:

Kendall correlation method measures the correspondence between the ranking of x and y variables. The total number of possible pairings of x with y observations is $n(n-1)/2$, where n is the size of x and y. Begin by ordering the pairs by the x values. If x and y are correlated, then they would have the same relative rank orders. Now, for each y_i, count the number of $y_j > y_i$ (concordant pairs (c)) and the number of $y_j < y_i$ (discordant pairs (d)).

Kendall correlation distance is defined as follow:

$$d_{kend}(x,y) = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

Where,

- n_c: total number of concordant pairs
- n_d: total number of discordant pairs
- n: size of x and y

Note that,

- Pearson correlation analysis is the most commonly used method. It is also known as a parametric correlation which depends on the distribution of the data.
- Kendall and Spearman correlations are non-parametric and they are used to perform rank-based correlation analysis.

In the formula above, *x* and *y* are two vectors of length *n* and, means $\bar{x}$ and $\bar{y}$, respectively. The distance between x and y is denoted $d(x, y)$.

What type of distance measures should we choose?

The choice of distance measures is very important, as it has a strong influence on the clustering results. For most common clustering software, the default distance measure is the Euclidean distance.

Depending on the type of the data and the researcher questions, other dissimilarity measures might be preferred. For example, correlation-based distance is often used in gene expression data analysis.

Correlation-based distance considers two objects to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance. The distance between two objects is 0 when they are perfectly correlated. Pearson's correlation is quite sensitive to outliers. This does not matter when clustering samples, because the correlation is over thousands of genes. When clustering genes, it is important to be aware of the possible impact of outliers. This can be mitigated by using Spearman's correlation instead of Pearson's correlation.

If we want to identify clusters of observations with the same overall profiles regardless of their magnitudes, then we should go with *correlation-based distance* as a dissimilarity measure. This is particularly the case in gene expression data analysis, where we might want to consider genes similar when they are “up” and “down” together. It is also the case, in marketing if we want to identify group of shoppers with the same preference in term of items, regardless of the volume of items they bought.

If Euclidean distance is chosen, then observations with high values of features will be clustered together. The same holds true for observations with low values of features.

Data standardization

The value of distance measures is intimately related to the scale on which measurements are made. Therefore, variables are often scaled (i.e. standardized) before measuring the inter-observation dissimilarities. This is particularly recommended when variables are measured in different scales (e.g: kilograms, kilometers, centimeters, …); otherwise, the dissimilarity measures obtained will be severely affected.

The goal is to make the variables comparable. Generally variables are scaled to have i) standard deviation one and ii) mean zero.

The standardization of data is an approach widely used in the context of gene expression data analysis before clustering. We might also want to scale the data when the mean and/or the standard deviation of variables are largely different.

When scaling variables, the data can be transformed as follow:

$$\frac{x_i - center(x)}{scale(x)}$$

Where $center(x)$ can be the mean or the median of x values, and $scale(x)$ can be the standard deviation (SD), the interquartile range, or the MAD (median absolute deviation).

The R base function *scale*() can be used to standardize the data. It takes a numeric matrix as an input and performs the scaling on the columns.

Standardization makes the four distance measure methods - Euclidean, Manhattan, Correlation and Eisen - more similar than they would be with non-transformed data.

Note that, when the data are standardized, there is a functional relationship between the Pearson correlation coefficient $r(x, y)$ and the Euclidean distance.

With some maths, the relationship can be defined as follow:

$$d_{euc}(x, y) = \sqrt{2m[1 - r(x, y)]}$$

Where x and y are two standardized m-vectors with zero mean and unit length.

Therefore, the result obtained with Pearson correlation measures and standardized Euclidean distances are comparable.

Distance matrix computation

Data preparation

We’ll use the USArrests data as demo data sets. We’ll use only a subset of the data by taking 15 random rows among the 50 rows in the data set. This is done by using the function *sample*(). Next, we standardize the data using the function *scale*():

```
# Subset of the data
set.seed(123)
ss <- sample(1:50, 15)   # Take 15 random rows
df <- USArrests[ss, ]    # Subset the 15 rows
df.scaled <- scale(df)   # Standardize the variables
```

R functions and packages

There are many R functions for computing distances between pairs of observations:

1. *dist*() R base function [*stats* package]: Accepts only numeric data as an input.
2. *get_dist*() function [*factoextra* package]: Accepts only numeric data as an input. Compared to the standard dist() function, it supports correlation-based distance measures including "pearson", "kendall" and "spearman" methods.
3. *daisy()* function [*cluster* package]: Able to handle other variable types (e.g. nominal, ordinal, (a)symmetric binary). In that case, the Gower's coefficient will be automatically used as the metric. It's one of the most popular measures of proximity for mixed data types. For more details, read the R documentation of the *daisy*() function (*?daisy*).

All these functions compute distances between rows of the data.

Computing euclidean distance

To compute Euclidean distance, you can use the R base *dist*() function, as follow:

```
dist.eucl <- dist(df.scaled, method = "euclidean")
```

Note that, allowed values for the option method include one of: "euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski".

To make it easier to see the distance information generated by the *dist*() function, you can reformat the distance vector into a matrix using the *as.matrix*() function.

```
# Reformat as a matrix
# Subset the first 3 columns and rows and Round the values
round(as.matrix(dist.eucl)[1:3, 1:3], 1)
```

```
##              Iowa Rhode Island Maryland
## Iowa          0.0          2.8      4.1
## Rhode Island  2.8          0.0      3.6
## Maryland      4.1          3.6      0.0
```

In this matrix, the value represent the distance between objects. The values on the diagonal of the matrix represent the distance between objects and themselves (which are zero).

In this data set, the columns are variables. Hence, if we want to compute pairwise distances between variables, we must start by transposing the data to have variables in the rows of the data set before using the *dist*() function. The function *t*() is used for transposing the data.

Computing correlation based distances

Correlation-based distances are commonly used in gene expression data analysis.

The function *get_dist*()[*factoextra* package] can be used to compute correlation-based distances. Correlation method can be either *pearson*, *spearman* or *kendall*.

```
# Compute
library("factoextra")
dist.cor <- get_dist(df.scaled, method = "pearson")
# Display a subset
round(as.matrix(dist.cor)[1:3, 1:3], 1)
```

```
##              Iowa Rhode Island Maryland
## Iowa          0.0          0.4      1.9
## Rhode Island  0.4          0.0      1.5
## Maryland      1.9          1.5      0.0
```

Computing distances for mixed data

The function *daisy*() [*cluster* package] provides a solution (*Gower's metric*) for computing the distance matrix, in the situation where the data contain no-numeric columns.

The R code below applies the *daisy*() function on *flower* data which contains *factor*, *ordered* and *numeric* variables:

```
library(cluster)
# Load data
data(flower)
head(flower, 3)
```

```
##   V1 V2 V3 V4 V5 V6  V7 V8
## 1  0  1  1  4  3 15  25 15
## 2  1  0  0  2  1  3 150 50
## 3  0  1  0  3  3  1 150 50
```

```
# Data structure
str(flower)
```

```
## 'data.frame':    18 obs. of  8 variables:
##  $ V1: Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 2 2 ...
##  $ V2: Factor w/ 2 levels "0","1": 2 1 2 1 2 2 1 1 2 2 ...
##  $ V3: Factor w/ 2 levels "0","1": 2 1 1 2 1 1 1 2 1 1 ...
##  $ V4: Factor w/ 5 levels "1","2","3","4",..: 4 2 3 4 5 4 4 2 3 5 ...
##  $ V5: Ord.factor w/ 3 levels "1"<"2"<"3": 3 1 3 2 2 3 3 2 1 2 ...
##  $ V6: Ord.factor w/ 18 levels "1"<"2"<"3"<"4"<..: 15 3 1 16 2 12 13 7 4 14 ...
##  $ V7: num  25 150 150 125 20 50 40 100 25 100 ...
##  $ V8: num  15 50 50 50 15 40 20 15 15 60 ...
```

```
# Distance matrix
dd <- daisy(flower)
round(as.matrix(dd)[1:3, 1:3], 2)
```

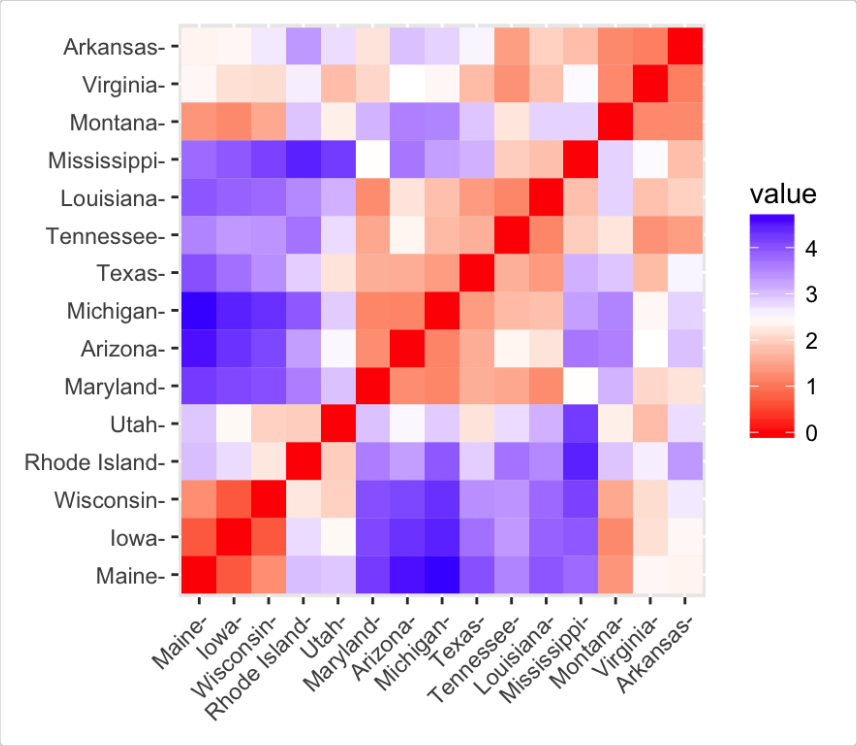
```
##      1    2    3
## 1 0.00 0.89 0.53
## 2 0.89 0.00 0.51
## 3 0.53 0.51 0.00
```

Visualizing distance matrices

A simple solution for visualizing the distance matrices is to use the function *fviz_dist*() [*factoextra* package]. Other specialized methods, such as agglomerative hierarchical clustering (Chapter @ref(agglomerative-clustering)) or heatmap (Chapter @ref(heatmap)) will be comprehensively described in the dedicated chapters.

To use *fviz_dist*() type this:

```
library(factoextra)
fviz_dist(dist.eucl)
```



- **Red**: high similarity (ie: low dissimilarity) | **Blue**: low similarity

The color level is proportional to the value of the dissimilarity between observations: pure red if $dist(x_i, x_j) = 0$ and pure blue corresponds to the highest value of euclidean distance computed. Objects belonging to the same cluster are displayed in consecutive order.

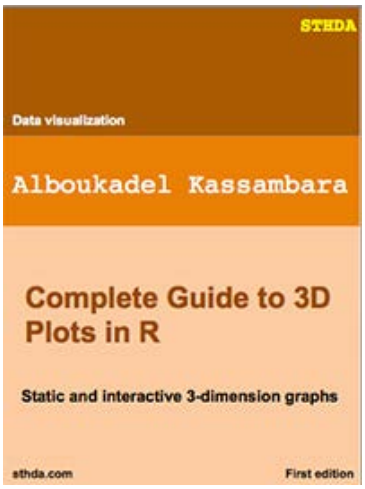
Summary

We described how to compute distance matrices using either Euclidean or correlation-based measures. It's generally recommended to standardize the variables before distance matrix computation. Standardization makes variable comparable, in the situation where they are measured in different scales.

Last update : 22/12/2017
□ □ □ □ □ 0 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!



Guest Book

Thanks Mr Kassambara for the great work and energy to share knowledge!! Really love your work!! Keep on going!!

By *Visitor*

Guest Book

Blogroll

- R-Bloggers

Actions menu for module Wiki

- Home
- Explorer

1. Home
2. Easy Guides
3. R software
4. R Basic Statistics
5. Normality Test in R

Normality Test in R

Tools

- Discussion (2)

- Install required R packages
- Load required R packages
- Import your data into R
- Check your data
- Assess the normality of the data in R
  - Case of large sample sizes
  - Visual methods
  - Normality test
- Infos

Many of statistical tests including correlation, regression, t-test, and analysis of variance (ANOVA) assume some certain characteristics about the data. They require the data to follow a **normal distribution** or **Gaussian distribution**. These tests are called **parametric tests**, because their validity depends on the distribution of the data.

Normality and the other assumptions made by these tests should be taken seriously to draw reliable interpretation and conclusions of the research.

Before using a parametric test, we should perform some **preleminary tests** to make sure that the test assumptions are met. In the situations where the assumptions are violated, **non-paramatric** tests are recommended.

Here, we'll describe how to check the normality of the data by visual inspection and by significance tests.

Install required R packages

1. **dplyr** for data manipulation

```
install.packages("dplyr")
```

2. ggpubr for an easy ggplot2-based data visualization

- Install the latest version from GitHub as follow:

```
# Install
if(!require(devtools)) install.packages("devtools")
devtools::install_github("kassambara/ggpubr")
```

- Or, install from CRAN as follow:

```
install.packages("ggpubr")
```

Load required R packages

```
library("dplyr")
library("ggpubr")
```

Import your data into R

1. **Prepare your data** as specified here: Best practices for preparing your data set for R
2. **Save your data** in an external .txt tab or .csv files
3. **Import your data into R** as follow:

```
# If .txt tab file, use this
my_data <- read.delim(file.choose())
# Or, if .csv file, use this
my_data <- read.csv(file.choose())
```

Here, we'll use the built-in R data set named ToothGrowth.

```
```

```
# Store the data in the variable my_data
my_data <- ToothGrowth
```

Check your data

We start by displaying a random sample of 10 rows using the function **sample_n**()[in **dplyr** package].

Show 10 random rows:

```
set.seed(1234)
dplyr::sample_n(my_data, 10)
```

```
    len supp dose
7  11.2   VC  0.5
37  8.2   OJ  0.5
36 10.0   OJ  0.5
58 27.3   OJ  2.0
49 14.5   OJ  1.0
57 26.4   OJ  2.0
1   4.2   VC  0.5
13 15.2   VC  1.0
35 14.5   OJ  0.5
27 26.7   VC  2.0
```

Assess the normality of the data in R

We want to test if the variable *len* (tooth length) is normally distributed.

Case of large sample sizes

If the sample size is large enough (n > 30), we can ignore the distribution of the data and use parametric tests.

The central limit theorem tells us that no matter what distribution things have, the sampling distribution tends to be normal if the sample is large enough (n > 30).

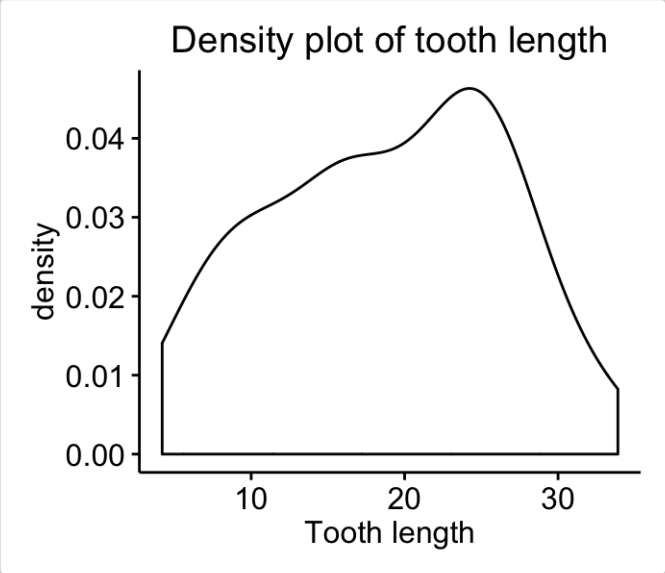
However, to be consistent, normality can be checked by visual inspection [**normal plots (histogram)**, **Q-Q plot** (quantile-quantile plot)] or by **significance tests**].

Visual methods

Density plot and **Q-Q plot** can be used to check normality visually.

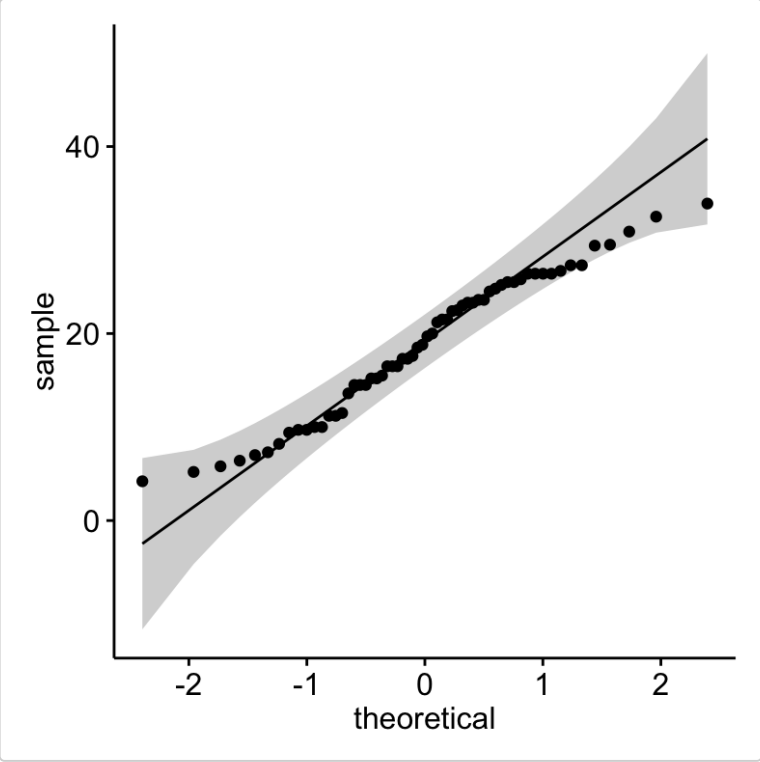
1. **Density plot**: the **density** plot provides a visual judgment about whether the distribution is bell shaped.

```
library("ggpubr")
ggdensity(my_data$len,
          main = "Density plot of tooth length",
          xlab = "Tooth length")
```



2. **Q-Q plot**: **Q-Q plot** (or quantile-quantile plot) draws the correlation between a given sample and the normal distribution. A 45-degree reference line is also plotted.

```
library(ggpubr)
ggqqplot(my_data$len)
```



It's also possible to use the function **qqPlot**() [in **car** package]:

```
library("car")
qqPlot(my_data$len)
```

As all the points fall approximately along this reference line, we can assume normality.

Normality test

Visual inspection, described in the previous section, is usually unreliable. It's possible to use a **significance test** comparing the sample distribution to a normal one in order to ascertain whether data show or not a serious deviation from normality.

There are several methods for **normality test** such as **Kolmogorov-Smirnov (K-S) normality test** and **Shapiro-Wilk's test**.

The null hypothesis of these tests is that "sample distribution is normal". If the test is **significant**, the distribution is non-normal.

Shapiro-Wilk's method is widely recommended for normality test and it provides better power than K-S. It is based on the correlation between the data and the corresponding normal scores.

Note that, normality test is sensitive to sample size. Small samples most often pass normality tests. Therefore, it's important to combine visual inspection and significance test in order to take the right decision.

The R function **shapiro.test**() can be used to perform the Shapiro-Wilk test of normality for one variable (univariate):

```
shapiro.test(my_data$len)
```

```
    Shapiro-Wilk normality test
data:  my_data$len
W = 0.96743, p-value = 0.1091
```

From the output, the p-value > 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

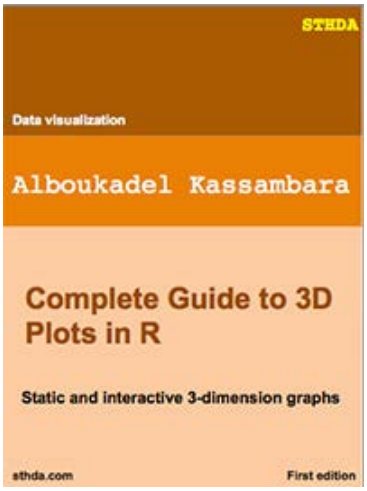
Infos

This analysis has been performed using **R software** (ver. 3.2.4).

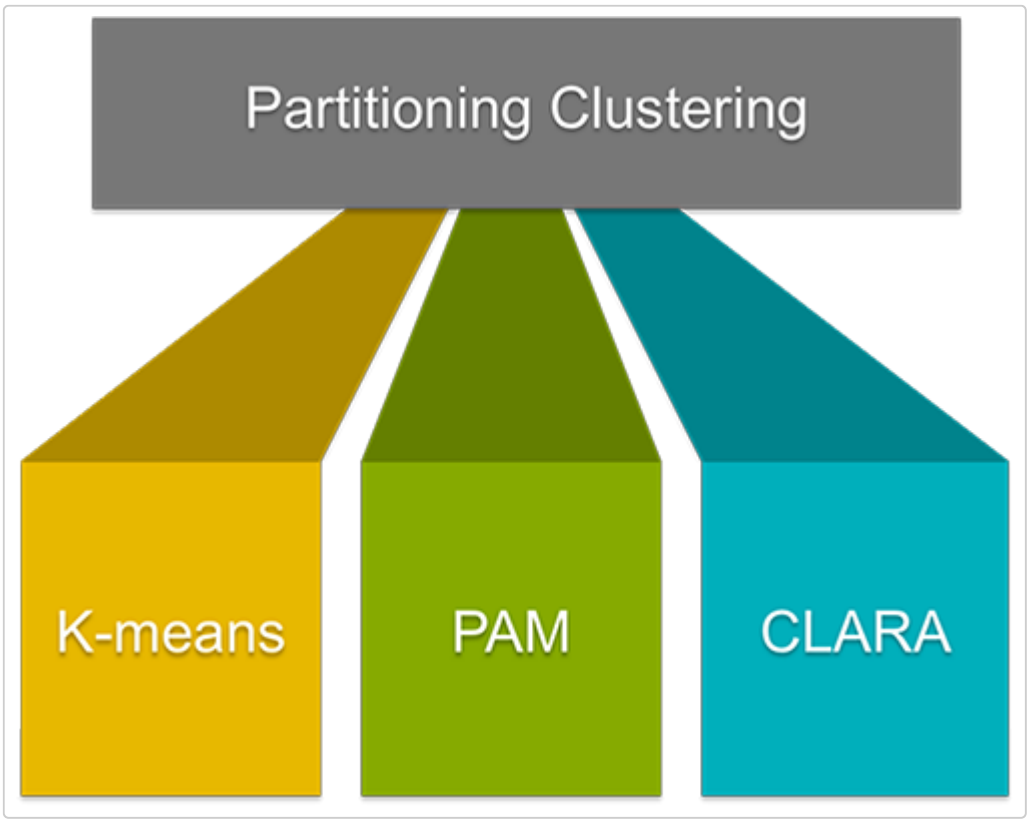
Enjoyed this article? I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!



Articles - Partitioning Clustering Essentials



Partitioning clustering are clustering methods used to classify observations, within a data set, into multiple groups based on their similarity. The algorithms require the analyst to specify the number of clusters to be generated.

This chapter describes the commonly used partitioning clustering, including:

- **K-means clustering** (MacQueen 1967), in which, each cluster is represented by the center or means of the data points belonging to the cluster. The K-means method is sensitive to anomalous data points and outliers.

- **K-medoids clustering** or **PAM** (*Partitioning Around Medoids*, Kaufman & Rousseeuw, 1990), in which, each cluster is represented by one of the objects in the cluster. PAM is less sensitive to outliers compared to k-means.
- **CLARA algorithm** (*Clustering Large Applications*), which is an extension to PAM adapted for large data sets.

For each of these methods, we provide:

- the basic idea and the key mathematical concepts
- the clustering algorithm and implementation in R software
- R lab sections with many examples for cluster analysis and visualization

The following R packages will be used to compute and visualize partitioning clustering:

- *stats* package for computing K-means
- *cluster* package for computing PAM and CLARA algorithms
- *factoextra* for beautiful visualization of clusters

Related articles

K-Means Clustering Essentials

Contents:

- K-means basic ideas
- K-means algorithm
- Computing k-means clustering in R
  - Data
  - Required R packages and functions: **stats::kmeans**()
  - Estimating the optimal number of clusters: **factoextra::fviz_nbclust**()
  - Computing k-means clustering
  - Accessing to the results of kmeans() function
  - Visualizing k-means clusters: **factoextra::fviz_cluster**()
- K-means clustering advantages and disadvantages
- Alternative to k-means clustering

K-Medoids Essentials: PAM clustering

Contents:

- PAM concept
- PAM algorithm
- Computing PAM in R
  - Data
  - Required R packages and functions: **cluster::pam**() or **fpc::pamk**()
  - Estimating the optimal number of clusters: **factoextra::fviz_nbclust**()
  - Computing PAM clustering
  - Accessing to the results of the pam() function
  - Visualizing PAM clusters: **factoextra::fviz_cluster**()

CLARA - Clustering Large Applications

Contents:

- CLARA concept
- CLARA Algorithm
- Computing CLARA in R
  - Data format and preparation
  - Required R packages and functions: **cluster::clara**()
  - Estimating the optimal number of clusters: **factoextra::fviz_nbclust**()
  - Computing CLARA
  - Visualizing CLARA clusters: **factoextra::fviz_cluster**()

Quick start

Data preparation:

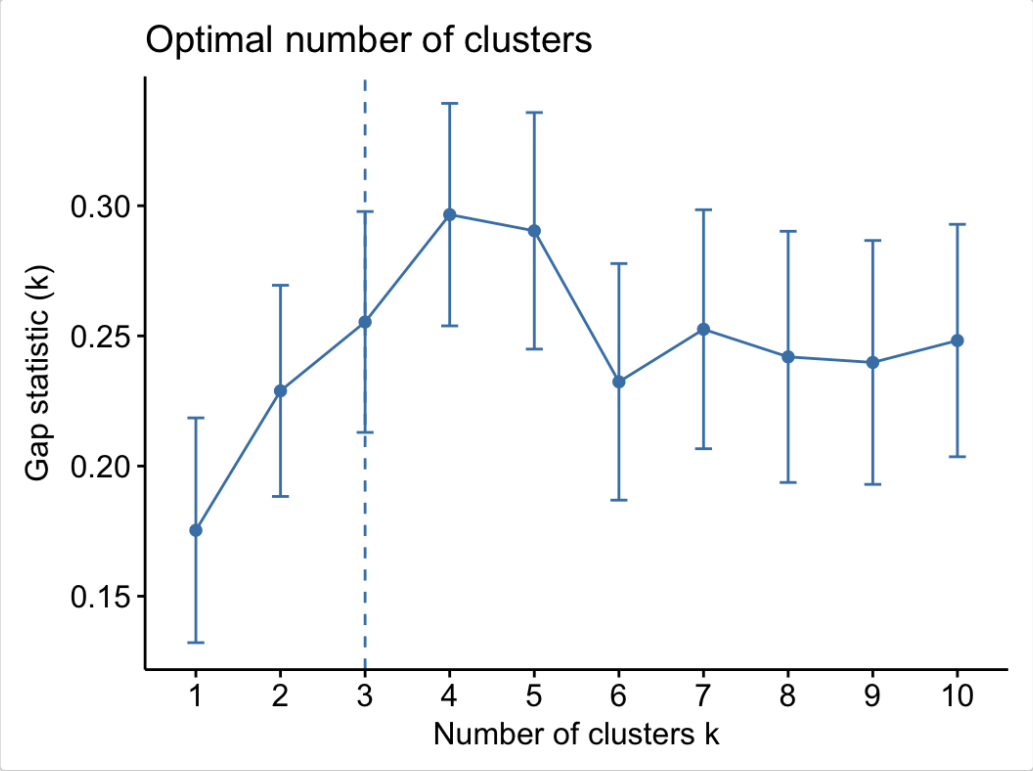
```
# Load data
data("USArrests")
my_data <- USArrests
# Remove any missing value (i.e, NA values for not available)
my_data <- na.omit(my_data)
# Scale variables
my_data <- scale(my_data)
# View the firt 3 rows
head(my_data, n = 3)
```

```
##            Murder Assault UrbanPop     Rape
## Alabama 1.2426   0.783   -0.521 -0.00342
## Alaska  0.5079   1.107   -1.212  2.48420
## Arizona 0.0716   1.479    0.999  1.04288
```

Determine the optimal number of clusters for k-means clustering:

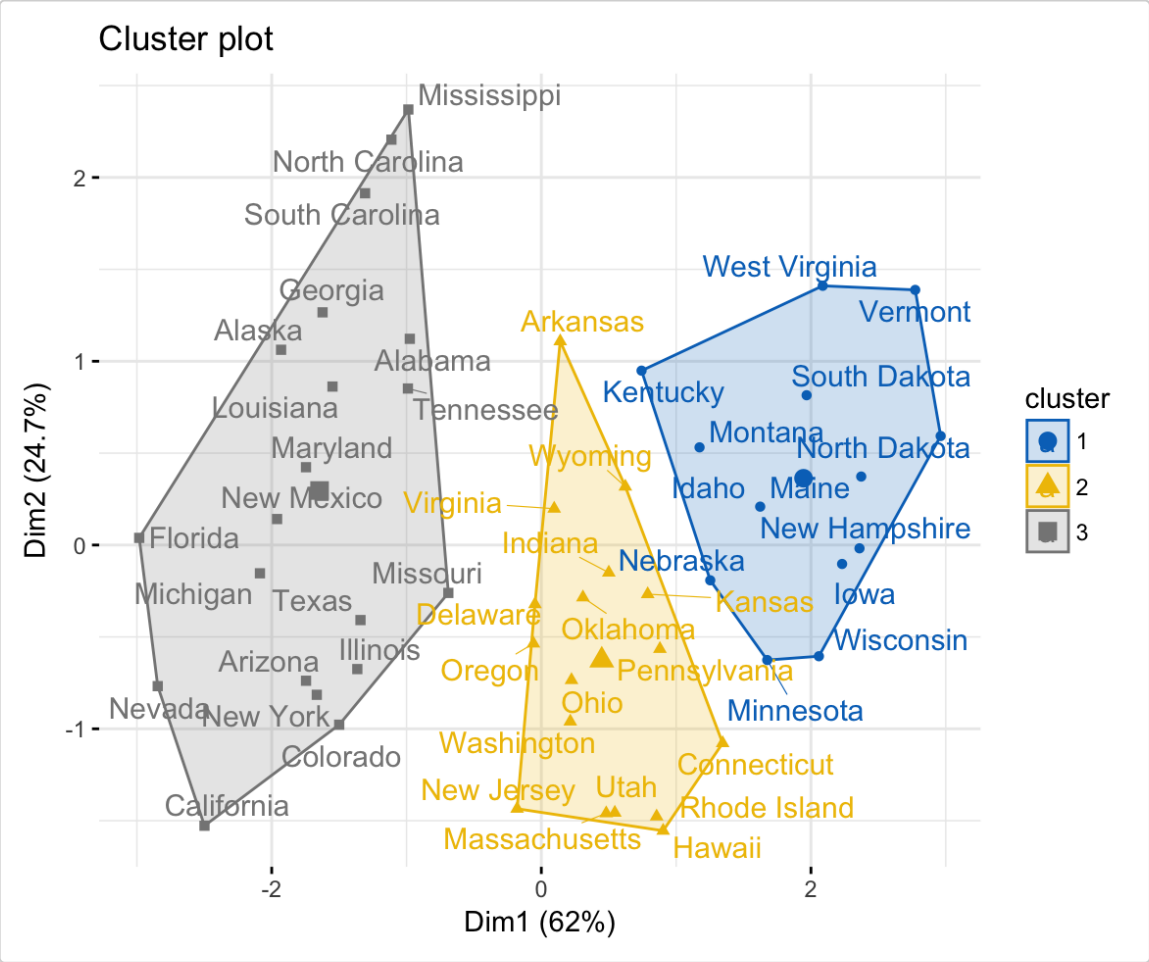
```
library("factoextra")
fviz_nbclust(my_data, kmeans,
             method = "gap_stat")
```

```
## Clustering k = 1,2,..., K.max (= 10): .. done
## Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
## .................................................. 50
## .................................................. 100
```



Compute and visualize k-means clustering:

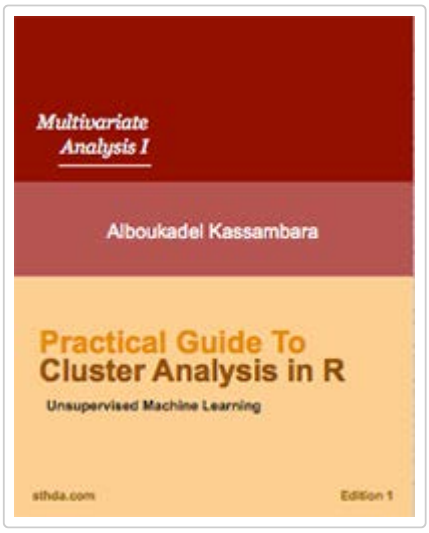
```
set.seed(123)
km.res <- kmeans(my_data, 3, nstart = 25)
# Visualize
library("factoextra")
fviz_cluster(km.res, data = my_data,
             ellipse.type = "convex",
             palette = "jco",
             repel = TRUE,
             ggtheme = theme_minimal())
```



Similarly, you can compute and visualize PAM clustering as follow:

```
# Compute PAM
library("cluster")
pam.res <- pam(my_data, 4)
# Visualize
fviz_cluster(pam.res)
```

Related Books:



Practical Guide to Cluster Analysis in R

References

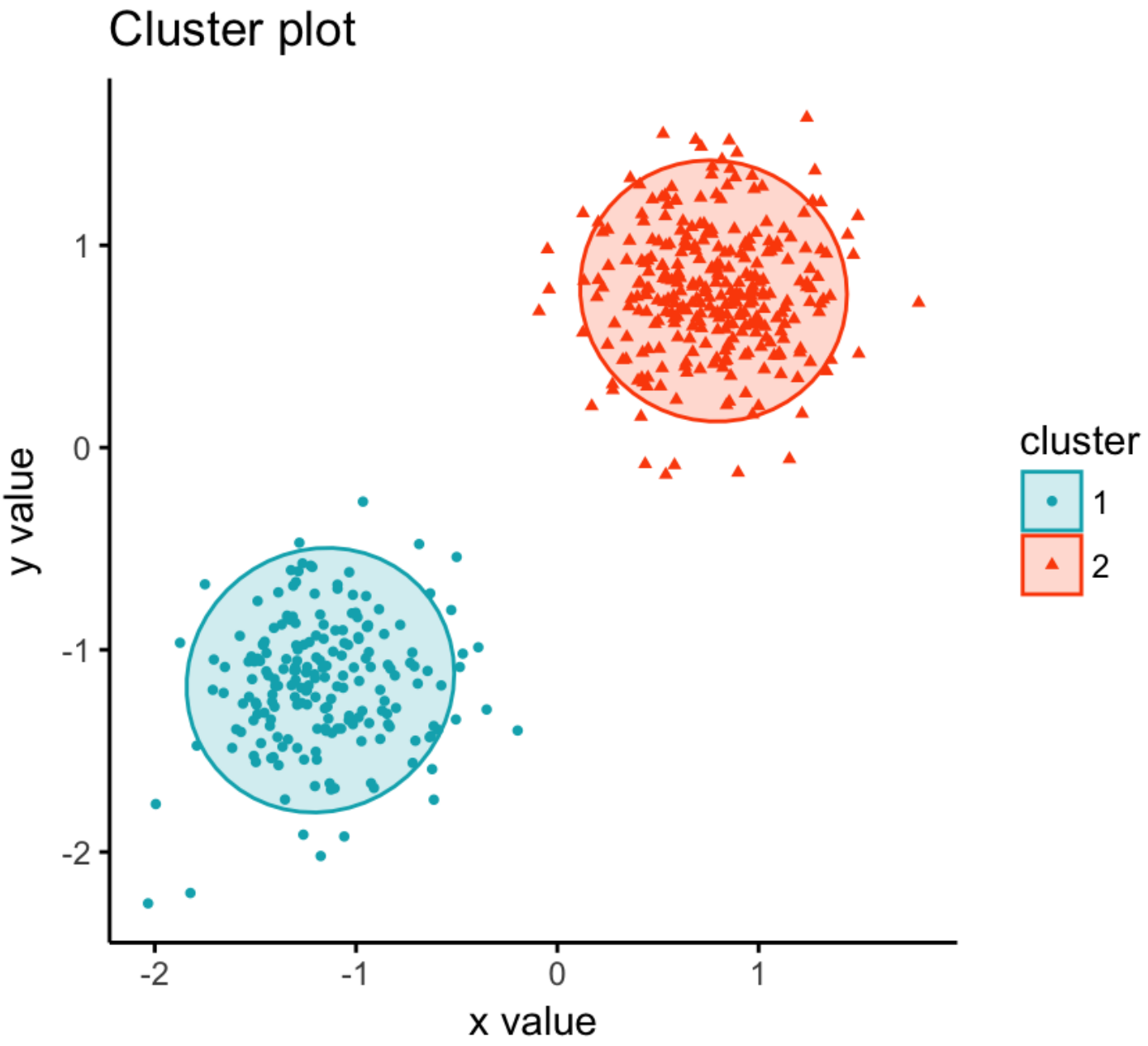
MacQueen, J. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–97. Berkeley, Calif.: University of California Press. http://projecteuclid.org:443/euclid.bsmsp/1200512992.

Sort by

Creation date

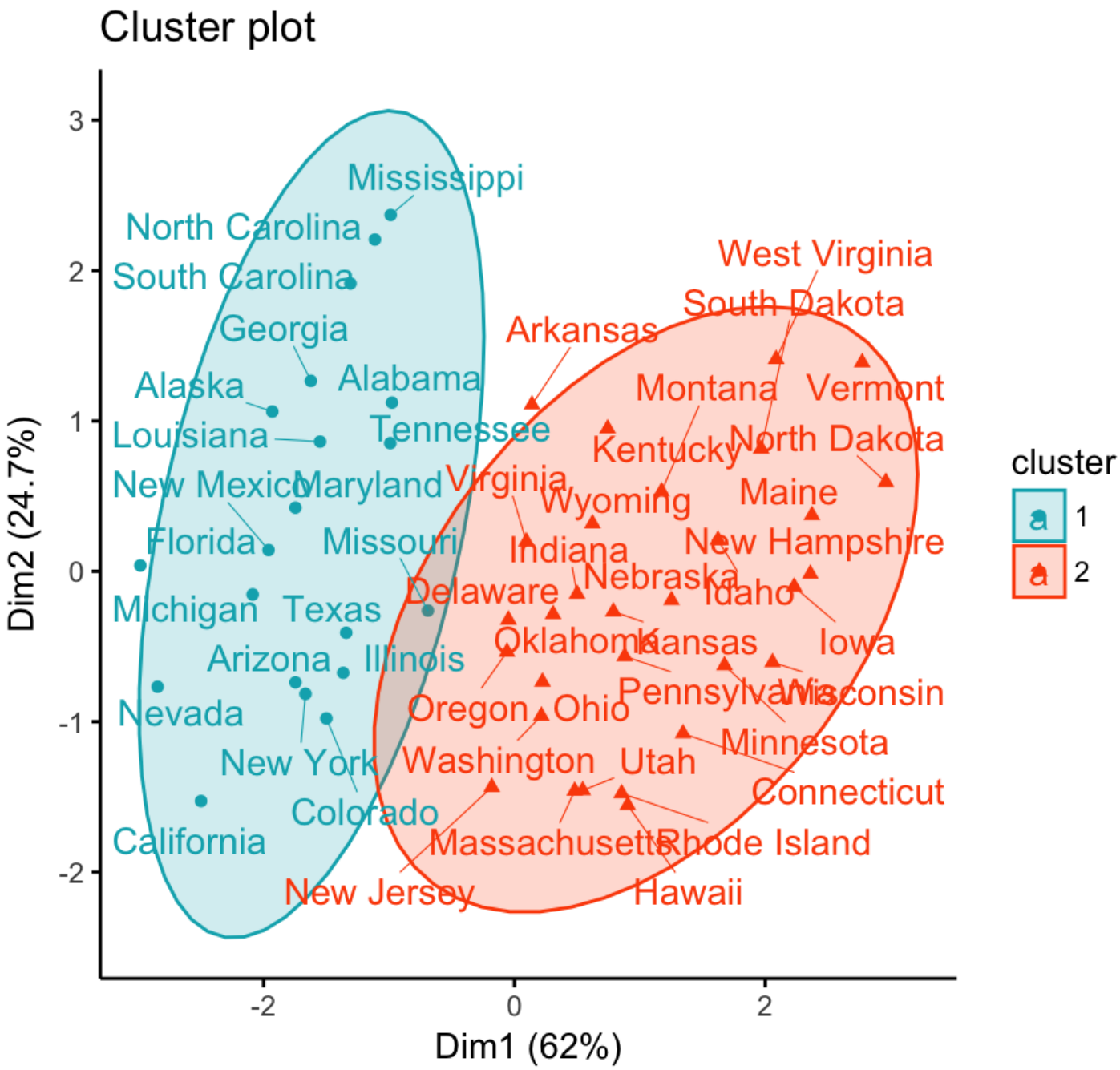
Descending

CLARA - Clustering Large Applications
By kassambara, The 04/09/2017 in Partitioning Clustering Essentials



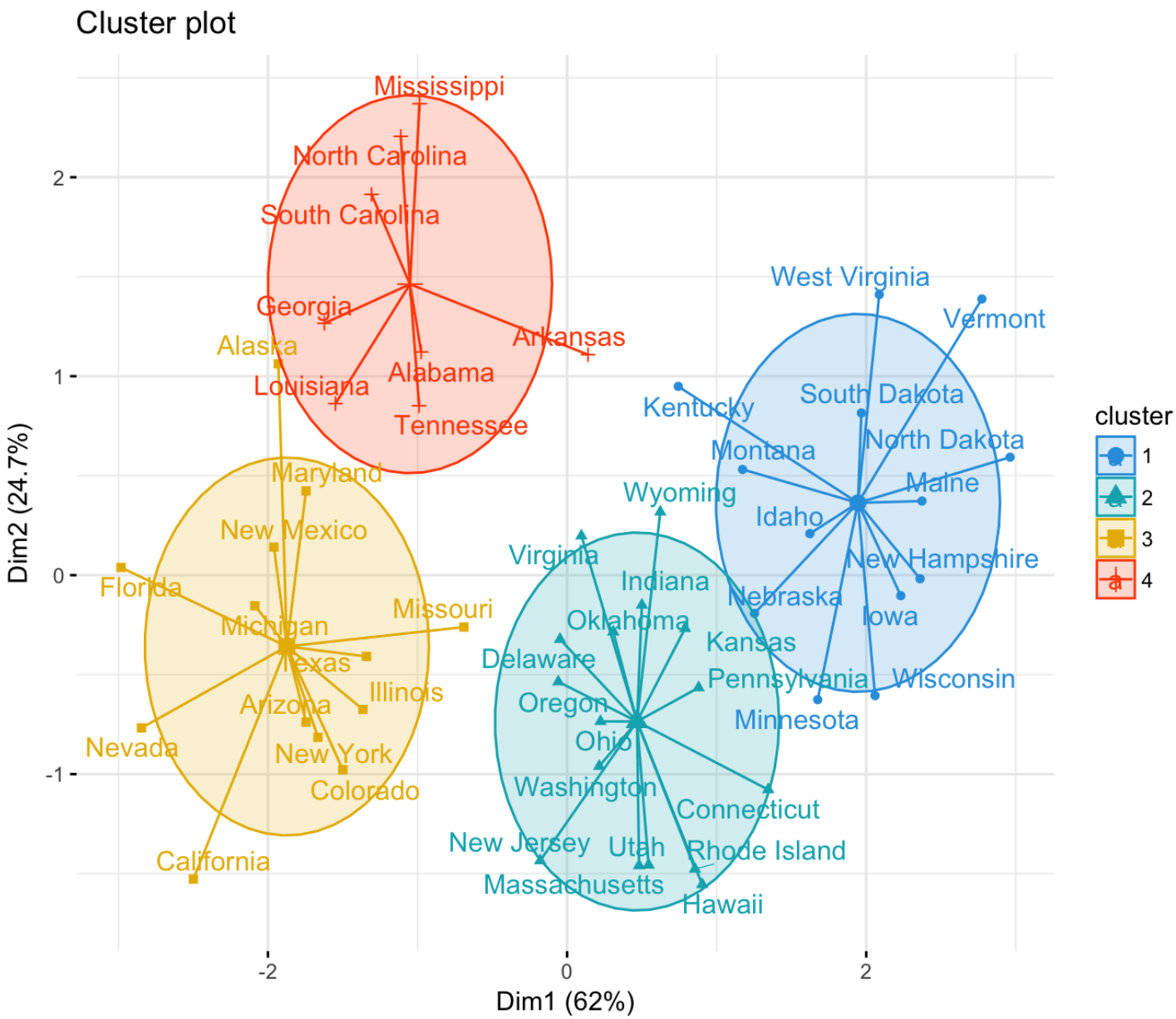
CLARA (Clustering Large Applications, (Kaufman and Rousseeuw 1990)) is an extension to k-medoids methods (Chapter @ref(k-medoids)) to deal with data containing a large number of objects (more... [Read more]

K-Medoids Essentials
By kassambara, The 04/09/2017 in Partitioning Clustering Essentials

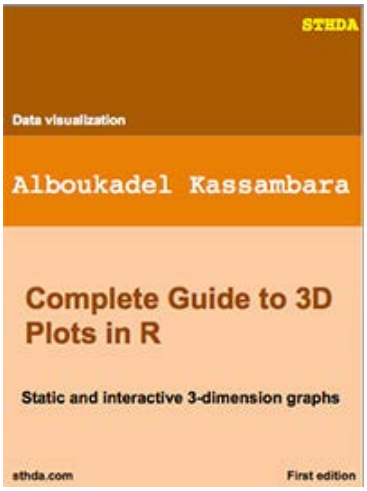


The k-medoids algorithm is a clustering approach related to k-means clustering (chapter @ref(kmeans-clustering)) for partitioning a data set into k groups or clusters. In k-medoids clustering,... [Read more]

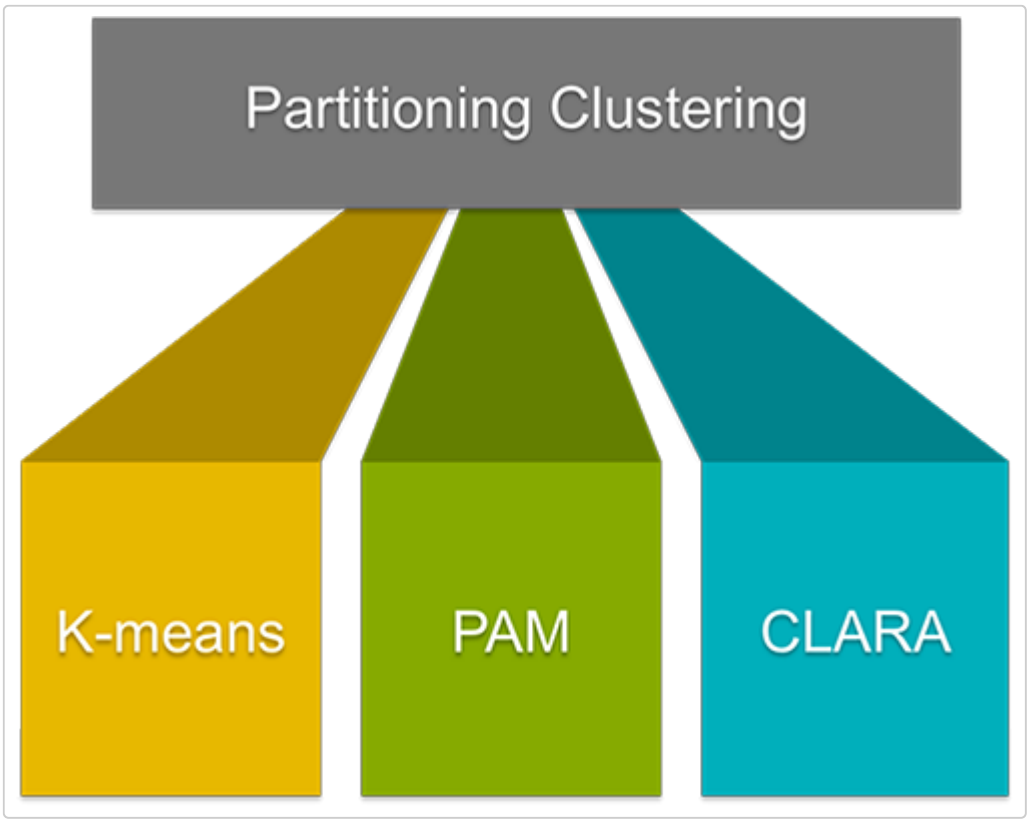
K-Means Clustering Essentials
By kassambara, The 04/09/2017 in Partitioning Clustering Essentials



K-means clustering (MacQueen 1967) is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents... [Read more]



Articles - Partitioning Clustering Essentials



Partitioning clustering are clustering methods used to classify observations, within a data set, into multiple groups based on their similarity. The algorithms require the analyst to specify the number of clusters to be generated.

This chapter describes the commonly used partitioning clustering, including:

- **K-means clustering** (MacQueen 1967), in which, each cluster is represented by the center or means of the data points belonging to the cluster. The K-means method is sensitive to anomalous data points and outliers.

- **K-medoids clustering** or **PAM** (*Partitioning Around Medoids*, Kaufman & Rousseeuw, 1990), in which, each cluster is represented by one of the objects in the cluster. PAM is less sensitive to outliers compared to k-means.
- **CLARA algorithm** (*Clustering Large Applications*), which is an extension to PAM adapted for large data sets.

For each of these methods, we provide:

- the basic idea and the key mathematical concepts
- the clustering algorithm and implementation in R software
- R lab sections with many examples for cluster analysis and visualization

The following R packages will be used to compute and visualize partitioning clustering:

- *stats* package for computing K-means
- *cluster* package for computing PAM and CLARA algorithms
- *factoextra* for beautiful visualization of clusters

Related articles

K-Means Clustering Essentials

Contents:

- K-means basic ideas
- K-means algorithm
- Computing k-means clustering in R
  - Data
  - Required R packages and functions: **stats::kmeans**()
  - Estimating the optimal number of clusters: **factoextra::fviz_nbclust**()
  - Computing k-means clustering
  - Accessing to the results of kmeans() function
  - Visualizing k-means clusters: **factoextra::fviz_cluster**()
- K-means clustering advantages and disadvantages
- Alternative to k-means clustering

K-Medoids Essentials: PAM clustering

Contents:

- PAM concept
- PAM algorithm
- Computing PAM in R
  - Data
  - Required R packages and functions: **cluster::pam**() or **fpc::pamk**()
  - Estimating the optimal number of clusters: **factoextra::fviz_nbclust**()
  - Computing PAM clustering
  - Accessing to the results of the pam() function
  - Visualizing PAM clusters: **factoextra::fviz_cluster**()

CLARA - Clustering Large Applications

Contents:

- CLARA concept
- CLARA Algorithm
- Computing CLARA in R
  - Data format and preparation
  - Required R packages and functions: **cluster::clara**()
  - Estimating the optimal number of clusters: **factoextra::fviz_nbclust**()
  - Computing CLARA
  - Visualizing CLARA clusters: **factoextra::fviz_cluster**()

Quick start

Data preparation:

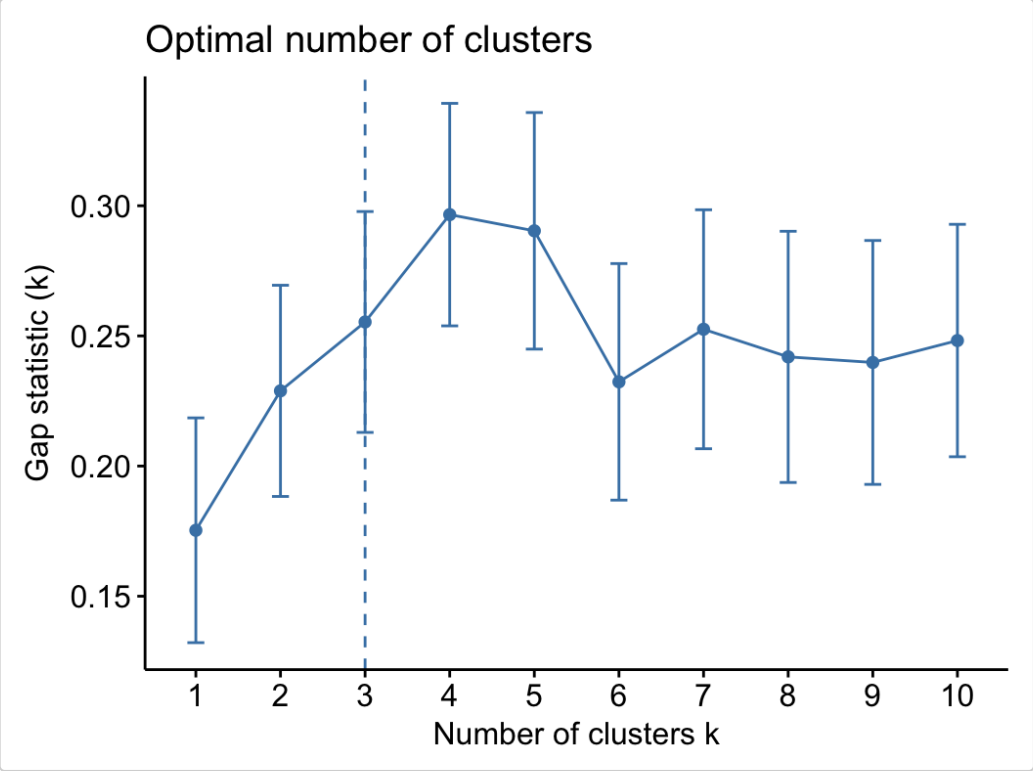
```
# Load data
data("USArrests")
my_data <- USArrests
# Remove any missing value (i.e, NA values for not available)
my_data <- na.omit(my_data)
# Scale variables
my_data <- scale(my_data)
# View the firt 3 rows
head(my_data, n = 3)
```

```
##          Murder Assault UrbanPop     Rape
## Alabama 1.2426   0.783   -0.521 -0.00342
## Alaska  0.5079   1.107   -1.212  2.48420
## Arizona 0.0716   1.479    0.999  1.04288
```

Determine the optimal number of clusters for k-means clustering:

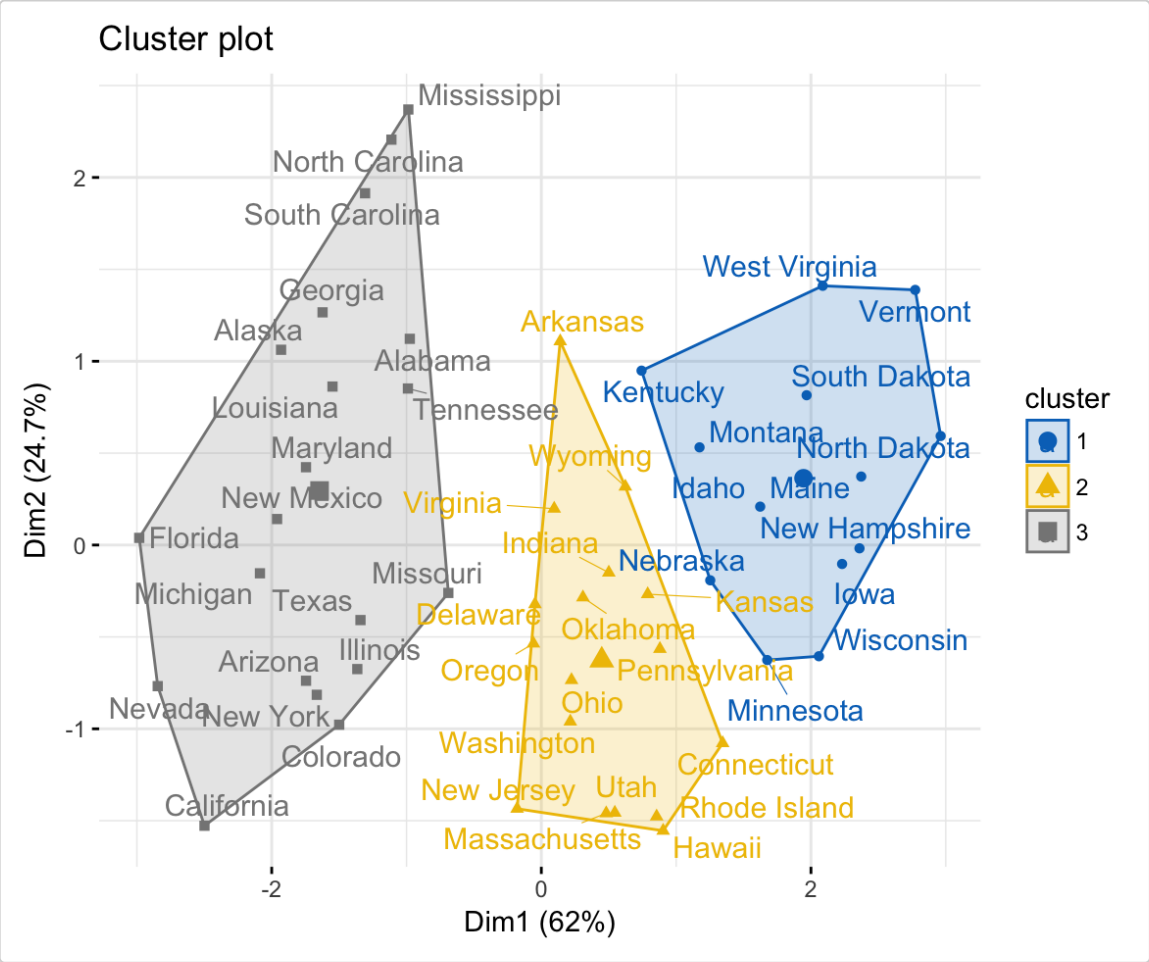
```
library("factoextra")
fviz_nbclust(my_data, kmeans,
             method = "gap_stat")
```

```
## Clustering k = 1,2,..., K.max (= 10): .. done
## Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
## .................................................. 50
## .................................................. 100
```



Compute and visualize k-means clustering:

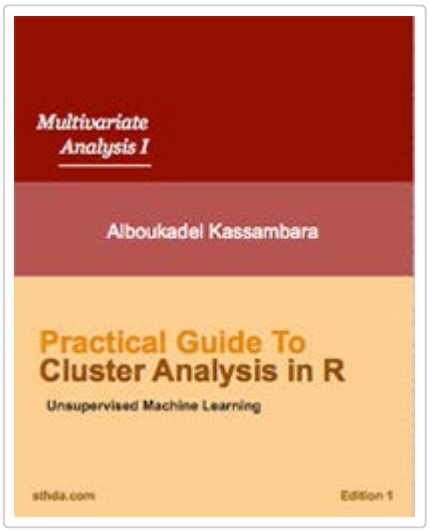
```
set.seed(123)
km.res <- kmeans(my_data, 3, nstart = 25)
# Visualize
library("factoextra")
fviz_cluster(km.res, data = my_data,
             ellipse.type = "convex",
             palette = "jco",
             repel = TRUE,
             ggtheme = theme_minimal())
```



Similarly, you can compute and visualize PAM clustering as follow:

```
# Compute PAM
library("cluster")
pam.res <- pam(my_data, 4)
# Visualize
fviz_cluster(pam.res)
```

Related Books:



Practical Guide to Cluster Analysis in R

References

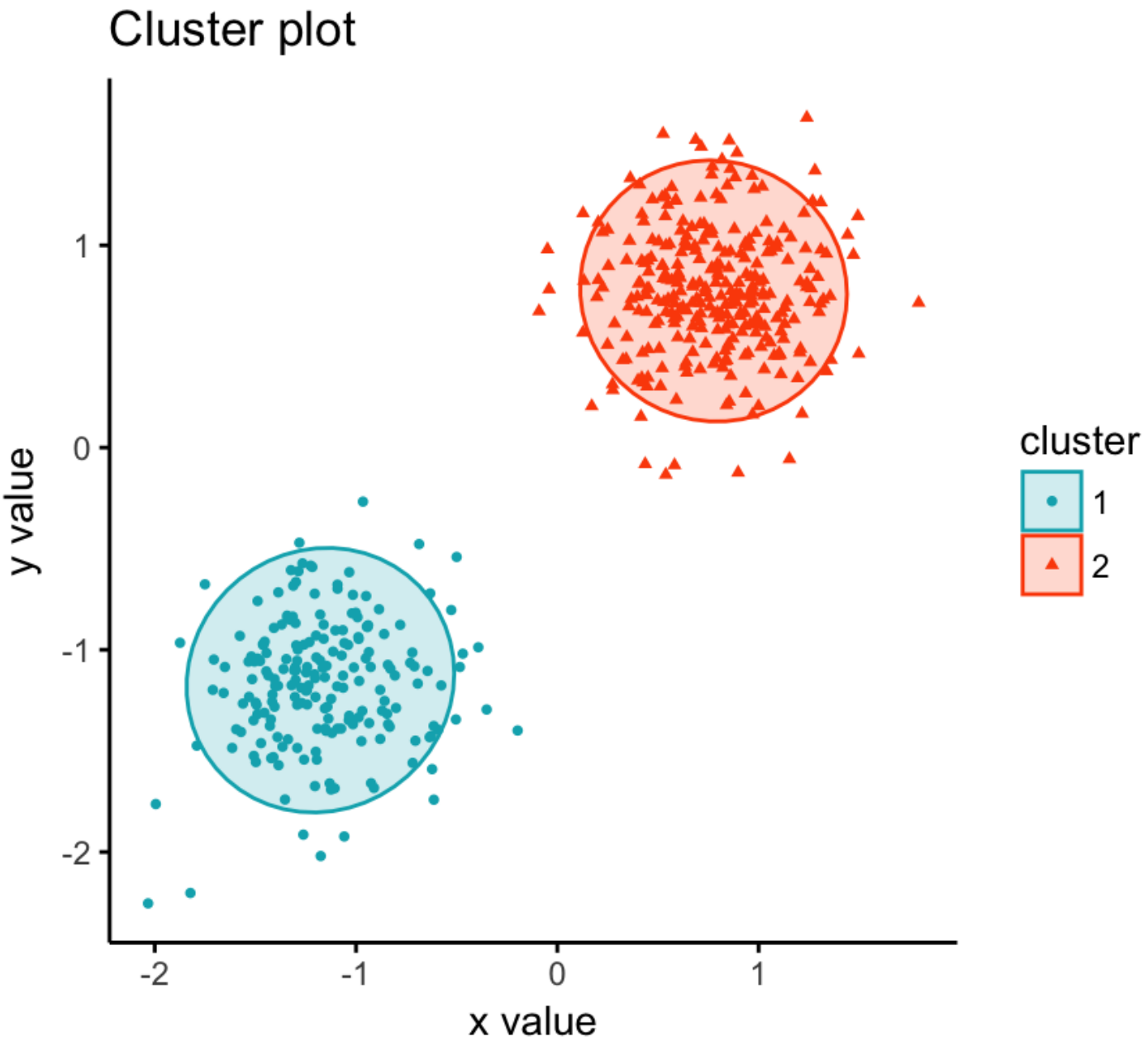
MacQueen, J. 1967. “Some Methods for Classification and Analysis of Multivariate Observations.” In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–97. Berkeley, Calif.: University of California Press. http://projecteuclid.org:443/euclid.bsmsp/1200512992.

Sort by

Creation date

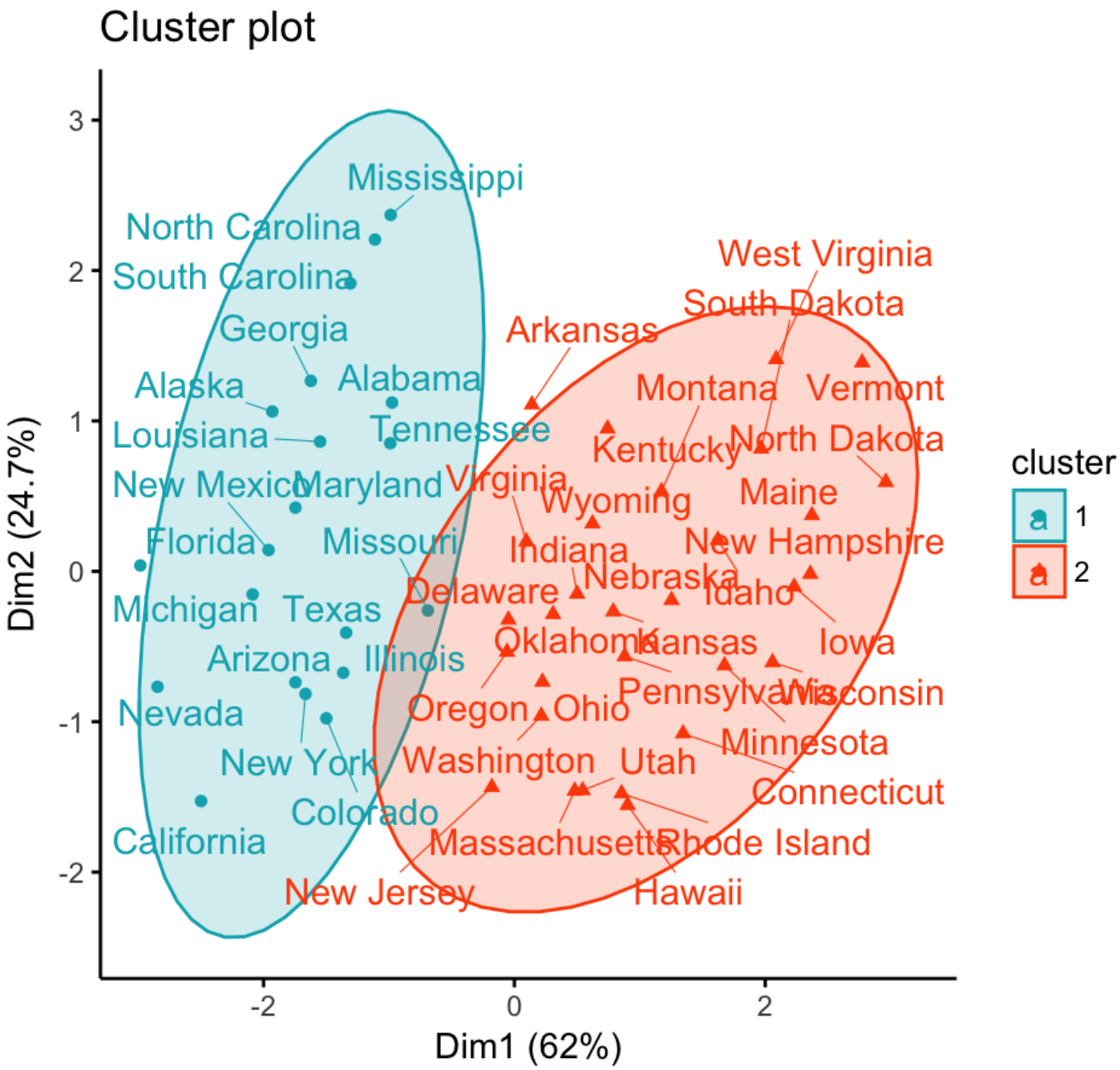
Descending

CLARA - Clustering Large Applications
By kassambara, The 04/09/2017 in Partitioning Clustering Essentials



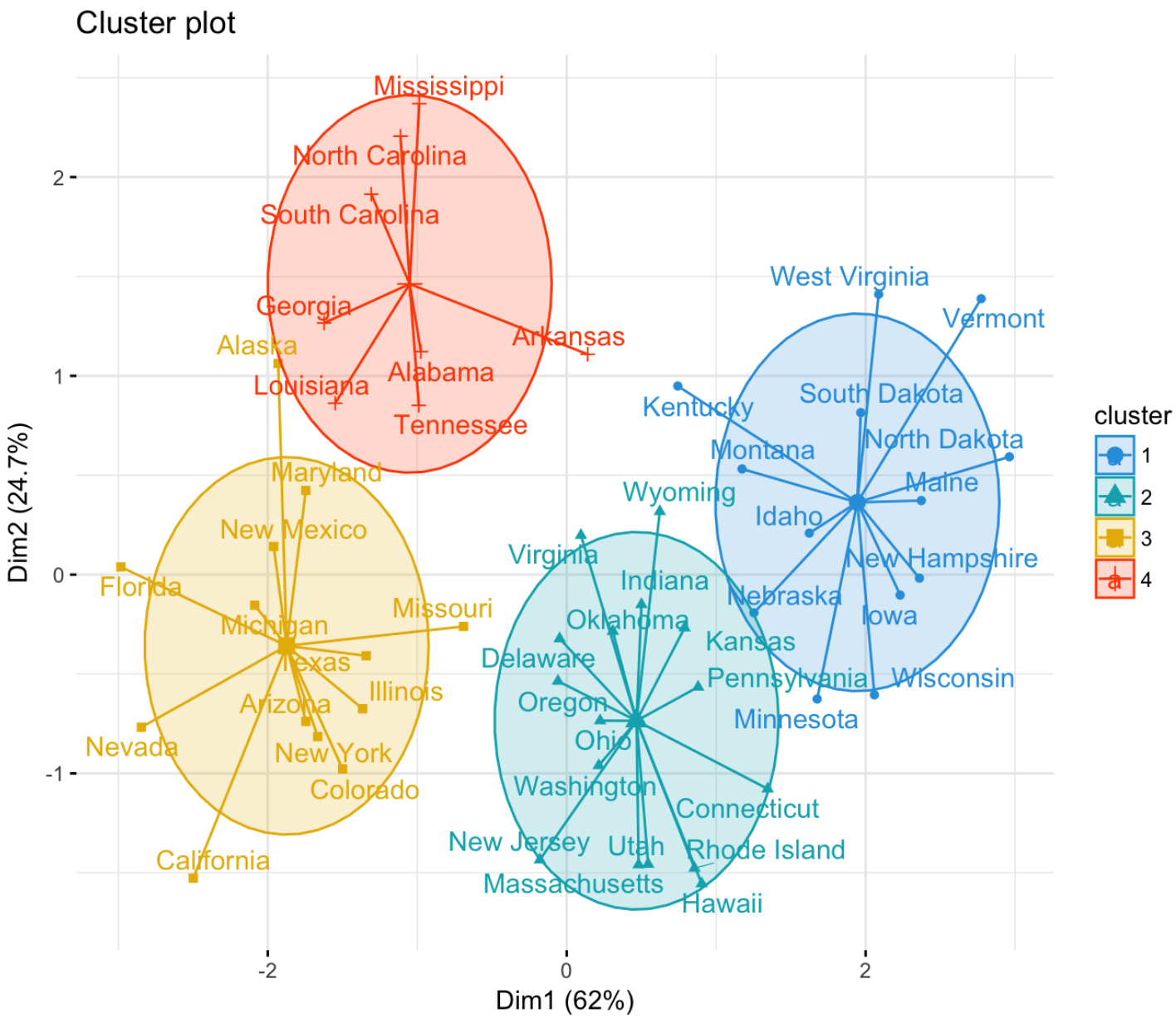
CLARA (Clustering Large Applications, (Kaufman and Rousseeuw 1990)) is an extension to k-medoids methods (Chapter @ref(k-medoids)) to deal with data containing a large number of objects (more... [Read more]

K-Medoids Essentials
By kassambara, The 04/09/2017 in Partitioning Clustering Essentials

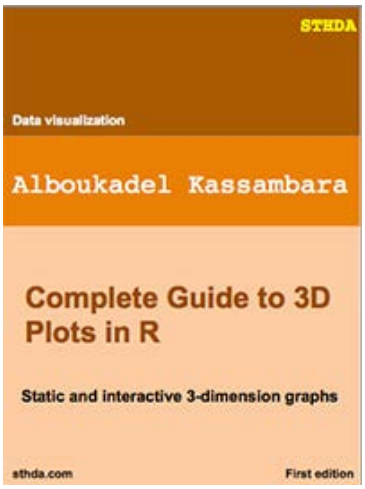


The k-medoids algorithm is a clustering approach related to k-means clustering (chapter @ref(kmeans-clustering)) for partitioning a data set into k groups or clusters. In k-medoids clustering,... [Read more]

K-Means Clustering Essentials
By kassambara, The 04/09/2017 in Partitioning Clustering Essentials



K-means clustering (MacQueen 1967) is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents... [Read more]



Guest Book

If you like this web site or if you have a suggestion, let us know. This encourages us to continue....

By kassambara

Guest Book
Blogroll
- R-Bloggers

1. Home
2. Articles
3. Cluster Analysis in R: Practical Guide
4. Partitioning Clustering Essentials
5. CLARA - Clustering Large Applications

Articles - Partitioning Clustering Essentials

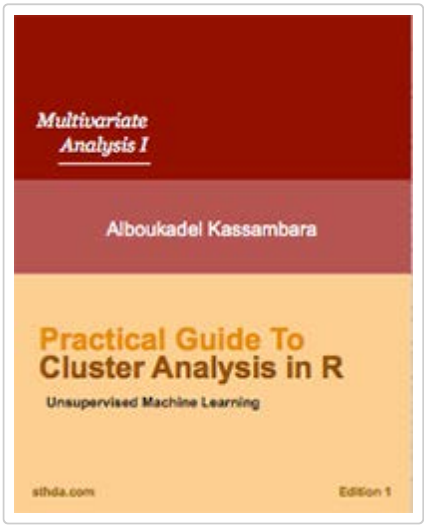
CLARA - Clustering Large Applications

kassambara | 04/09/2017 | 6434 | Comments (8) | Partitioning Clustering Essentials | PAM clustering, K medoids

CLARA (Clustering Large Applications, (Kaufman and Rousseeuw 1990)) is an extension to k-medoids methods (Chapter @ref(k-medoids)) to deal with data containing a large number of objects (more than several thousand observations) in order to reduce computing time and RAM storage problem. This is achieved using the sampling approach.

Contents:

Related Books:



Practical Guide to Cluster Analysis in R

CLARA concept

Instead of finding medoids for the entire data set, CLARA considers a small sample of the data with fixed size (*sampsize*) and applies the PAM algorithm (Chapter @ref(k-medoids)) to generate an optimal set of medoids for the sample. The quality of resulting medoids is measured by the average dissimilarity between every object in the entire data set and the medoid of its cluster, defined as the cost function.

CLARA repeats the sampling and clustering processes a pre-specified number of times in order to minimize the sampling bias. The final clustering results correspond to the set of medoids with the minimal cost. The CLARA algorithm is summarized in the next section.

CLARA Algorithm

The algorithm is as follow:

1. Split randomly the data sets in multiple subsets with fixed size (sampsize)
2. Compute PAM algorithm on each subset and choose the corresponding k representative objects (medoids). Assign each observation of the entire data set to the closest medoid.
3. Calculate the mean (or the sum) of the dissimilarities of the observations to their closest medoid. This is used as a measure of the goodness of the clustering.
4. Retain the sub-dataset for which the mean (or sum) is minimal. A further analysis is carried out on the final partition.

Note that, each sub-data set is forced to contain the medoids obtained from the best sub-data set until then. Randomly drawn observations are added to this set until sampsize has been reached.

Computing CLARA in R

Data format and preparation

To compute the CLARA algorithm in R, the data should be prepared as indicated in Chapter @ref(data-preparation-and-r-packages).

Here, we'll generate use a random data set. To make the result reproducible, we start by using the function *set.seed*().

```
set.seed(1234)
# Generate 500 objects, divided into 2 clusters.
df <- rbind(cbind(rnorm(200,0,8), rnorm(200,0,8)),
           cbind(rnorm(300,50,8), rnorm(300,50,8)))
# Specify column and row names
colnames(df) <- c("x", "y")
rownames(df) <- paste0("S", 1:nrow(df))
# Previewing the data
head(df, nrow = 6)
```

```
##          x    y
## S1  -9.66 3.88
## S2   2.22 5.57
## S3   8.68 1.48
## S4 -18.77 5.61
## S5   3.43 2.49
## S6   4.05 6.08
```

Required R packages and functions

The function *clara*() [*cluster* package] can be used to compute *CLARA*. The simplified format is as follow:

```
clara(x, k, metric = "euclidean", stand = FALSE,
      samples = 5, pamLike = FALSE)
```

- **x**: a numeric data matrix or data frame, each row corresponds to an observation, and each column corresponds to a variable. Missing values (NAs) are allowed.
- **k**: the number of clusters.
- **metric**: the distance metrics to be used. Available options are “euclidean” and “manhattan”. Euclidean distances are root sum-of-squares of differences, and manhattan distances are the sum of absolute differences. Read more on distance measures (Chapter). Note that, manhattan distance is less sensitive to outliers.
- **stand**: logical value; if true, the variables (columns) in x are standardized before calculating the dissimilarities. Note that, it’s recommended to standardize variables before clustering.
- **samples**: number of samples to be drawn from the data set. Default value is 5 but it’s recommended a much larger value.
- **pamLike**: logical indicating if the same algorithm in the **pam**() function should be used. This should be always true.

To create a beautiful graph of the clusters generated with the *pam*() function, will use the *factoextra* package.

1. Installing required packages:

```
install.packages(c("cluster", "factoextra"))
```

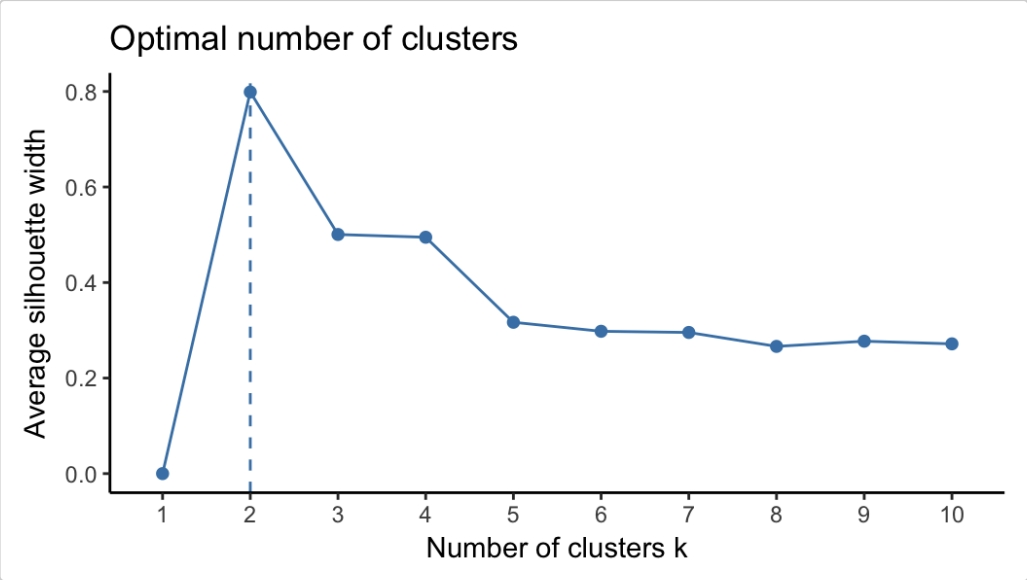
2. Loading the packages:

```
library(cluster)
library(factoextra)
```

Estimating the optimal number of clusters

To estimate the optimal number of clusters in your data, it's possible to use the average silhouette method as described in PAM clustering chapter (Chapter @ref(k-medoids)). The R function *fviz_nbclust*() [*factoextra* package] provides a solution to facilitate this step.

```
library(cluster)
library(factoextra)
fviz_nbclust(df, clara, method = "silhouette")+
  theme_classic()
```



From the plot, the suggested number of clusters is 2. In the next section, we'll classify the observations into 2 clusters.

Computing CLARA

The R code below computes PAM algorithm with k = 2:

```
# Compute CLARA
clara.res <- clara(df, 2, samples = 50, pamLike = TRUE)
# Print components of clara.res
print(clara.res)
```

```
## Call:     clara(x = df, k = 2, samples = 50, pamLike = TRUE)
## Medoids:
##          x     y
## S121 -1.53  1.15
## S455 48.36 50.23
## Objective function:   9.88
## Clustering vector:    Named int [1:500] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "names")= chr [1:500] "S1" "S2" "S3" "S4" "S5" "S6" "S7" ...
## Cluster sizes:            200 300
## Best sample:
##  [1] S37  S49  S54  S63  S68  S71  S76  S80  S82  S101 S103 S108 S109 S118
## [15] S121 S128 S132 S138 S144 S162 S203 S210 S216 S231 S234 S249 S260 S261
## [29] S286 S299 S304 S305 S312 S315 S322 S350 S403 S450 S454 S455 S456 S465
## [43] S488 S497
##
## Available components:
##  [1] "sample"     "medoids"    "i.med"      "clustering" "objective"
##  [6] "clusinfo"   "diss"       "call"       "silinfo"    "data"
```

The output of the function *clara*() includes the following components:

- **medoids**: Objects that represent clusters
- **clustering**: a vector containing the cluster number of each object
- **sample**: labels or case numbers of the observations in the best sample, that is, the sample used by the clara algorithm for the final partition.

If you want to add the point classifications to the original data, use this:

```
dd <- cbind(df, cluster = clara.res$cluster)
head(dd, n = 4)
```

```
##          x    y cluster
## S1  -9.66 3.88       1
## S2   2.22 5.57       1
## S3   8.68 1.48       1
## S4 -18.77 5.61       1
```

You can access to the results returned by *clara*() as follow:

```
# Medoids
clara.res$medoids
```

```
##           x     y
## S121 -1.53  1.15
## S455 48.36 50.23
```

```
# Clustering
head(clara.res$clustering, 10)
```

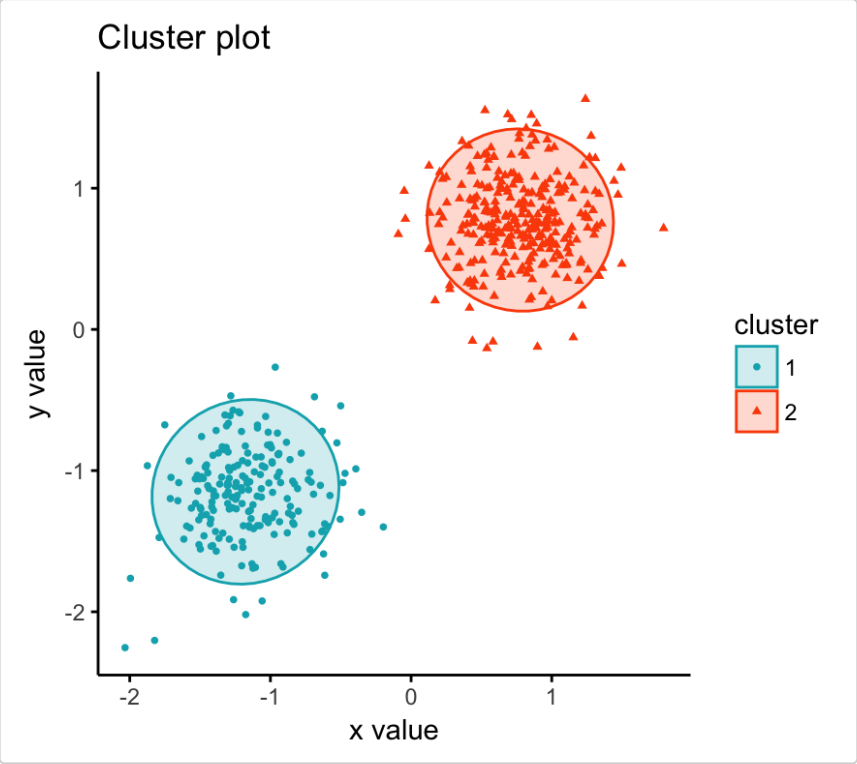
```
##  S1  S2  S3  S4  S5  S6  S7  S8  S9 S10
##   1   1   1   1   1   1   1   1   1   1
```

The **medoids** are S121, S455

Visualizing CLARA clusters

To visualize the partitioning results, we'll use the function *fviz_cluster*() [*factoextra* package]. It draws a scatter plot of data points colored by cluster numbers.

```
fviz_cluster(clara.res,
             palette = c("#00AFBB", "#FC4E07"), # color palette
             ellipse.type = "t", # Concentration ellipse
             geom = "point", pointsize = 1,
             ggtheme = theme_classic()
             )
```



Summary

The CLARA (Clustering Large Applications) algorithm is an extension to the PAM (Partitioning Around Medoids) clustering method for large data sets. It intended to reduce the computation time in the case of large data set.

As almost all partitioning algorithm, it requires the user to specify the appropriate number of clusters to be produced. This can be estimated using the function *fviz_nbclust* [in *factoextra* R package].

The R function *clara*() [*cluster* package] can be used to compute CLARA algorithm. The simplified format is clara(x, k, pamLike = TRUE), where "x" is the data and k is the number of clusters to be generated.

After, computing CLARA, the R function *fviz_cluster*() [*factoextra* package] can be used to visualize the results. The format is fviz_cluster(clara.res), where clara.res is the CLARA results.

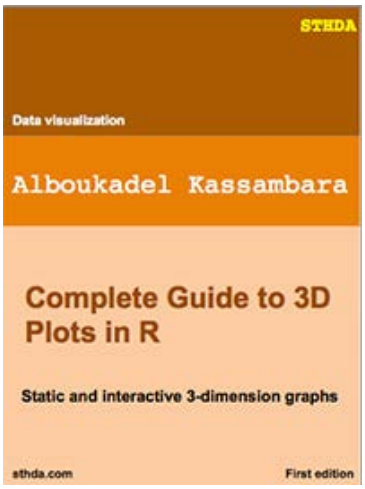
References

Kaufman, Leonard, and Peter Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*.

Last update : 24/09/2017
□ □ □ □ □ 1 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!



Articles - Partitioning Clustering Essentials

K-Medoids Essentials

kassambara | 04/09/2017 | 9328 | Comments (4) | Partitioning Clustering Essentials | Unsupervised machine learning, Multivariate Analysis, Genomic data visualization, PAM clustering

The **k-medoids algorithm** is a clustering approach related to k-means clustering (chapter @ref(kmeans-clustering)) for partitioning a data set into k groups or clusters. In k-medoids clustering, each cluster is represented by one of the data point in the cluster. These points are named cluster medoids.

The term medoid refers to an object within a cluster for which average dissimilarity between it and all the other the members of the cluster is minimal. It corresponds to the most centrally located point in the cluster. These objects (one per cluster) can be considered as a representative example of the members of that cluster which may be useful in some situations. Recall that, in k-means clustering, the center of a given cluster is calculated as the mean value of all the data points in the cluster.

K-medoid is a robust alternative to k-means clustering. This means that, the algorithm is less sensitive to noise and outliers, compared to k-means, because it uses medoids as cluster centers instead of means (used in k-means).

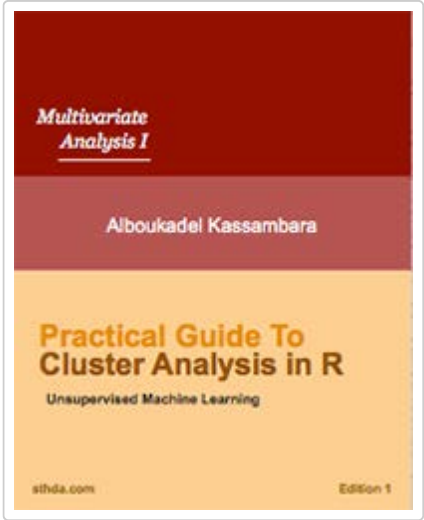
The k-medoids algorithm requires the user to specify k, the number of clusters to be generated (like in k-means clustering). A useful approach to determine the optimal number of clusters is the **silhouette** method, described in the next sections.

The most common k-medoids clustering methods is the **PAM** algorithm (**Partitioning Around Medoids**, (Kaufman and Rousseeuw 1990)).

In this article, We'll describe the PAM algorithm and provide practical examples using **R** software. In the next chapter, we'll also discuss a variant of PAM named **CLARA** (Clustering Large Applications) which is used for analyzing large data sets.

Contents:

Related Books:



Practical Guide to Cluster Analysis in R

PAM concept

The use of means implies that k-means clustering is highly sensitive to outliers. This can severely affects the assignment of observations to clusters. A more robust algorithm is provided by the **PAM** algorithm.

PAM algorithm

The PAM algorithm is based on the search for k representative objects or medoids among the observations of the data set.

After finding a set of k medoids, clusters are constructed by assigning each observation to the nearest medoid. Next, each selected medoid m and each non-medoid data point are swapped and the objective function is computed. The objective function corresponds to the sum of the dissimilarities of all objects to their nearest medoid.

The SWAP step attempts to improve the quality of the clustering by exchanging selected objects (medoids) and non-selected objects. If the objective function can be reduced by interchanging a selected object with an unselected object, then the swap is carried out. This is continued until the objective function can no longer be decreased. The goal is to find k representative objects which minimize the sum of the dissimilarities of the observations to their closest representative object.

In summary, PAM algorithm proceeds in two phases as follow:

Build phase:

1. Select k objects to become the medoids, or in case these objects were provided use them as the medoids;
2. Calculate the dissimilarity matrix if it was not provided;
3. Assign every object to its closest medoid;

Swap phase: 4. For each cluster search if any of the object of the cluster decreases the average dissimilarity coefficient; if it does, select the entity that decreases this coefficient the most as the medoid for this cluster; 5. If at least one medoid has changed go to (3), else end the algorithm.

As mentioned above, the PAM algorithm works with a matrix of dissimilarity, and to compute this matrix the algorithm can use two metrics:

1. The euclidean distances, that are the root sum-of-squares of differences;
2. And, the Manhattan distance that are the sum of absolute distances.

Note that, in practice, you should get similar results most of the time, using either euclidean or Manhattan distance. If your data contains outliers, Manhattan distance should give more robust results, whereas euclidean would be influenced by unusual values.

Read more on distance measures in Chapter @ref(clustering-distance-measures).

Computing PAM in R

Data

We'll use the demo data sets "USArrests", which we start by scaling (Chapter @ref(data-preparation-and-r-packages) using the R function *scale()* as follow:

```
data("USArrests")      # Load the data set
df <- scale(USArrests) # Scale the data
head(df, n = 3)        # View the firt 3 rows of the data
```

```
##            Murder Assault UrbanPop     Rape
## Alabama 1.2426   0.783   -0.521 -0.00342
## Alaska  0.5079   1.107   -1.212  2.48420
## Arizona 0.0716   1.479    0.999  1.04288
```

Required R packages and functions

The function *pam*() [*cluster* package] and *pamk()* [*fpc* package] can be used to compute **PAM**.

The function *pamk*() does not require a user to decide the number of clusters K.

In the following examples, we'll describe only the function *pam*(), which simplified format is:

```
pam(x, k, metric = "euclidean", stand = FALSE)
```

- **x**: possible values includes:
  - Numeric data matrix or numeric data frame: each row corresponds to an observation, and each column corresponds to a variable.
  - Dissimilarity matrix: in this case x is typically the output of **daisy()** or **dist()**
- **k**: The number of clusters
- **metric**: the distance metrics to be used. Available options are "euclidean" and "manhattan".

- **stand**: logical value; if true, the variables (columns) in x are standardized before calculating the dissimilarities. Ignored when x is a dissimilarity matrix.

To create a beautiful graph of the clusters generated with the *pam*() function, will use the *factoextra* package.

1. Installing required packages:

```
install.packages(c("cluster", "factoextra"))
```

2. Loading the packages:

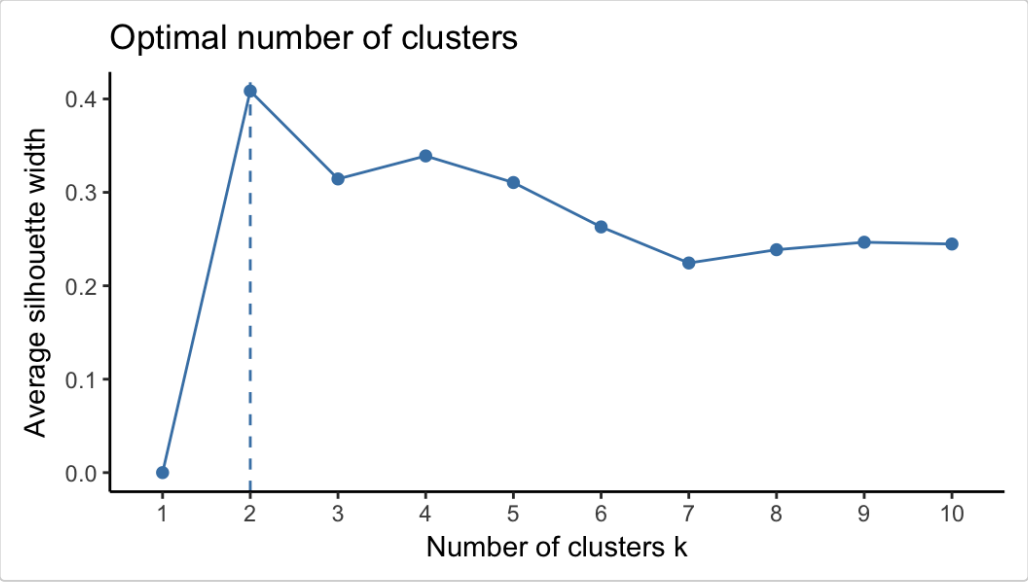
```
library(cluster)
library(factoextra)
```

Estimating the optimal number of clusters

To estimate the optimal number of clusters, we'll use the average silhouette method. The idea is to compute PAM algorithm using different values of clusters k. Next, the average clusters silhouette is drawn according to the number of clusters. The average silhouette measures the quality of a clustering. A high average silhouette width indicates a good clustering. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990).

The R function *fviz_nbclust*() [*factoextra* package] provides a convenient solution to estimate the optimal number of clusters.

```
library(cluster)
library(factoextra)
fviz_nbclust(df, pam, method = "silhouette")+
  theme_classic()
```



From the plot, the suggested number of clusters is 2. In the next section, we'll classify the observations into 2 clusters.

Computing PAM clustering

The R code below computes PAM algorithm with k = 2:

```
pam.res <- pam(df, 2)
print(pam.res)
```

```
## Medoids:
##            ID Murder Assault UrbanPop   Rape
## New Mexico 31  0.829   1.371    0.308  1.160
## Nebraska   27 -0.801  -0.825   -0.245 -0.505
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              2              2              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              1              2              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              2              1              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              2              1              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              1              2              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              2              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              2              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              1              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              2              2              2
## Objective function:
## build  swap
##  1.44  1.37
##
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
##  [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

The printed output shows:

- the cluster medoids: a matrix, which rows are the medoids and columns are variables
- the clustering vector: A vector of integers (from 1:k) indicating the cluster to which each point is allocated

If you want to add the point classifications to the original data, use this:

```
dd <- cbind(USArrests, cluster = pam.res$cluster)
head(dd, n = 3)
```

```
##         Murder Assault UrbanPop Rape cluster
## Alabama   13.2     236       58 21.2       1
## Alaska    10.0     263       48 44.5       1
## Arizona    8.1     294       80 31.0       1
```

Accessing to the results of the pam() function

The function *pam*() returns an object of class *pam* which components include:

- **medoids**: Objects that represent clusters
- **clustering**: a vector containing the cluster number of each object

These components can be accessed as follow:

```
# Cluster medoids: New Mexico, Nebraska
pam.res$medoids
```

```
##            Murder Assault UrbanPop   Rape
## New Mexico  0.829   1.371    0.308  1.160
## Nebraska   -0.801  -0.825   -0.245 -0.505
```

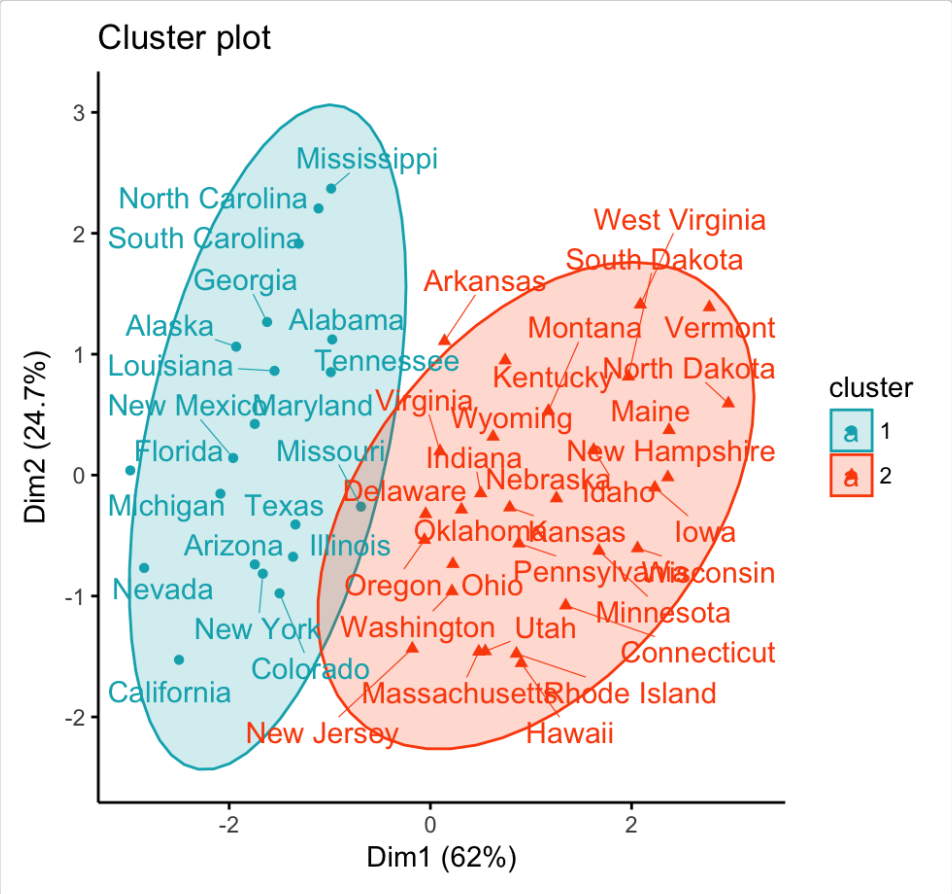
```
# Cluster numbers
head(pam.res$clustering)
```

```
##    Alabama     Alaska    Arizona   Arkansas California   Colorado
##          1          1          1          2          1          1
```

Visualizing PAM clusters

To visualize the partitioning results, we'll use the function *fviz_cluster*() [*factoextra* package]. It draws a scatter plot of data points colored by cluster numbers. If the data contains more than 2 variables, the *Principal Component Analysis (PCA)* algorithm is used to reduce the dimensionality of the data. In this case, the first two principal dimensions are used to plot the data.

```
fviz_cluster(pam.res,
             palette = c("#00AFBB", "#FC4E07"), # color palette
             ellipse.type = "t", # Concentration ellipse
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_classic()
             )
```



Summary

The K-medoids algorithm, PAM, is a robust alternative to k-means for partitioning a data set into clusters of observation.

In k-medoids method, each cluster is represented by a selected object within the cluster. The selected objects are named medoids and corresponds to the most centrally located points within the cluster.

The PAM algorithm requires the user to know the data and to indicate the appropriate number of clusters to be produced. This can be estimated using the function *fviz_nbclust* [in *factoextra* R package].

The R function *pam*() [*cluster* package] can be used to compute PAM algorithm. The simplified format is pam(x, k), where "x" is the data and k is the number of clusters to be generated.

After, performing PAM clustering, the R function *fviz_cluster*() [**factoextra** package] can be used to visualize the results. The format is fviz_cluster(pam.res), where pam.res is the PAM results.

Note that, for large data sets, () may need too much memory or too much computation time. In this case, the function () is preferable. This should not be a problem for modern computers.

References

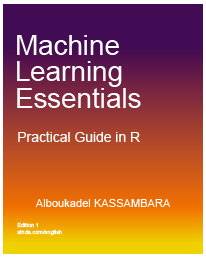
Kaufman, Leonard, and Peter Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*.

Last update : 24/09/2017
□ □ □ □ □ 0 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

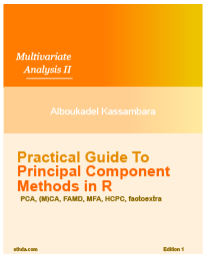
Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!

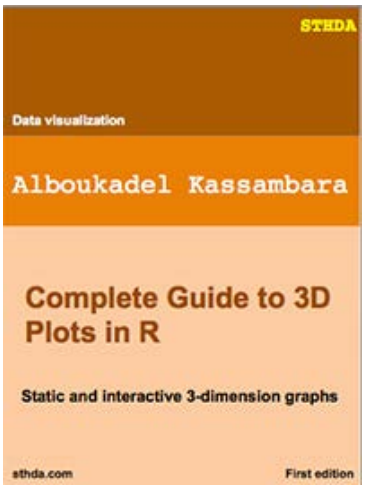


Machine Learning Essentials: Practical Guide in R

Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R



Blogroll

- R-Bloggers

1. Home
2. Articles
3. Cluster Analysis in R: Practical Guide
4. Partitioning Clustering Essentials
5. K-Means Clustering Essentials

Articles - Partitioning Clustering Essentials

K-Means Clustering Essentials

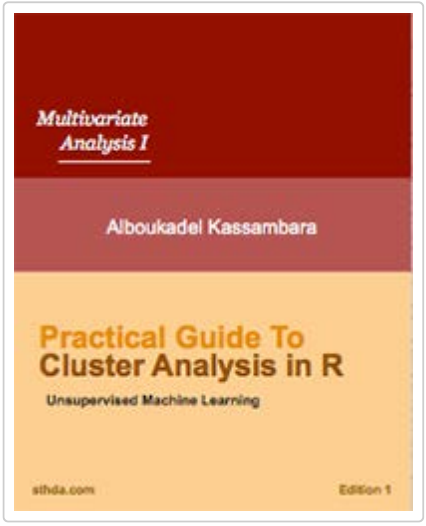
kassambara | 04/09/2017 | 10778 | Comments (7) | Partitioning Clustering Essentials | Unsupervised machine learning, Multivariate Analysis, Genomic data visualization

K-means clustering (MacQueen 1967) is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e. *k clusters*), where k represents the number of groups pre-specified by the analyst. It classifies objects in multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible (i.e., high *intra-class similarity*), whereas objects from different clusters are as dissimilar as possible (i.e., low *inter-class similarity*). In k-means clustering, each cluster is represented by its center (i.e, *centroid*) which corresponds to the mean of points assigned to the cluster.

In this article, we'll describe the **k-means algorithm** and provide practical examples using **R** software.

Contents:

Related Books:



Practical Guide to Cluster Analysis in R

K-means basic ideas

The basic idea behind k-means clustering consists of defining clusters so that the total intra-cluster variation (known as total within-cluster variation) is minimized.

There are several k-means algorithms available. The standard algorithm is the Hartigan-Wong algorithm (Hartigan and Wong 1979), which defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

- x_i design a data point belonging to the cluster C_k
- μ_k is the mean value of the points assigned to the cluster C_k

Each observation (x_i) is assigned to a given cluster such that the sum of squares (SS) distance of the observation to their assigned cluster centers μ_k is a minimum.

We define the total within-cluster variation as follow:

$$tot.withinss = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

The *total within-cluster sum of square* measures the compactness (i.e *goodness*) of the clustering and we want it to be as small as possible.

K-means algorithm

The first step when using k-means clustering is to indicate the number of clusters (k) that will be generated in the final solution.

The algorithm starts by randomly selecting k objects from the data set to serve as the initial centers for the clusters. The selected objects are also known as cluster means or centroids.

Next, each of the remaining objects is assigned to it's closest centroid, where closest is defined using the Euclidean distance (Chapter @ref(clustering-distance-measures) between the object and the cluster mean. This step is called "cluster assignment step". Note that, to use correlation distance, the data are input as z-scores.

After the assignment step, the algorithm computes the new mean value of each cluster. The term cluster "centroid update" is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it might be closer to a different cluster. All the objects are reassigned again using the updated cluster means.

The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e until *convergence* is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration.

K-means algorithm can be summarized as follow:

1. Specify the number of clusters (K) to be created (by the analyst)
2. Select randomly k objects from the dataset as the initial cluster centers or means
3. Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid
4. For each of the k clusters update the *cluster centroid* by calculating the new mean values of all the data points in the cluster. The centoid of a K_{th} cluster is a vector of length p containing the means of all variables for the observations in the k_{th} cluster; p is the number of variables.
5. Iteratively minimize the total within sum of square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the **R** software uses 10 as the default value for the maximum number of iterations.

Computing k-means clustering in R

Data

We'll use the demo data sets "USArrests". The data should be prepared as described in chapter @ref(data-preparation-and-r-packages). The data must contains only continuous variables, as the k-means algorithm uses variable means. As we don't want the k-means algorithm to depend to an arbitrary variable unit, we start by scaling the data using the R function *scale()* as follow:

```
data("USArrests")      # Loading the data set
df <- scale(USArrests) # Scaling the data
# View the firt 3 rows of the data
head(df, n = 3)
```

```
##            Murder Assault UrbanPop     Rape
## Alabama 1.2426   0.783   -0.521 -0.00342
## Alaska  0.5079   1.107   -1.212  2.48420
## Arizona 0.0716   1.479    0.999  1.04288
```

Required R packages and functions

The standard R function for k-means clustering is *kmeans*() [*stats* package], which simplified format is as follow:

```
kmeans(x, centers, iter.max = 10, nstart = 1)
```

- **x**: numeric matrix, numeric data frame or a numeric vector

centers: Possible values are the number of clusters (k) or a set of initial (distinct) cluster centers. If a number, a random set of (distinct) rows in x is chosen as the initial centers.
- **iter.max**: The maximum number of iterations allowed. Default value is 10.
- **nstart**: The number of random starting partitions when centers is a number. Trying nstart > 1 is often recommended.

To create a beautiful graph of the clusters generated with the *kmeans*() function, will use the *factoextra* package.

- Installing *factoextra* package as:

```
install.packages("factoextra")
```

- Loading *factoextra*:

```
library(factoextra)
```

Estimating the optimal number of clusters

The k-means clustering requires the users to specify the number of clusters to be generated.

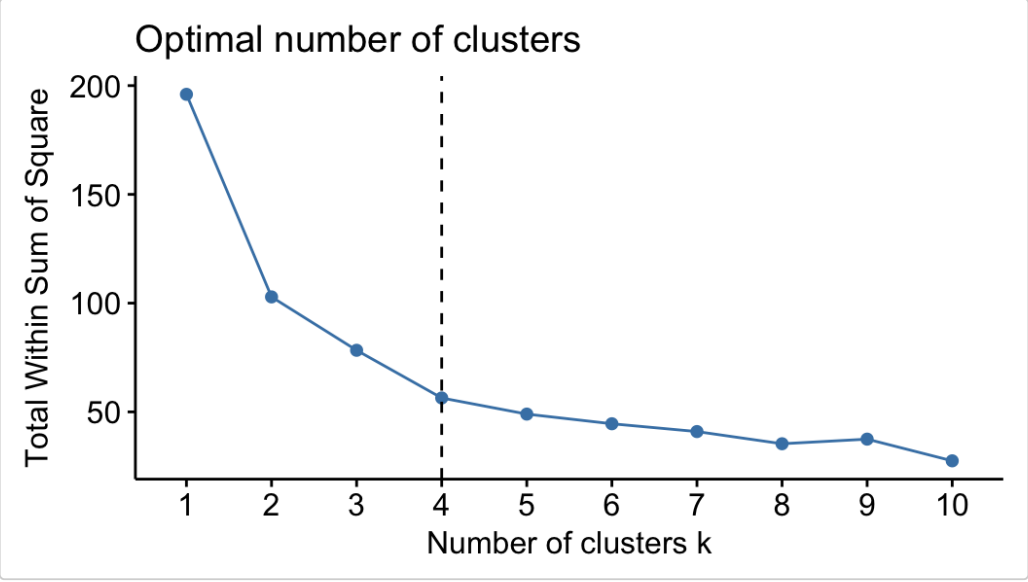
One fundamental question is: How to choose the right number of expected clusters (k)?

Different methods will be presented in the chapter “cluster evaluation and validation statistics”.

Here, we provide a simple solution. The idea is to compute k-means clustering using different values of clusters k. Next, the wss (within sum of square) is drawn according to the number of clusters. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

The R function *fviz_nbclust*() [in *factoextra* package] provides a convenient solution to estimate the optimal number of clusters.

```
library(factoextra)
fviz_nbclust(df, kmeans, method = "wss") +
    geom_vline(xintercept = 4, linetype = 2)
```



The plot above represents the variance within the clusters. It decreases as k increases, but it can be seen a bend (or

"elbow") at k = 4. This bend indicates that additional clusters beyond the fourth have little value.. In the next section, we'll classify the observations into 4 clusters.

Computing k-means clustering

As k-means clustering algorithm starts with k randomly selected centroids, it's always recommended to use the *set.seed()* function in order to set a seed for *R's random number generator*. The aim is to make reproducible the results, so that the reader of this article will obtain exactly the same results as those shown below.

The R code below performs *k-means clustering* with k = 4:

```
# Compute k-means with k = 4
set.seed(123)
km.res <- kmeans(df, 4, nstart = 25)
```

As the final result of k-means clustering result is sensitive to the random starting assignments, we specify *nstart = 25*. This means that R will try 25 different random starting assignments and then select the best results corresponding to the one with the lowest within cluster variation. The default value of *nstart* in R is one. But, it's strongly recommended to compute *k-means clustering* with a large value of *nstart* such as 25 or 50, in order to have a more stable result.

```
# Print the results
print(km.res)
```

```
## K-means clustering with 4 clusters of sizes 13, 16, 13, 8
##
## Cluster means:
##   Murder Assault UrbanPop    Rape
## 1 -0.962  -1.107   -0.930 -0.9668
## 2 -0.489  -0.383    0.576 -0.2617
## 3  0.695   1.039    0.723  1.2769
## 4  1.412   0.874   -0.815  0.0193
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              4              3              3              4              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              2              2              3              4
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              1              3              2              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              1              4              1              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              3              1              4              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              3              1              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              3              4              1              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              2              2              4
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              4              3              2              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              1              1              2
##
## Within cluster sum of squares by cluster:
## [1] 11.95 16.21 19.92  8.32
##  (between_SS / total_SS =  71.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

The printed output displays:

the cluster means or centers: a matrix, which rows are cluster number (1 to 4) and columns are variables
- the clustering vector: A vector of integers (from 1:k) indicating the cluster to which each point is allocated

It's possible to compute the mean of each variables by clusters using the original data:

```
aggregate(USArrests, by=list(cluster=km.res$cluster), mean)
```

```
##   cluster Murder Assault UrbanPop Rape
## 1       1   3.60    78.5     52.1 12.2
## 2       2   5.66   138.9     73.9 18.8
## 3       3  10.82   257.4     76.0 33.2
## 4       4  13.94   243.6     53.8 21.4
```

If you want to add the point classifications to the original data, use this:

```
dd <- cbind(USArrests, cluster = km.res$cluster)
head(dd)
```

```
##            Murder Assault UrbanPop Rape cluster
## Alabama      13.2     236       58 21.2       4
## Alaska       10.0     263       48 44.5       3
## Arizona       8.1     294       80 31.0       3
## Arkansas      8.8     190       50 19.5       4
## California    9.0     276       91 40.6       3
## Colorado      7.9     204       78 38.7       3
```

Accessing to the results of kmeans() function

kmeans() function returns a list of components, including:

- **cluster**: A vector of integers (from 1:k) indicating the cluster to which each point is allocated
- **centers**: A matrix of cluster centers (cluster means)
- **totss**: The total sum of squares (TSS), i.e $\sum (x_i - \bar{x})^2$. TSS measures the total variance in the data.
- **withinss**: Vector of within-cluster sum of squares, one component per cluster
- **tot.withinss**: Total within-cluster sum of squares, i.e. $sum(withinss)$
- **betweenss**: The between-cluster sum of squares, i.e. $totss - tot.withinss$
- **size**: The number of observations in each cluster

These components can be accessed as follow:

```
# Cluster number for each of the observations
km.res$cluster
```

```
head(km.res$cluster, 4)
```

```
##  Alabama   Alaska  Arizona Arkansas
##        4        3        3        4
```

.....

```
# Cluster size
km.res$size
```

```
## [1] 13 16 13  8
```

```
# Cluster means
km.res$centers
```

```
##   Murder Assault UrbanPop    Rape
## 1 -0.962  -1.107   -0.930 -0.9668
## 2 -0.489  -0.383    0.576 -0.2617
## 3  0.695   1.039    0.723  1.2769
## 4  1.412   0.874   -0.815  0.0193
```

Visualizing k-means clusters

It is a good idea to plot the cluster results. These can be used to assess the choice of the number of clusters as well as comparing two different cluster analyses.

Now, we want to visualize the data in a scatter plot with coloring each data point according to its cluster assignment.

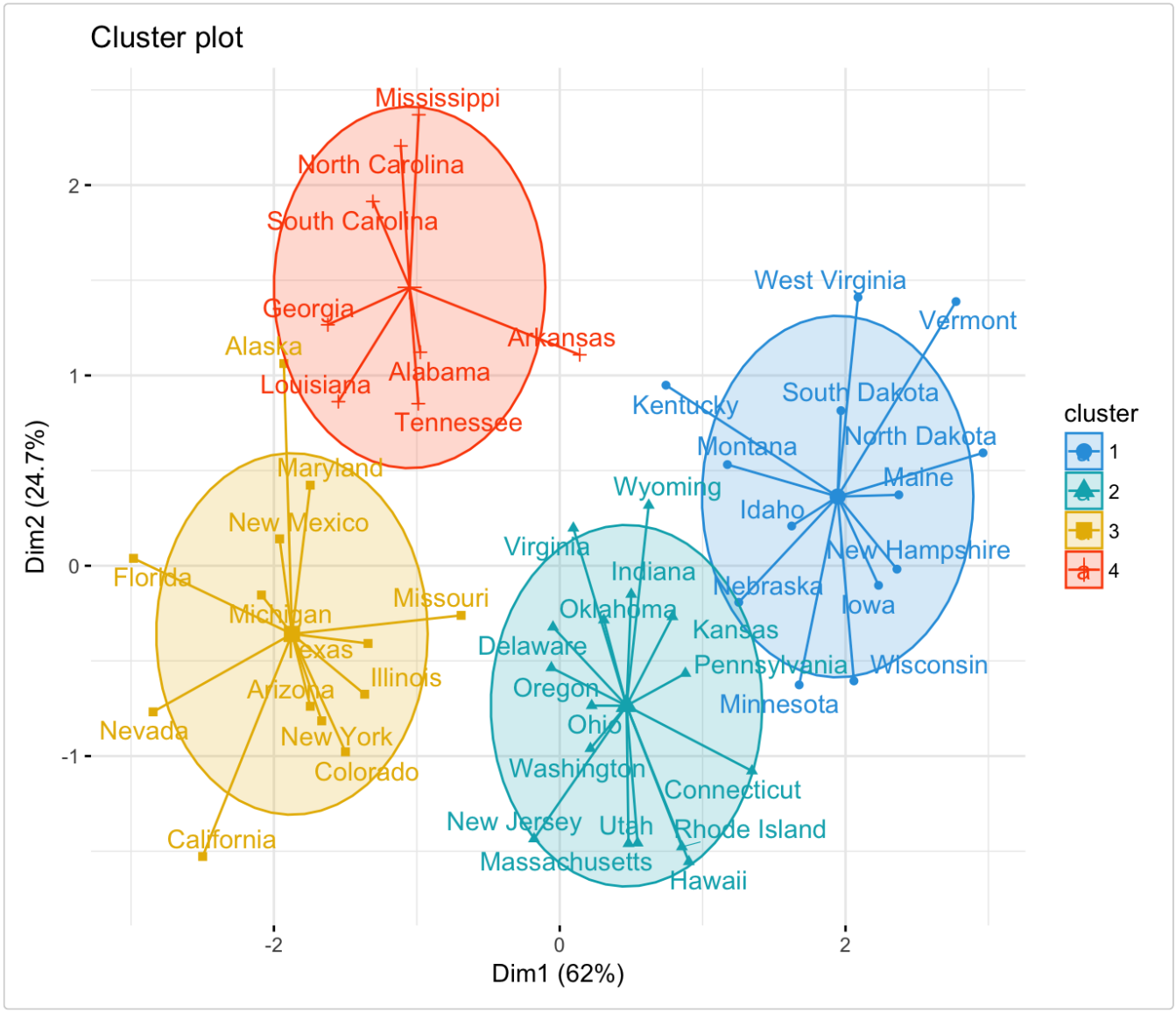
The problem is that the data contains more than 2 variables and the question is what variables to choose for the xy scatter plot.

A solution is to reduce the number of dimensions by applying a dimensionality reduction algorithm, such as **Principal Component Analysis (PCA)**, that operates on the four variables and outputs two new variables (that represent the original variables) that you can use to do the plot.

In other words, if we have a multi-dimensional data set, a solution is to perform Principal Component Analysis (PCA) and to plot data points according to the first two principal components coordinates. }

The function *fviz_cluster*() [*factoextra* package] can be used to easily visualize k-means clusters. It takes k-means results and the original data as arguments. In the resulting plot, observations are represented by points, using principal components if the number of variables is greater than 2. It's also possible to draw concentration ellipse around each cluster.

```
fviz_cluster(km.res, data = df,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
             ellipse.type = "euclid", # Concentration ellipse
             star.plot = TRUE, # Add segments from centroids to items
             repel = TRUE, # Avoid label overplotting (slow)
             ggtheme = theme_minimal()
             )
```



K-means clustering advantages and disadvantages

K-means clustering is very simple and fast algorithm. It can efficiently deal with very large data sets. However there are some weaknesses, including:

1. It assumes prior knowledge of the data and requires the analyst to choose the appropriate number of cluster (k) in advance
2. The final results obtained is sensitive to the initial random selection of cluster centers. Why is it a problem? Because, for every different run of the algorithm on the same dataset, you may choose different set of initial centers. This may lead to different clustering results on different runs of the algorithm.
3. It's sensitive to outliers.
4. If you rearrange your data, it's very possible that you'll get a different solution every time you change the ordering of your data.

Possible solutions to these weaknesses, include:

1. Solution to issue 1: Compute k-means for a range of k values, for example by varying k between 2 and 10. Then, choose the best k by comparing the clustering results obtained for the different k values.
2. Solution to issue 2: Compute K-means algorithm several times with different initial cluster centers. The run with the lowest total within-cluster sum of square is selected as the final clustering solution.
3. To avoid distortions caused by excessive outliers, it's possible to use PAM algorithm, which is less sensitive to outliers.

Alternative to k-means clustering

A robust alternative to k-means is PAM, which is based on medoids. As discussed in the next chapter, the PAM clustering can be computed using the function *pam*() [*cluster* package]. The function *pamk*( ) [fpc package] is a wrapper for PAM that also prints the suggested number of clusters based on optimum average silhouette width.

Summary

K-means clustering can be used to classify observations into k groups, based on their similarity. Each group is represented by the mean value of points in the group, known as the cluster centroid.

K-means algorithm requires users to specify the number of cluster to generate. The R function *kmeans*() [*stats* package] can be used to compute k-means algorithm. The simplified format is kmeans(x, centers), where "x" is the data and centers is the number of clusters to be produced.

After, computing k-means clustering, the R function *fviz_cluster*() [*factoextra* package] can be used to visualize the results. The format is fviz_cluster(km.res, data), where km.res is k-means results and data corresponds to the original data sets.

References

Hartigan, JA, and MA Wong. 1979. "Algorithm AS 136: A K-means clustering algorithm." *Applied Statistics*. Royal Statistical Society, 100–108.

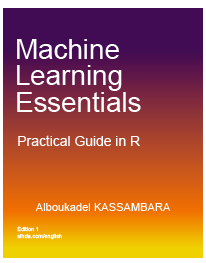
MacQueen, J. 1967. "Some Methods for Classification and Analysis of Multivariate Observations." In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–97. Berkeley, Calif.: University of California Press. http://projecteuclid.org:443/euclid.bsmsp/1200512992.

Last update : 24/09/2017
□ □ □ □ □ 0 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

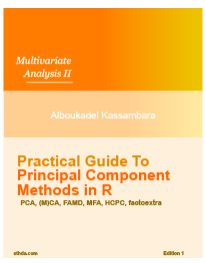
Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!

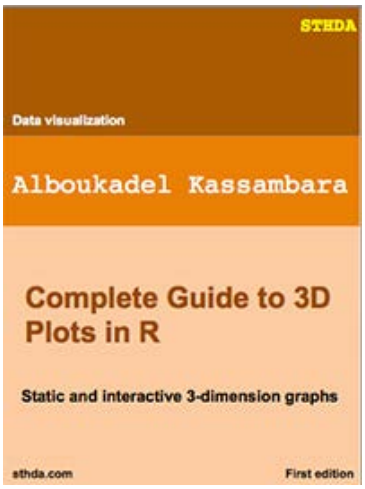


Machine Learning Essentials:

Practical Guide to Cluster



Practical Guide to Principal



Guest Book

Very helpful guide for using the xlsx package.

R.H.

By *Visitor*

Guest Book

Blogroll

- R-Bloggers

Articles - Hierarchical Clustering Essentials

Divisive Hierarchical Clustering Essentials

kassambara | 06/09/2017 | 2827 | Post a comment | Hierarchical Clustering Essentials | Unsupervised machine learning, Multivariate Analysis

The **divisive hierarchical clustering**, also known as *DIANA* (*DIvisive ANAlysis*) is the inverse of agglomerative clustering (Chapter @ref(agglomerative-clustering)).

This article introduces the divisive clustering algorithms and provides practical examples showing how to compute divise clustering using R.

Algorithm

It starts by including all objects in a single large cluster. At each step of iteration, the most heterogeneous cluster is divided into two. The process is iterated until all objects are in their own cluster.

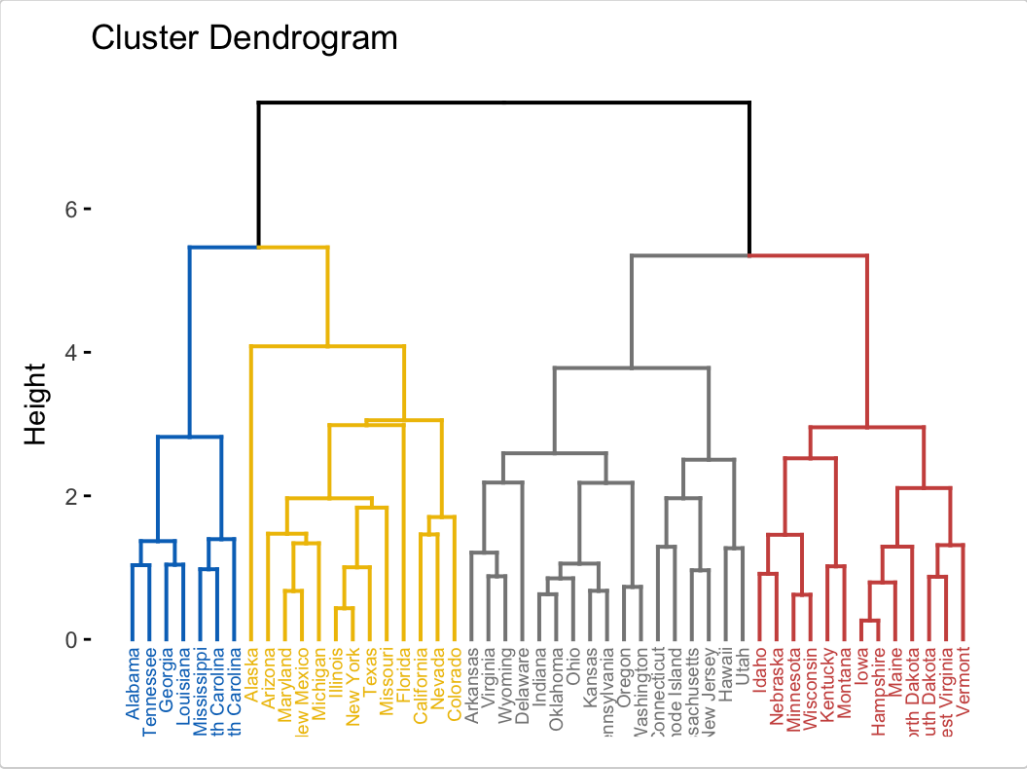
Recall that, divisive clustering is good at identifying large clusters while agglomerative clustering is good at identifying small clusters.

Computation

The R function **diana**() [*cluster* package] can be used to compute divisive clustering. It returns an object of class “diana” (see ?diana.object) which has also methods for the functions: print(), summary(), plot(), pltree(), as.dendrogram(), as.hclust() and cutree().

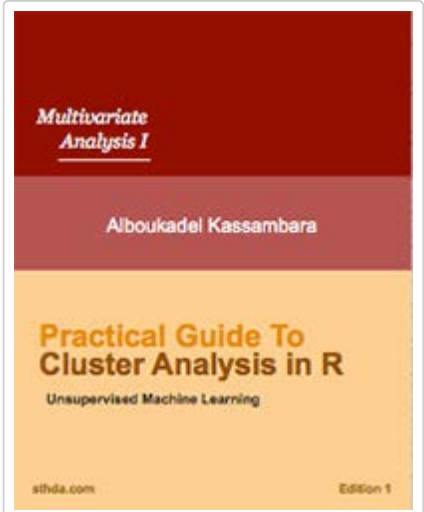
The output of DIANA can be visualized as *dendrograms* using the function *fviz_dend*() [*factoextra* package]. For example, the following R code shows how to computes and visualize divise clustering:

```
# Compute diana()
library(cluster)
res.diana <- diana(USArrests, stand = TRUE)
# Plot the dendrogram
library(factoextra)
fviz_dend(res.diana, cex = 0.5,
          k = 4, # Cut in four groups
          palette = "jco" # Color palette
          )
```



For interpreting dendrograms, read the “agglomerative clustering” chapter.

Related Book:



Practical Guide to Cluster

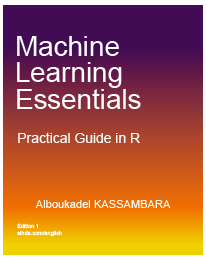
Analysis in R

Last update : 24/09/2017
0 Note

Enjoyed this article? Give us 5 stars (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

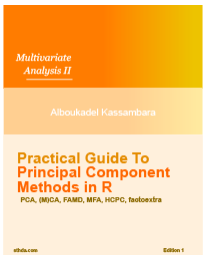
Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!



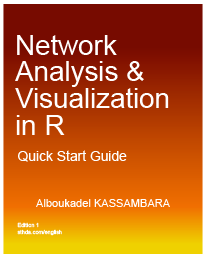
Machine Learning Essentials: Practical Guide in R

Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R

R Graphics Essentials for Great Data Visualization



Network Analysis and Visualization in R

More books on R and data science

Comments

The fields marked with a * are required !

Add a comment
Name
Visitor

By *Visitor*

Articles - Hierarchical Clustering Essentials

Heatmap - Static and Interactive: Absolute Guide

kassambara | 06/09/2017 | 10065 | Comments (7) | Hierarchical Clustering Essentials | Unsupervised machine learning, Multivariate Analysis, Genomic data visualization

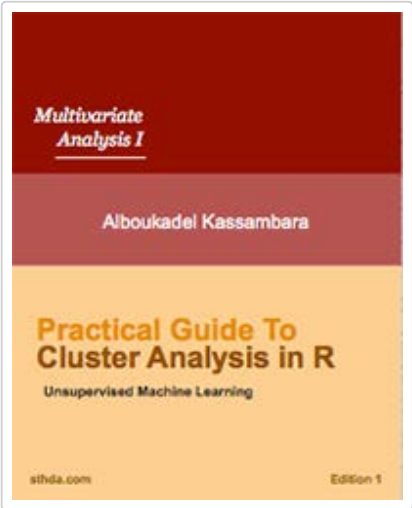
A **heatmap** (or **heat map**) is another way to visualize hierarchical clustering. It's also called a false colored image, where data values are transformed to color scale.

Heat maps allow us to simultaneously visualize clusters of samples and features. First hierarchical clustering is done of both the rows and the columns of the data matrix. The columns/rows of the data matrix are re-ordered according to the hierarchical clustering result, putting similar observations close to each other. The blocks of 'high' and 'low' values are adjacent in the data matrix. Finally, a color scheme is applied for the visualization and the data matrix is displayed. Visualizing the data matrix in this way can help to find the variables that appear to be characteristic for each sample cluster.

Previously, we described how to visualize dendrograms. Here, we'll demonstrate how to draw and arrange a heatmap in R.

Contents:

Related Book:



Practical Guide to Cluster Analysis in R

R Packages/functions for drawing heatmaps

There are a multiple numbers of R packages and functions for drawing interactive and static heatmaps, including:

- *heatmap*() [R base function, *stats* package]: Draws a simple heatmap
- *heatmap.2*() [*gplots* R package]: Draws an enhanced heatmap compared to the R base function.
- *pheatmap*() [*pheatmap* R package]: Draws pretty heatmaps and provides more control to change the appearance of heatmaps.
- *d3heatmap*() [*d3heatmap* R package]: Draws an interactive/clickable heatmap
- *Heatmap*() [*ComplexHeatmap* R/Bioconductor package]: Draws, annotates and arranges complex heatmaps (very useful for genomic data analysis)

Here, we start by describing the 5 R functions for drawing heatmaps. Next, we'll focus on the *ComplexHeatmap* package, which provides a flexible solution to arrange and annotate multiple heatmaps. It allows also to visualize the association between different data from different sources.

Data preparation

We use mtcars data as a demo data set. We start by standardizing the data to make variables comparable:

```
df <- scale(mtcars)
```

R base heatmap: heatmap()

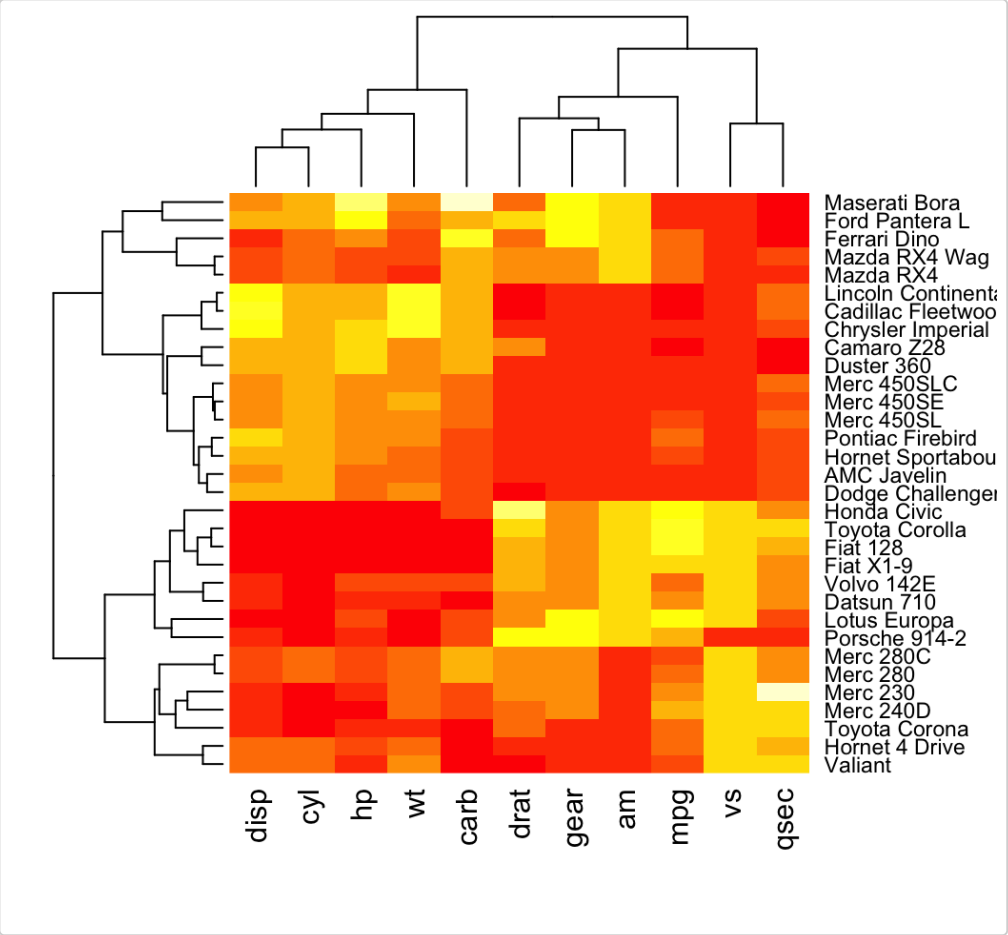
The built-in R *heatmap*() function [in *stats* package] can be used.

A simplified format is:

```
heatmap(x, scale = "row")
```

- **x**: a numeric matrix
- **scale**: a character indicating if the values should be centered and scaled in either the row direction or the column direction, or none. Allowed values are in c("row", "column", "none"). Default is "row".

```
# Default plot
heatmap(df, scale = "none")
```



In the plot above, high values are in red and low values are in yellow.

It's possible to specify a color palette using the argument *col*, which can be defined as follow:

- Using custom colors:

```
col<- colorRampPalette(c("red", "white", "blue"))(256)
```

- Or, using RColorBrewer color palette:

```
library("RColorBrewer")
col <- colorRampPalette(brewer.pal(10, "RdYlBu"))(256)
```

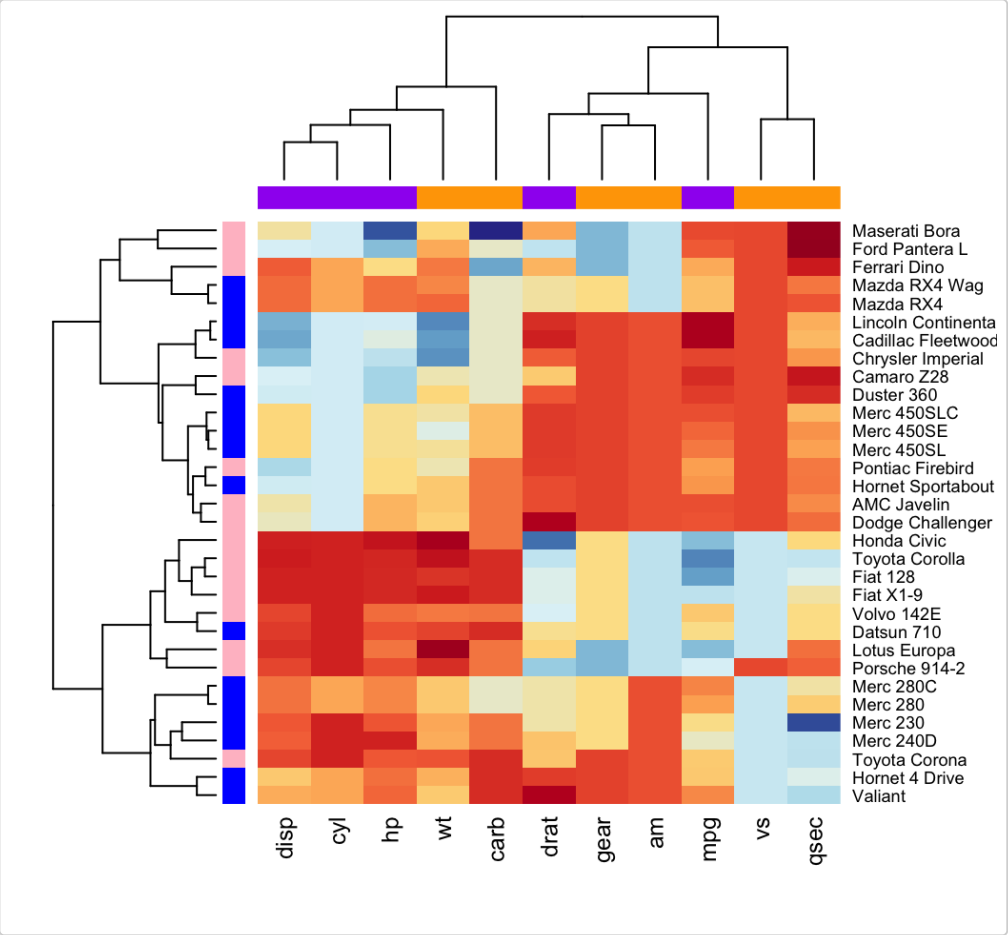
Additionally, you can use the argument *RowSideColors* and *ColSideColors* to annotate rows and columns, respectively.

For example, in the the R code below will customize the heatmap as follow:

1. An RColorBrewer color palette name is used to change the appearance
2. The argument *RowSideColors* and *ColSideColors* are used to annotate rows and columns respectively. The expected values for these options are a vector containing color names specifying the classes for rows/columns.

```
# Use RColorBrewer color palette names
library("RColorBrewer")
col <- colorRampPalette(brewer.pal(10, "RdYlBu"))(256)
heatmap(df, scale = "none", col =  col,
```

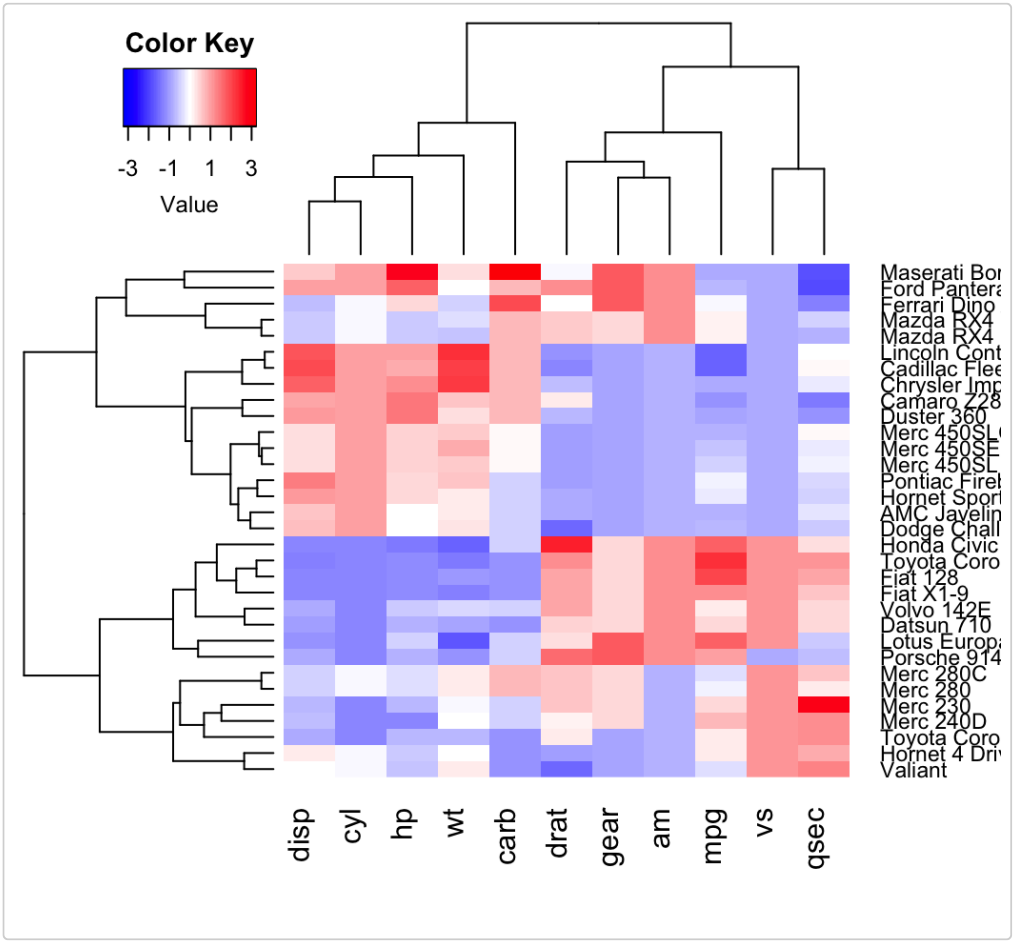
```
        RowSideColors = rep(c("blue", "pink"), each = 16),
        ColSideColors = c(rep("purple", 5), rep("orange", 6)))
```



Enhanced heat maps: heatmap.2()

The function *heatmap.2*() [in *gplots* package] provides many extensions to the standard R *heatmap*() function presented in the previous section.

```
# install.packages("gplots")
library("gplots")
heatmap.2(df, scale = "none", col = bluered(100),
          trace = "none", density.info = "none")
```



Other arguments can be used including:

- *labRow*, *labCol*
- *hclustfun*: hclustfun=function(x) hclust(x, method="ward")

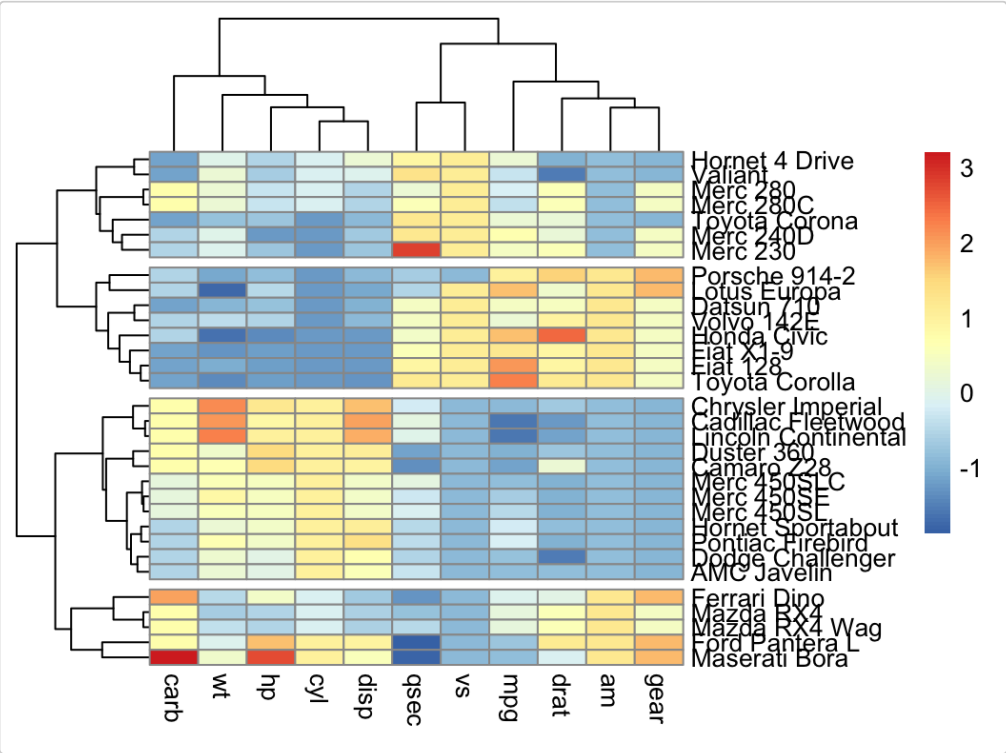
In the R code above, the *bluered*() function [in *gplots* package] is used to generate a smoothly varying set of colors. You can also use the following color generator functions:

- *colorpanel*(n, low, mid, high)
  - *n*: Desired number of color elements to be generated
  - *low*, *mid*, *high*: Colors to use for the Lowest, middle, and highest values. mid may be omitted.
- *redgreen*(n), *greenred*(n), *bluered*(n) and *redblue*(n)

Pretty heat maps: pheatmap()

First, install the *pheatmap* package: install.packages("pheatmap"); then type this:

```
library("pheatmap")
pheatmap(df, cutree_rows = 4)
```

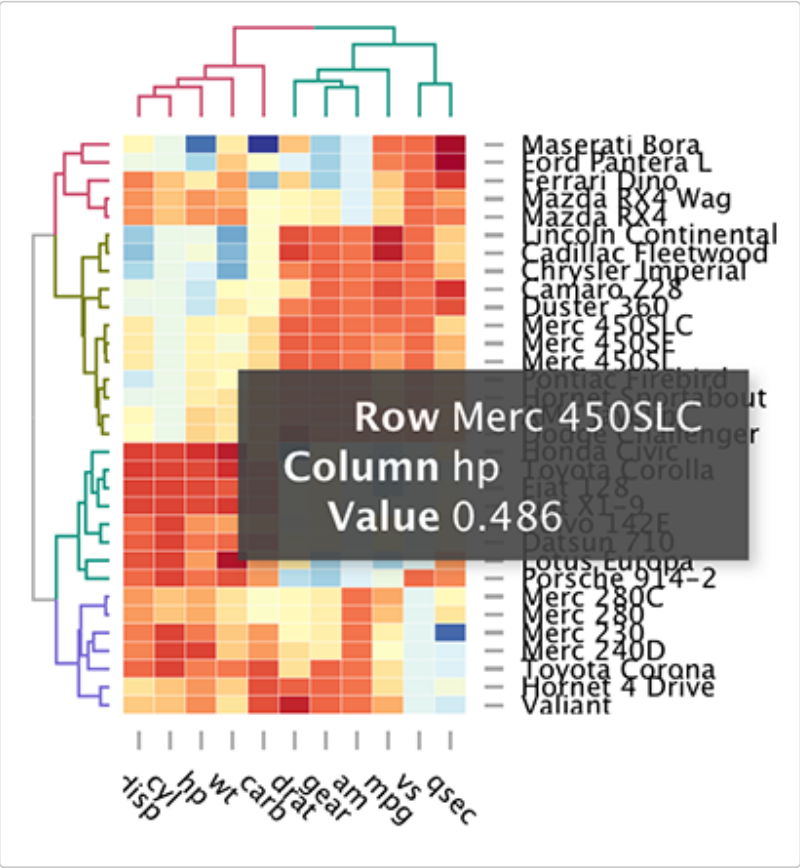


Arguments are available for changing the default clustering metric (“euclidean”) and method (“complete”). It’s also possible to annotate rows and columns using grouping variables.

Interactive heat maps: d3heatmap()

First, install the *d3heatmap* package: install.packages(“d3heatmap”); then type this:

```
library("d3heatmap")
d3heatmap(scale(mtcars), colors = "RdYlBu",
          k_row = 4, # Number of groups in rows
          k_col = 2 # Number of groups in columns
          )
```



The *d3heamap*() function makes it possible to:

- Put the mouse on a heatmap cell of interest to view the row and the column names as well as the corresponding value.
- Select an area for zooming. After zooming, click on the heatmap again to go back to the previous display

Enhancing heatmaps using dendextend

The package *dendextend* can be used to enhance functions from other packages. The *mtcars* data is used in the following sections. We'll start by defining the order and the appearance for rows and columns using dendextend. These results are used in others functions from others packages.

The order and the appearance for rows and columns can be defined as follow:

```
library(dendextend)
# order for rows
Rowv  <- mtcars %>% scale %>% dist %>% hclust %>% as.dendrogram %>%
   set("branches_k_color", k = 3) %>% set("branches_lwd", 1.2) %>%
   ladderize
# Order for columns: We must transpose the data
Colv  <- mtcars %>% scale %>% t %>% dist %>% hclust %>% as.dendrogram %>%
   set("branches_k_color", k = 2, value = c("orange", "blue")) %>%
   set("branches_lwd", 1.2) %>%
   ladderize
```

The arguments above can be used in the functions below:

1. The standard *heatmap*() function [in *stats* package]:

```
heatmap(scale(mtcars), Rowv = Rowv, Colv = Colv,
```

```
        scale = "none")
```

2. The enhanced *heatmap.2*() function [in *gplots* package]:

```
library(gplots)
heatmap.2(scale(mtcars), scale = "none", col = bluered(100),
          Rowv = Rowv, Colv = Colv,
          trace = "none", density.info = "none")
```

3. The interactive heatmap generator *d3heatmap*() function [in *d3heatmap* package]:

```
library("d3heatmap")
d3heatmap(scale(mtcars), colors = "RdBu",
          Rowv = Rowv, Colv = Colv)
```

Complex heatmap

ComplexHeatmap is an R/bioconductor package, developed by Zuguang Gu, which provides a flexible solution to arrange and annotate multiple heatmaps. It allows also to visualize the association between different data from different sources.

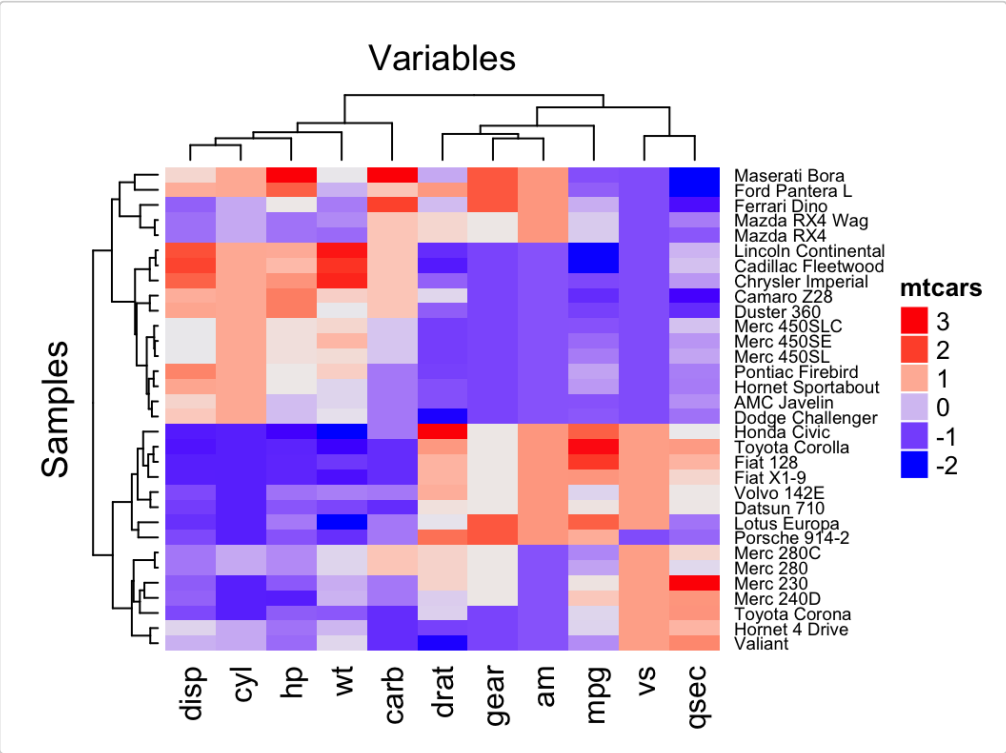
It can be installed as follow:

```
source("https://bioconductor.org/biocLite.R")
biocLite("ComplexHeatmap")
```

Simple heatmap

You can draw a simple heatmap as follow:

```
library(ComplexHeatmap)
Heatmap(df,
        name = "mtcars", #title of legend
        column_title = "Variables", row_title = "Samples",
        row_names_gp = gpar(fontsize = 7) # Text size for row names
        )
```



Additional arguments:

1. show_row_names, show_column_names: whether to show row and column names, respectively. Default value is TRUE
2. show_row_hclust, show_column_hclust: logical value; whether to show row and column clusters. Default is TRUE
3. clustering_distance_rows, clustering_distance_columns: metric for clustering: “euclidean”, “maximum”, “manhattan”, “canberra”, “binary”, “minkowski”, “pearson”, “spearman”, “kendall”)
4. clustering_method_rows, clustering_method_columns: clustering methods: “ward.D”, “ward.D2”, “single”, “complete”, “average”, … (see **?hclust**).

To specify a custom colors, you must use the the *colorRamp2*() function [*circlize* package], as follow:

```
library(circlize)
mycols <- colorRamp2(breaks = c(-2, 0, 2),
                     colors = c("green", "white", "red"))
Heatmap(df, name = "mtcars", col = mycols)
```

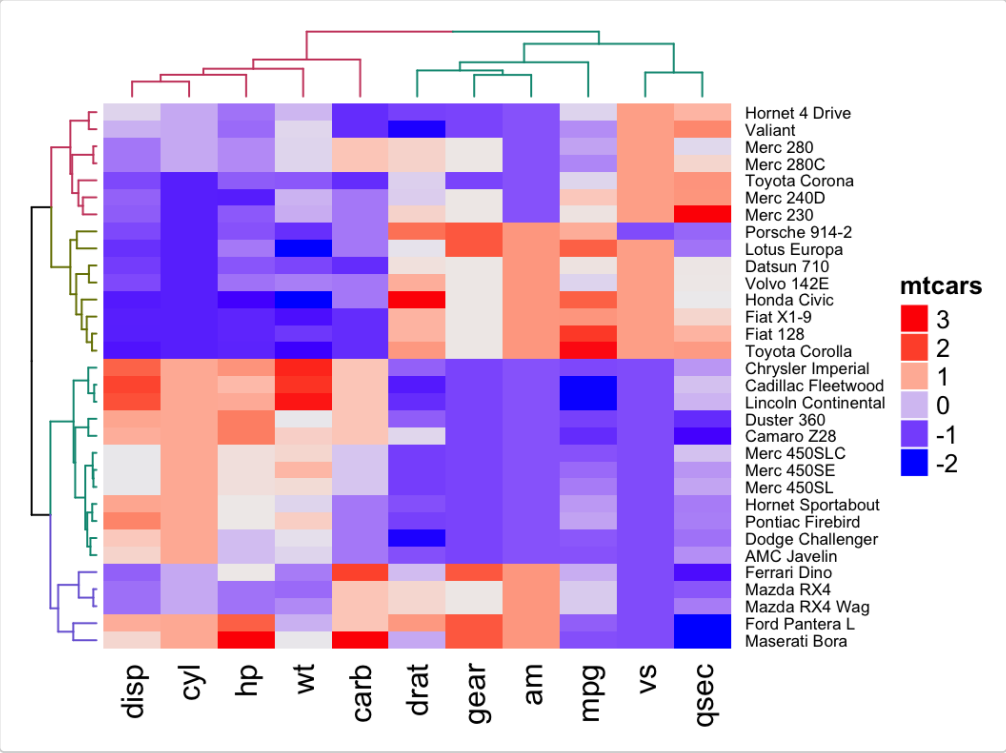
It’s also possible to use **RColorBrewer** color palettes:

```
library("circlize")
library("RColorBrewer")
Heatmap(df, name = "mtcars",
        col = colorRamp2(c(-2, 0, 2), brewer.pal(n=3, name="RdBu")))
```

We can also customize the appearance of dendograms using the function *color_branches*() [*dendextend* package]:

```
library(dendextend)
row_dend = hclust(dist(df)) # row clustering
col_dend = hclust(dist(t(df))) # column clustering
Heatmap(df, name = "mtcars",
        row_names_gp = gpar(fontsize = 6.5),
```

```
        cluster_rows = color_branches(row_dend, k = 4),
        cluster_columns = color_branches(col_dend, k = 2))
```



Splitting heatmap by rows

You can split the heatmap using either the k-means algorithm or a grouping variable.

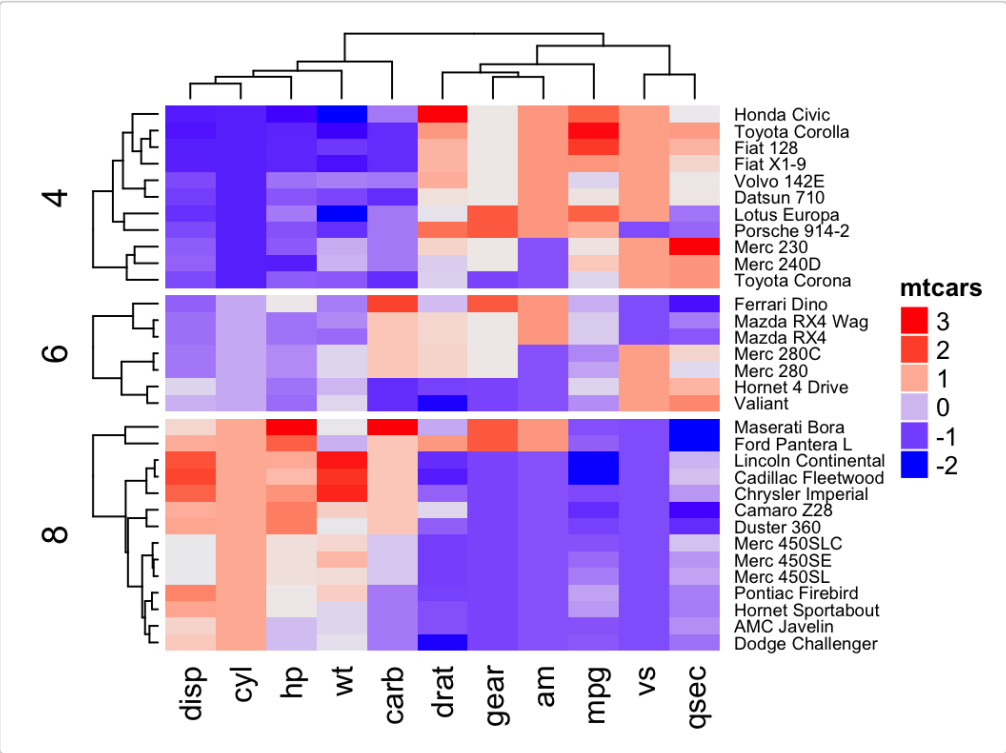
It's important to use the set.seed() function when performing k-means so that the results obtained can be reproduced precisely at a later time.

- To split the dendrogram using k-means, type this:

```
# Divide into 2 groups
set.seed(2)
Heatmap(df, name = "mtcars", k = 2)
```

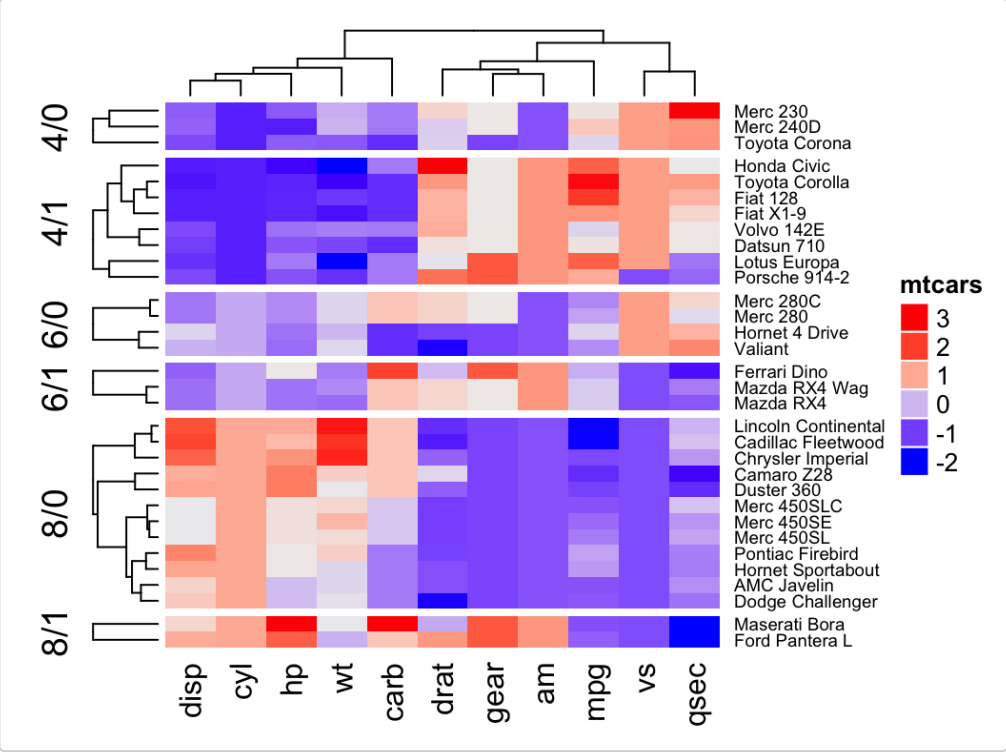
- To split by a grouping variable, use the argument *split*. In the following example we'll use the levels of the factor variable cyl [in mtcars data set] to split the heatmap by rows. Recall that the column cyl corresponds to the number of cylinders.

```
# split by a vector specifying rowgroups
Heatmap(df, name = "mtcars", split = mtcars$cyl,
        row_names_gp = gpar(fontsize = 7))
```



Note that, *split* can be also a data frame in which different combinations of levels split the rows of the heatmap.

```
# Split by combining multiple variables
Heatmap(df, name ="mtcars",
        split = data.frame(cyl = mtcars$cyl, am = mtcars$am),
        row_names_gp = gpar(fontsize = 7))
```



- It's also possible to combine km and split:

```
Heatmap(df, name ="mtcars", col = mycol,
        km = 2, split =  mtcars$cyl)
```

- If you want to use other partitioning method, rather than k-means, you can easily do it by just assigning the partitioning vector to split. In the R code below, we'll use *pam*() function [cluster package]. pam() stands for Partitioning of the data into k clusters "around medoids", a more robust version of K-means.

```
# install.packages("cluster")
library("cluster")
set.seed(2)
pa = pam(df, k = 3)
Heatmap(df, name = "mtcars", col = mycol,
        split = paste0("pam", pa$clustering))
```

Heatmap annotation

The *HeatmapAnnotation* class is used to define annotation on row or column. A simplified format is:

```
HeatmapAnnotation(df, name, col, show_legend)
```

- **df**: a data.frame with column names
- **name**: the name of the heatmap annotation
- **col**: a list of colors which contains color mapping to columns in df

For the example below, we'll transpose our data to have the observations in columns and the variables in rows.

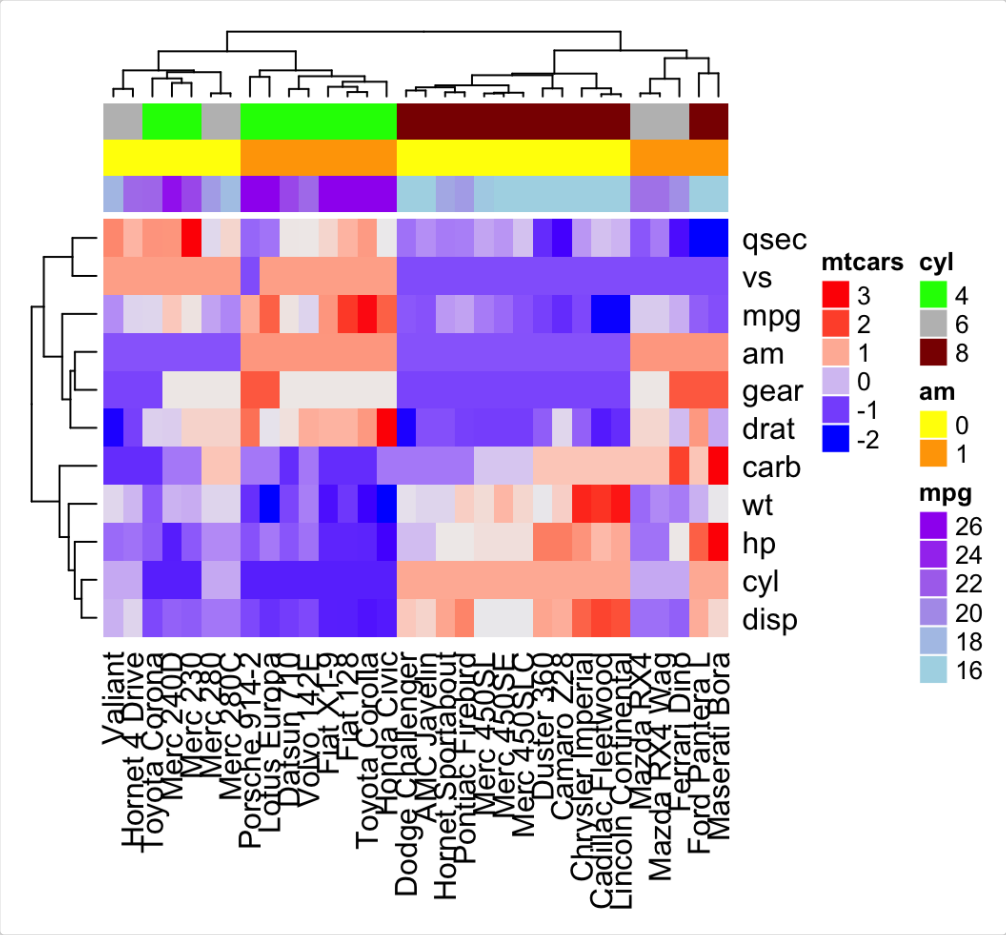
```
df <- t(df)
```

Simple annotation

A vector, containing discrete or continuous values, is used to annotate rows or columns. We'll use the qualitative variables *cyl* (levels = "4", "5" and "8") and *am* (levels = "0" and "1"), and the continuous variable *mpg* to annotate columns.

For each of these 3 variables, custom colors are defined as follow:

```
# Annotation data frame
annot_df <- data.frame(cyl = mtcars$cyl, am = mtcars$am,
                       mpg = mtcars$mpg)
# Define colors for each levels of qualitative variables
# Define gradient color for continuous variable (mpg)
col = list(cyl = c("4" = "green", "6" = "gray", "8" = "darkred"),
           am = c("0" = "yellow", "1" = "orange"),
           mpg = circlize::colorRamp2(c(17, 25),
                                      c("lightblue", "purple")) )
# Create the heatmap annotation
ha <- HeatmapAnnotation(annot_df, col = col)
# Combine the heatmap and the annotation
Heatmap(df, name = "mtcars",
        top_annotation = ha)
```



It's possible to hide the annotation legend using the argument *show_legend = FALSE* as follow:

```
ha <- HeatmapAnnotation(annot_df, col = col, show_legend = FALSE)
Heatmap(df, name = "mtcars", top_annotation = ha)
```

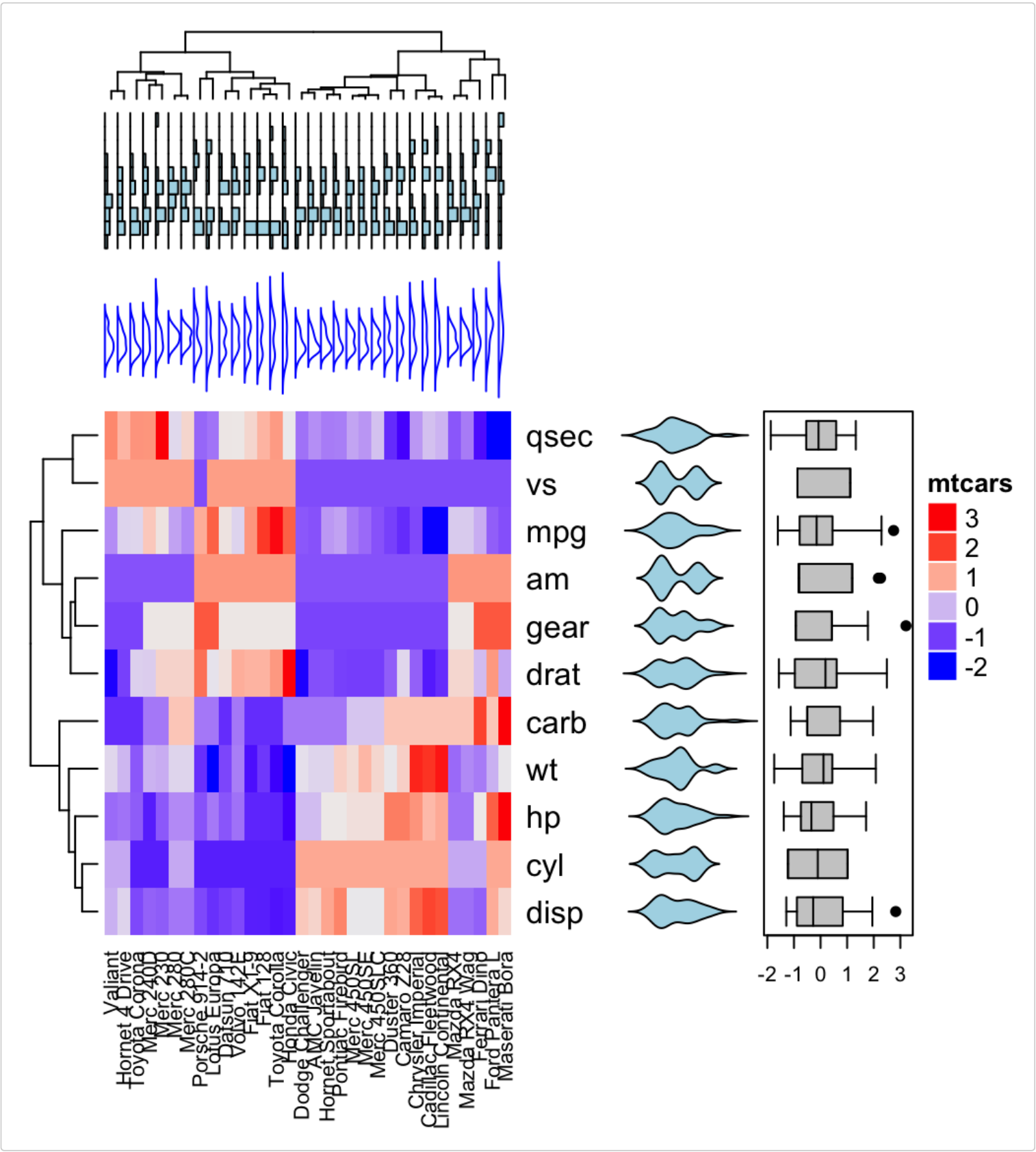
Complex annotation

In this section we'll see how to combine heatmap and some basic graphs to show the data distribution. For simple annotation graphics, the following functions can be used: *anno_points*(), *anno_barplot*(), *anno_boxplot*(), *anno_density*() and *anno_histogram*().

An example is shown below:

```
# Define some graphics to display the distribution of columns
.hist = anno_histogram(df, gp = gpar(fill = "lightblue"))
.density = anno_density(df, type = "line", gp = gpar(col = "blue"))
ha_mix_top = HeatmapAnnotation(hist = .hist, density = .density)
# Define some graphics to display the distribution of rows
.violin = anno_density(df, type = "violin",
                       gp = gpar(fill = "lightblue"), which = "row")
.boxplot = anno_boxplot(df, which = "row")
ha_mix_right = HeatmapAnnotation(violin = .violin, bxplt = .boxplot,
                                 which = "row", width = unit(4, "cm"))
# Combine annotation with heatmap
```

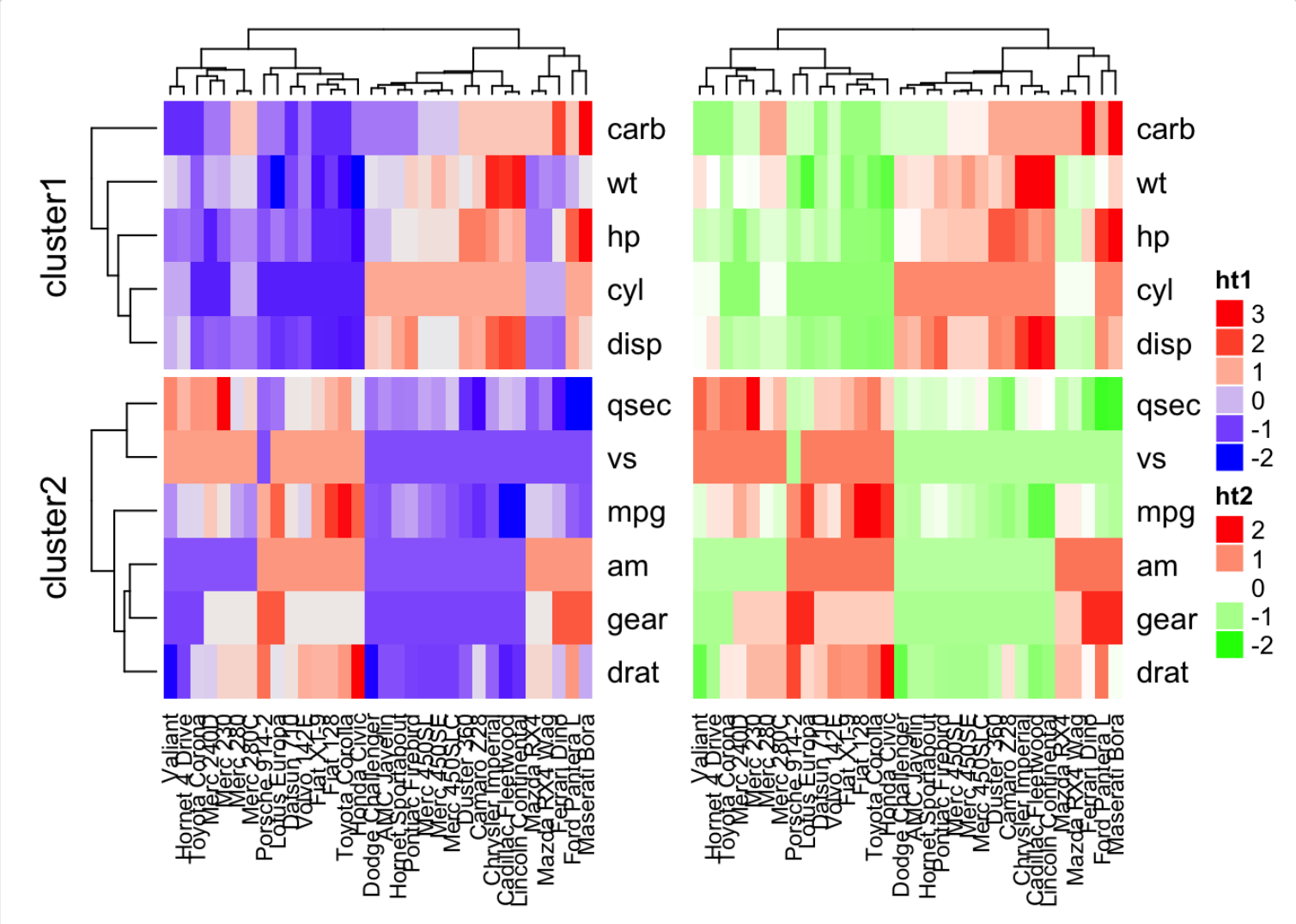
```
Heatmap(df, name = "mtcars",
        column_names_gp = gpar(fontsize = 8),
        top_annotation = ha_mix_top,
        top_annotation_height = unit(3.8, "cm")) + ha_mix_right
```



Combining multiple heatmaps

Multiple heatmaps can be arranged as follow:

```
# Heatmap 1
ht1 = Heatmap(df, name = "ht1", km = 2,
              column_names_gp = gpar(fontsize = 9))
# Heatmap 2
ht2 = Heatmap(df, name = "ht2",
        col = circlize::colorRamp2(c(-2, 0, 2), c("green", "white", "red")),
        column_names_gp = gpar(fontsize = 9))
# Combine the two heatmaps
ht1 + ht2
```



You can use the option width = unit(3, “cm”)) to control the size of the heatmaps.

Note that when combining multiple heatmaps, the first heatmap is considered as the main heatmap. Some settings of the remaining heatmaps are auto-adjusted according to the setting of the main heatmap. These include: removing row clusters and titles, and adding splitting.

The *draw*() function can be used to customize the appearance of the final image:

```
draw(ht1 + ht2,
    row_title = "Two heatmaps, row title",
    row_title_gp = gpar(col = "red"),
    column_title = "Two heatmaps, column title",
```

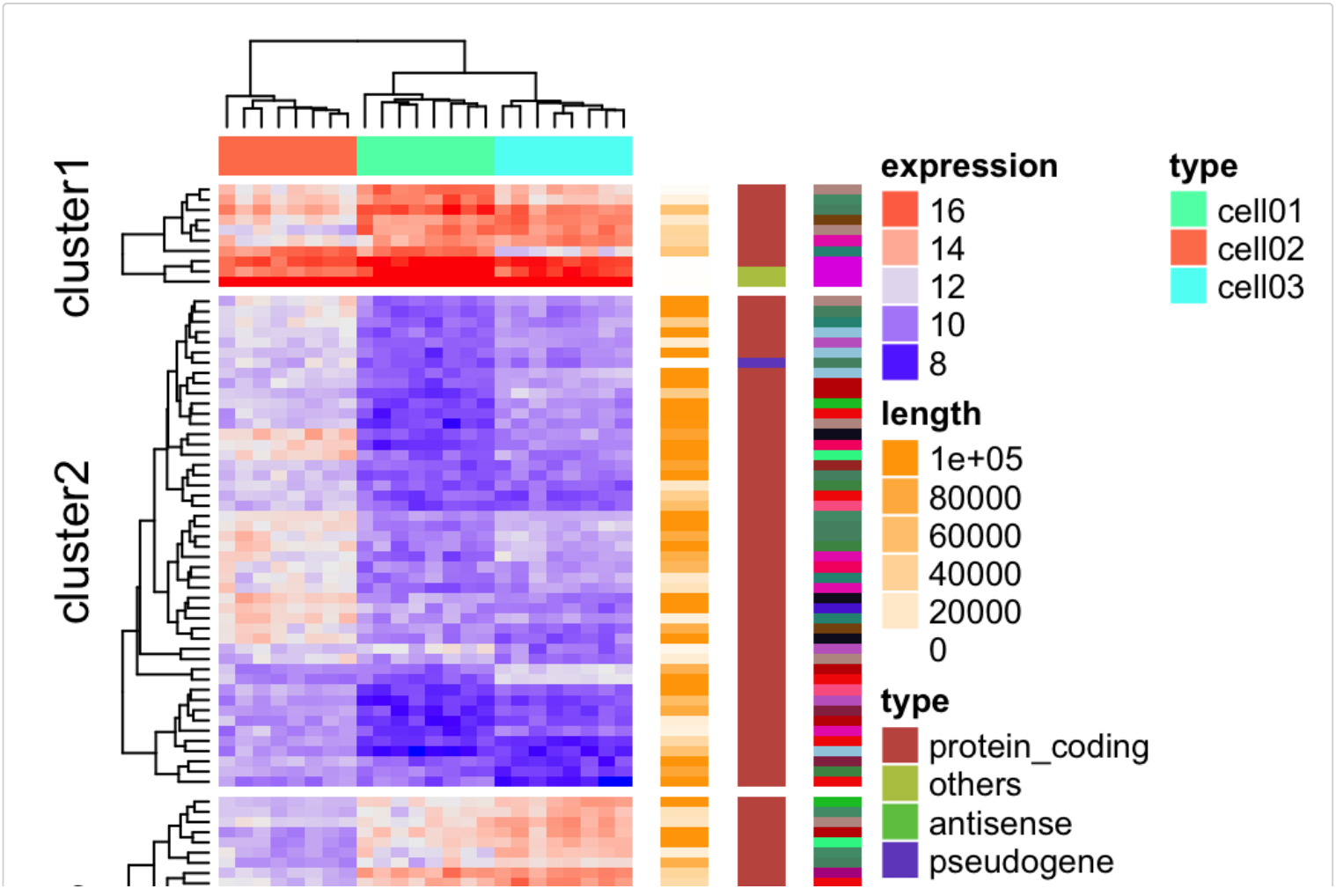
```
    column_title_side = "bottom",
    # Gap between heatmaps
    gap = unit(0.5, "cm"))
```

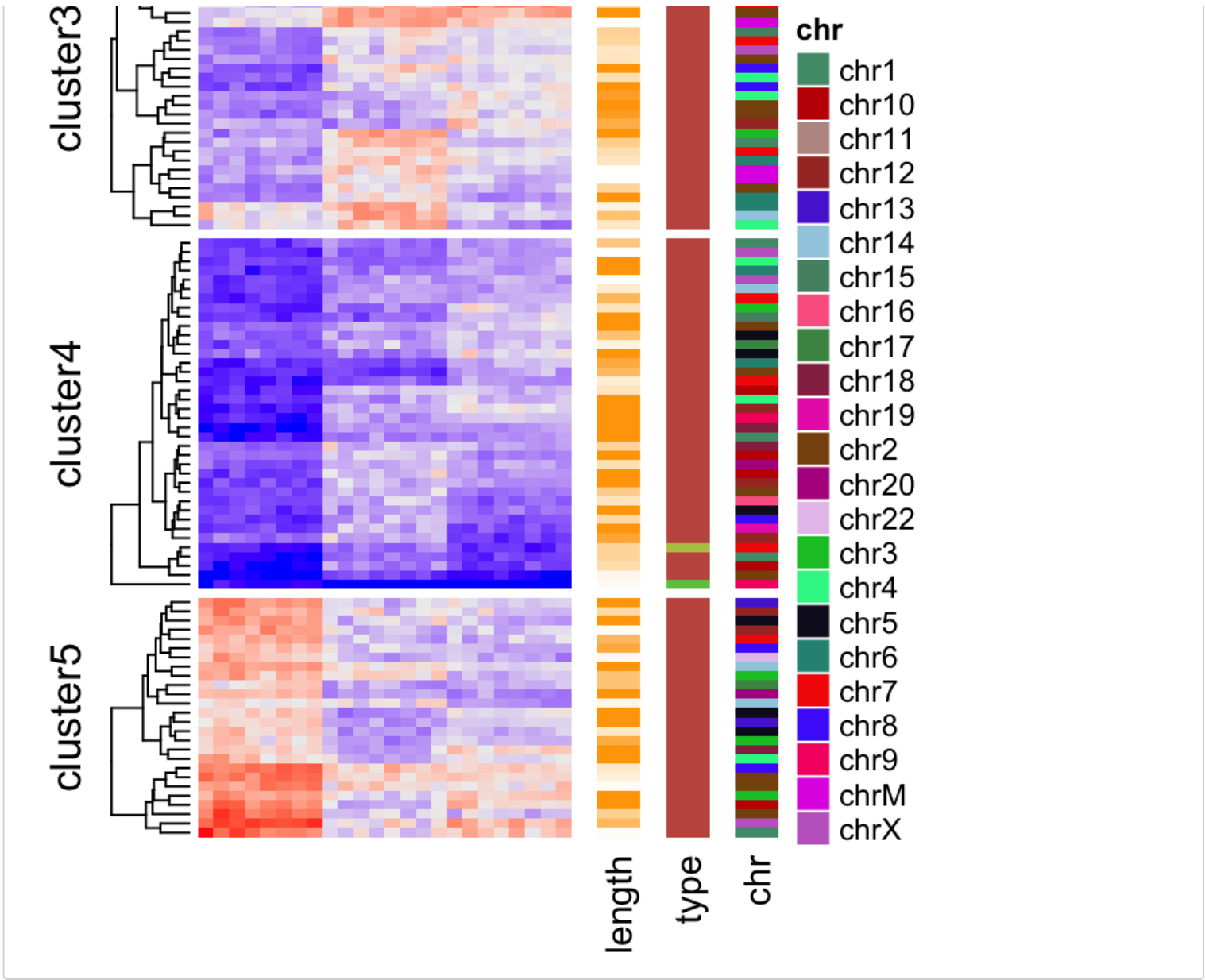
Legends can be removed using the arguments *show_heatmap_legend = FALSE*, *show_annotation_legend = FALSE*.

Application to gene expression matrix

In gene expression data, rows are genes and columns are samples. More information about genes can be attached after the expression heatmap such as gene length and type of genes.

```
expr <- readRDS(paste0(system.file(package = "ComplexHeatmap"),
                       "/extdata/gene_expression.rds"))
mat <- as.matrix(expr[, grep("cell", colnames(expr))])
type <- gsub("s\\d+_", "", colnames(mat))
ha = HeatmapAnnotation(df = data.frame(type = type))
Heatmap(mat, name = "expression", km = 5, top_annotation = ha,
    top_annotation_height = unit(4, "mm"),
    show_row_names = FALSE, show_column_names = FALSE) +
Heatmap(expr$length, name = "length", width = unit(5, "mm"),
    col = circlize::colorRamp2(c(0, 100000), c("white", "orange"))) +
Heatmap(expr$type, name = "type", width = unit(5, "mm")) +
Heatmap(expr$chr, name = "chr", width = unit(5, "mm"),
    col = circlize::rand_color(length(unique(expr$chr))))
```

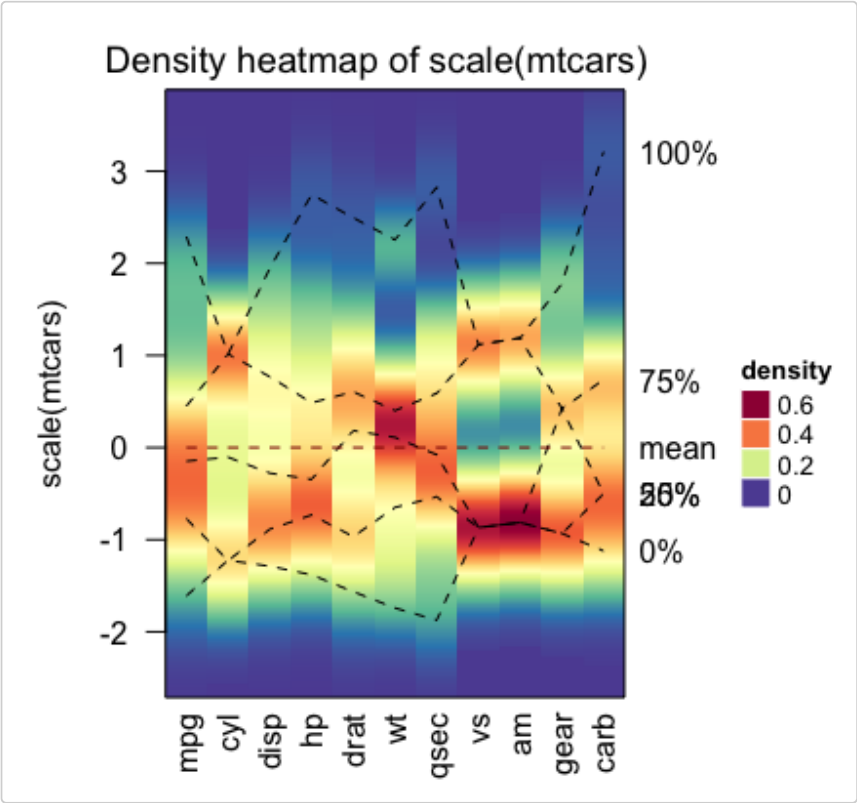




It's also possible to visualize genomic alterations and to integrate different molecular levels (gene expression, DNA methylation, …). Read the vignette, on Bioconductor, for further examples.

Visualizing the distribution of columns in matrix

```
densityHeatmap(scale(mtcars))
```



The dashed lines on the heatmap correspond to the five quantile numbers. The text for the five quantile levels are added in the right of the heatmap.

Summary

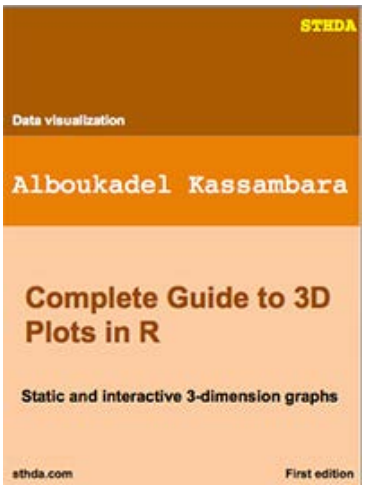
We described many functions for drawing heatmaps in R (from basic to complex heatmaps). A basic heatmap can be produced using either the R base function *heatmap*() or the function *heatmap.2*() [in the *gplots* package]. The *pheatmap*() function, in the package of the same name, creates pretty heatmaps, where ones has better control over some graphical parameters such as cell size.

The *Heatmap*() function [in *ComplexHeatmap* package] allows us to easily, draw, annotate and arrange complex heatmaps. This might be very useful in genomic fields.

Last update : 24/09/2017
□ □ □ □ □ 0 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!



Articles - Hierarchical Clustering Essentials

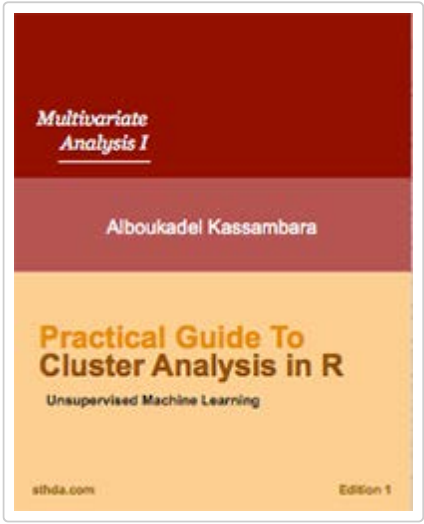
Visualizing Dendrograms: Ultimate Guide

kassambara | 06/09/2017 | 5095 | Post a comment | Hierarchical Clustering Essentials | Unsupervised machine learning, Multivariate Analysis, Genomic data visualization

As described in previous chapters, a **dendrogram** is a tree-based representation of a data created using hierarchical clustering methods (Chapter @ref(agglomerative-clustering)). In this article, we provide R code for **visualizing** and customizing dendrograms. Additionally, we show how to save and to zoom a large dendrogram.

Contents:

The Book:



Practical Guide to Cluster Analysis in R

We start by computing hierarchical clustering using the USArrests data sets:

```
# Load data
data(USArrests)
# Compute distances and hierarchical clustering
dd <- dist(scale(USArrests), method = "euclidean")
hc <- hclust(dd, method = "ward.D2")
```

To visualize the dendrogram, we'll use the following R functions and packages:

- *fviz_dend*()[in factoextra R package] to create easily a ggplot2-based beautiful dendrogram.
- *dendextend* package to manipulate dendrograms

Before continuing, install the required package as follow:

```
install.packages(c("factoextra", "dendextend"))
```

Visualizing dendrograms

We'll use the function *fviz_dend*()[in *factoextra* R package] to create easily a beautiful dendrogram using either the R base plot or ggplot2. It provides also an option for drawing circular dendrograms and phylogenic-like trees.

To create a basic dendrograms, type this:

```
library(factoextra)
fviz_dend(hc, cex = 0.5)
```

You can use the arguments main, sub, xlab, ylab to change plot titles as follow:

```
fviz_dend(hc, cex = 0.5,
          main = "Dendrogram - ward.D2",
          xlab = "Objects", ylab = "Distance", sub = "")
```

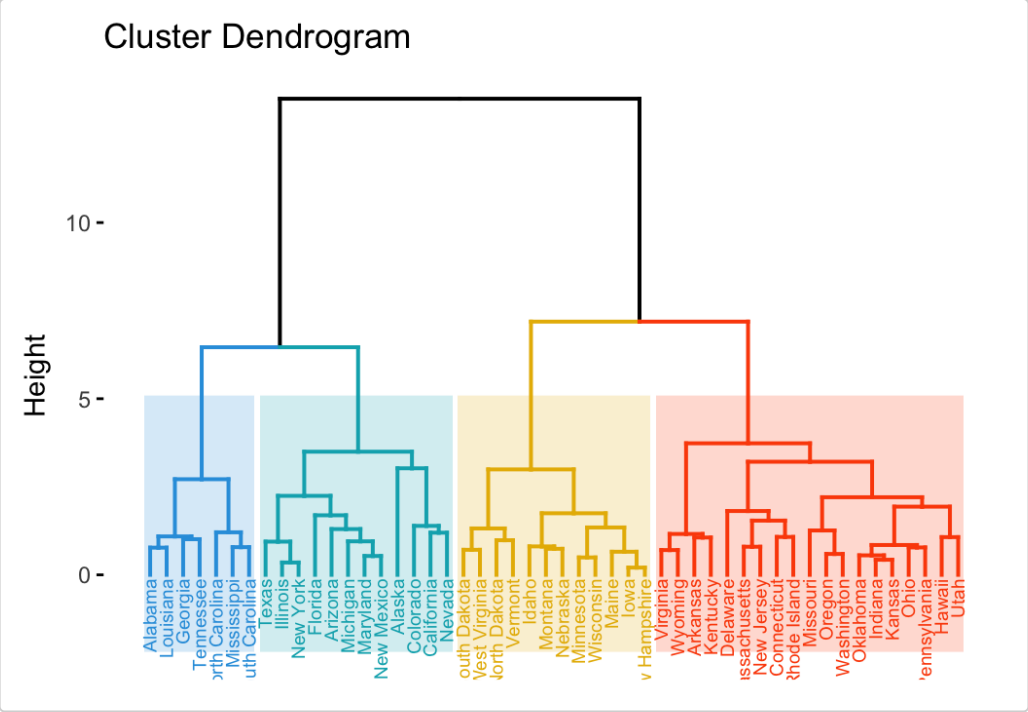
To draw a horizontal dendrogram, type this:

```
fviz_dend(hc, cex = 0.5, horiz = TRUE)
```

It's also possible to cut the tree at a given height for partitioning the data into multiple groups as described in the previous chapter: Hierarchical clustering (Chapter @ref(agglomerative-clustering)). In this case, it's possible to color branches by groups and to add rectangle around each group.

For example:

```
fviz_dend(hc, k = 4, # Cut in four groups
          cex = 0.5, # label size
          k_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
          color_labels_by_k = TRUE, # color labels by groups
          rect = TRUE, # Add rectangle around groups
          rect_border = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
          rect_fill = TRUE)
```



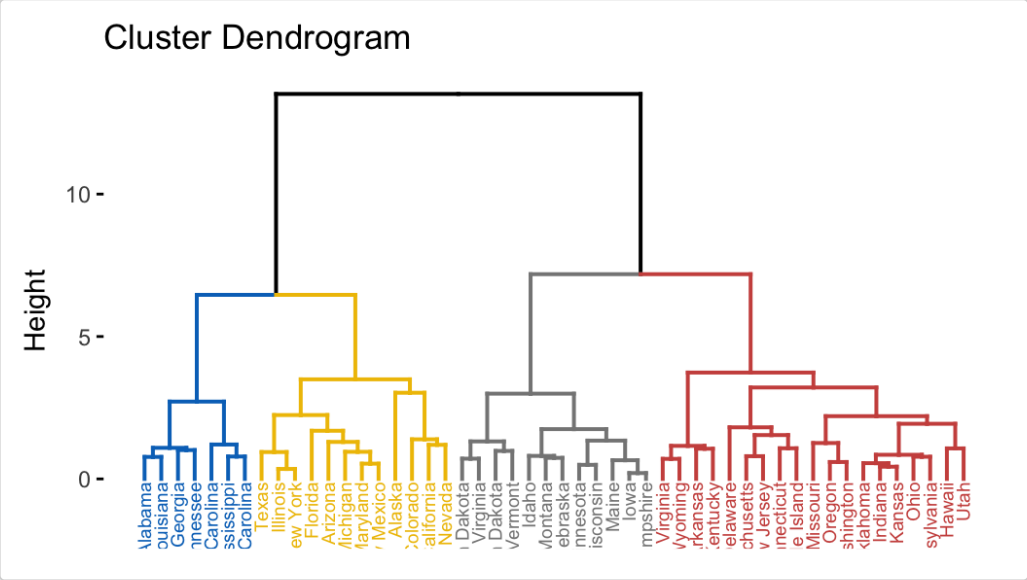
To change the plot theme, use the argument ggtheme, which allowed values include ggplot2 official themes [*theme_gray*(), *theme_bw*(), *theme_minimal*(), *theme_classic*(), *theme_void*()] or any other user-defined ggplot2 themes.

```
fviz_dend(hc, k = 4,                  # Cut in four groups
          cex = 0.5,                  # label size
          k_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
          color_labels_by_k = TRUE,   # color labels by groups
          ggtheme = theme_gray()      # Change theme
          )
```

Allowed values for k_color include brewer palettes from *RColorBrewer* Package (e.g. “RdBu”, “Blues”, “Dark2”, “Set2”, …;) and scientific journal palettes from *ggsci* R package (e.g.: “npg”, “aaas”, “lancet”, “jco”, “ucscgb”, “uchicago”, “simpsons” and “rickandmorty”).

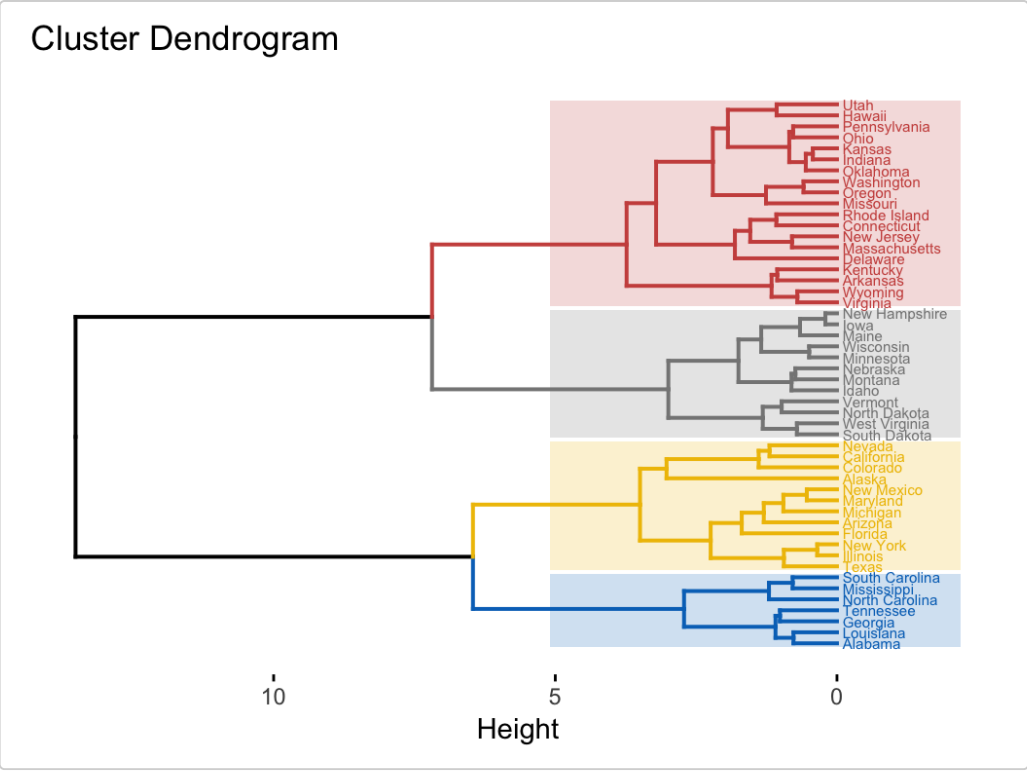
In the R code below, we’ll change group colors using “jco” (journal of clinical oncology) color palette:

```
fviz_dend(hc, cex = 0.5, k = 4, # Cut in four groups
          k_colors = "jco")
```



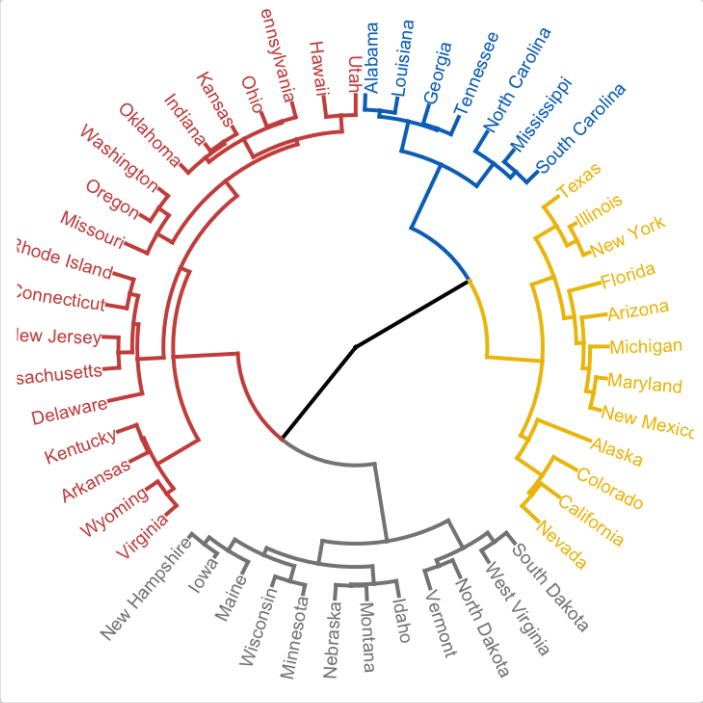
If you want to draw a horizontal dendrogram with rectangle around clusters, use this:

```
fviz_dend(hc, k = 4, cex = 0.4, horiz = TRUE,  k_colors = "jco",
          rect = TRUE, rect_border = "jco", rect_fill = TRUE)
```



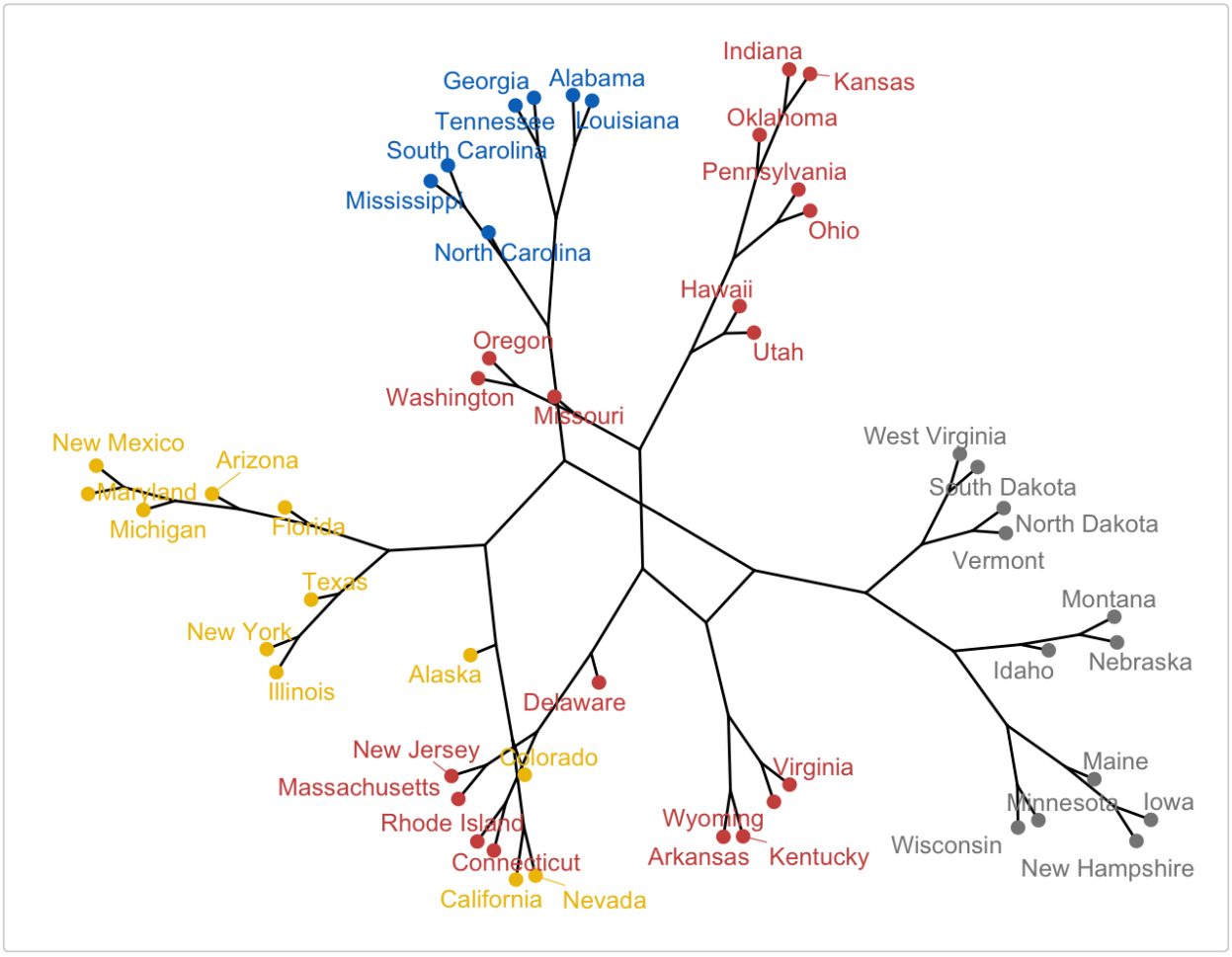
Additionally, you can plot a circular dendrogram using the option type = “circular”.

```
fviz_dend(hc, cex = 0.5, k = 4,
          k_colors = "jco", type = "circular")
```



To plot a phylogenic-like tree, use type = “phylogenic” and repel = TRUE (to avoid labels overplotting). This functionality requires the R package *igraph*. Make sure that it’s installed before typing the following R code.

```
require("igraph")
fviz_dend(hc, k = 4, k_colors = "jco",
          type = "phylogenic", repel = TRUE)
```



The default layout for phylogenic trees is “layout.auto”. Allowed values are one of: c(“layout.auto”, “layout_with_drl”, “layout_as_tree”, “layout.gem”, “layout.mds”, “layout_with_lgl”). To read more about these layouts, read the documentation of the igraph R package.

Let’s try phylo.layout = “layout.gem”:

```
require("igraph")
fviz_dend(hc, k = 4, # Cut in four groups
          k_colors = "jco",
          type = "phylogenic", repel = TRUE,
          phylo_layout = "layout.gem")
```

Case of dendrogram with large data sets

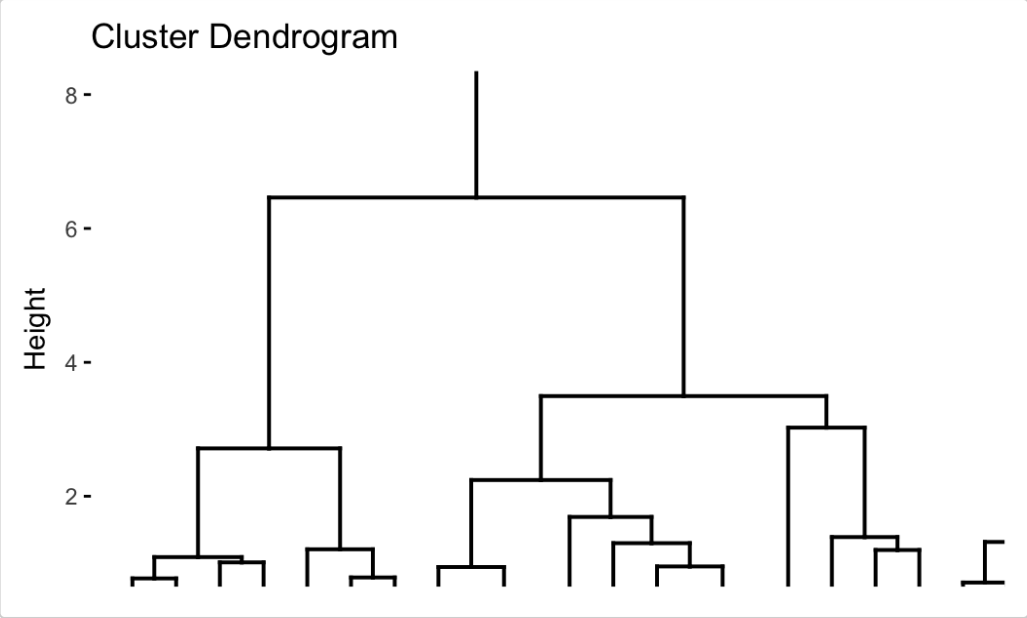
If you compute hierarchical clustering on a large data set, you might want to zoom in the dendrogram or to plot only a subset of the dendrogram.

Alternatively, you could also plot the dendrogram to a large page on a PDF, which can be zoomed without loss of resolution.

Zooming in the dendrogram

If you want to zoom in the first clusters, its possible to use the option xlim and ylim to limit the plot area. For example, type the code below:

```
fviz_dend(hc, xlim = c(1, 20), ylim = c(1, 8))
```



Plotting a sub-tree of dendrograms

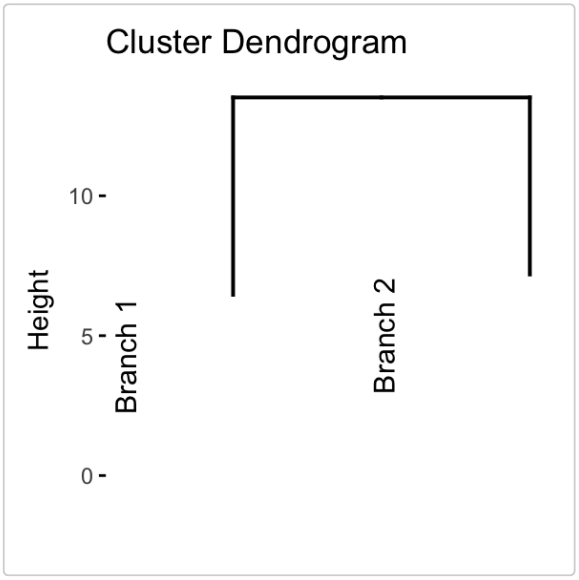
To plot a sub-tree, we'll follow the procedure below:

1. Create the whole dendrogram using *fviz_dend*() and save the result into an object, named dend_plot for example.
2. Use the R base function *cut.dendrogram*() to cut the dendrogram, at a given height (h), into multiple sub-trees. This returns a list with components $upper and $lower, the first is a truncated version of the original tree, also of class dendrogram, the latter a list with the branches obtained from cutting the tree, each a dendrogram.
3. Visualize sub-trees using *fviz_dend*().

The R code is as follow.

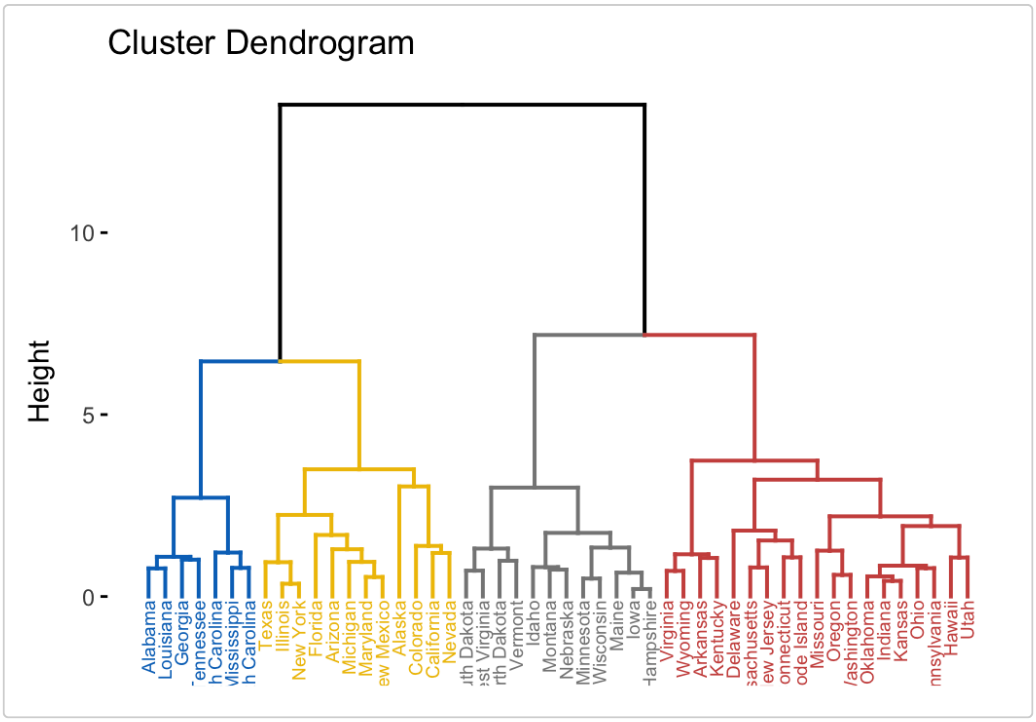
- Cut the dendrogram and visualize the truncated version:

```
# Create a plot of the whole dendrogram,
# and extract the dendrogram data
dend_plot <- fviz_dend(hc, k = 4, # Cut in four groups
          cex = 0.5, # label size
          k_colors = "jco"
          )
dend_data <- attr(dend_plot, "dendrogram") # Extract dendrogram data
# Cut the dendrogram at height h = 10
dend_cuts <- cut(dend_data, h = 10)
# Visualize the truncated version containing
# two branches
fviz_dend(dend_cuts$upper)
```

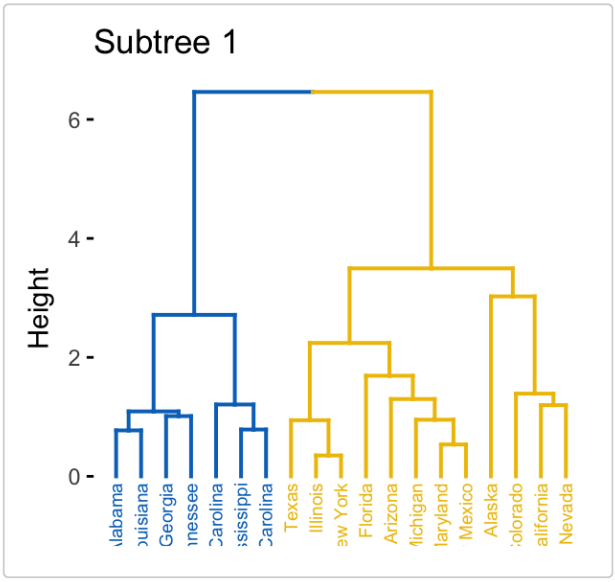


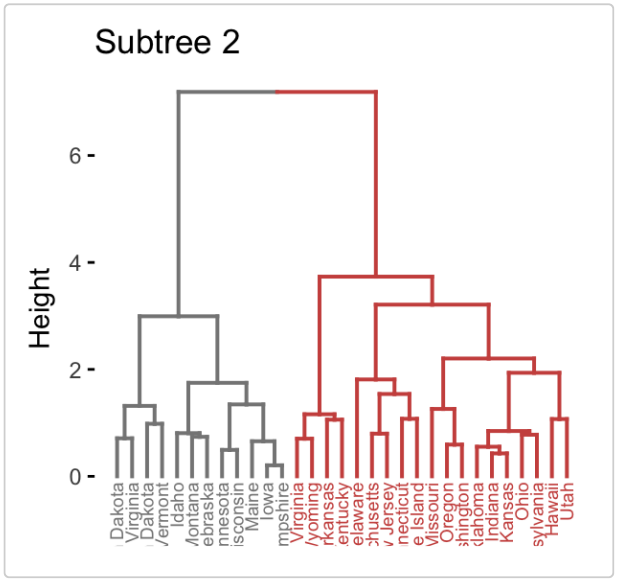
- Plot dendrograms sub-trees:

```
# Plot the whole dendrogram
print(dend_plot)
```



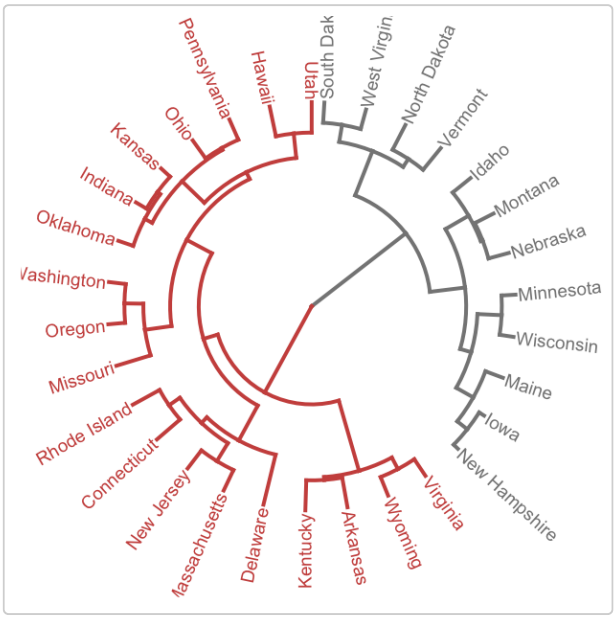

```
# Plot subtree 1
fviz_dend(dend_cuts$lower[[1]], main = "Subtree 1")
# Plot subtree 2
fviz_dend(dend_cuts$lower[[2]], main = "Subtree 2")
```





You can also plot circular trees as follow:

```
fviz_dend(dend_cuts$lower[[2]], type = "circular")
```



Saving dendrogram into a large PDF page

If you have a large dendrogram, you can save it to a large PDF page, which can be zoomed without loss of resolution.

```
pdf("dendrogram.pdf", width=30, height=15)            # Open a PDF
p <- fviz_dend(hc, k = 4, cex = 1, k_colors = "jco" ) # Do plotting
print(p)
dev.off()                                             # Close the PDF
```

Manipulating dendrograms using dendextend

The package *dendextend* provide functions for changing easily the appearance of a dendrogram and for comparing dendrograms.

In this section we'll use the chaining operator (*%>%*) to simplify our code. The chaining operator turns x %>% f(y) into f(x, y) so you can use it to rewrite multiple operations such that they can be read from left-to-right, top-to-bottom. For instance, the results of the two R codes below are equivalent.

- Standard R code for creating a dendrogram:

```
data <- scale(USArrests)
dist.res <- dist(data)
hc <- hclust(dist.res, method = "ward.D2")
dend <- as.dendrogram(hc)
plot(dend)
```

- R code for creating a dendrogram using chaining operator:

```
library(dendextend)
dend <- USArrests[1:5,] %>% # data
        scale %>% # Scale the data
        dist %>% # calculate a distance matrix,
        hclust(method = "ward.D2") %>% # Hierarchical clustering
        as.dendrogram # Turn the object into a dendrogram.
plot(dend)
```

- Functions to customize dendrograms: The function *set*() [in dendextend package] can be used to change the parameters of a dendrogram. The format is:

```
set(object, what, value)
```

1. **object**: a dendrogram object
2. **what**: a character indicating what is the property of the tree that should be set/updated
3. **value**: a vector with the value to set in the tree (the type of the value depends on the "what").

Possible values for the argument **what** include:

Value for the argument what	Description
labels	set the labels
labels_colors and **labels_cex**	Set the color and the size of labels, respectively
leaves_pch, **leaves_cex** and **leaves_col**	set the point type, size and color for leaves, respectively
nodes_pch, **nodes_cex** and **nodes_col**	set the point type, size and color for nodes, respectively
hang_leaves	hang the leaves
branches_k_color	color the branches
branches_col, **branches_lwd** , **branches_lty**	Set the color, the line width and the line type of branches, respectively
by_labels_branches_col, **by_labels_branches_lwd** and **by_labels_branches_lty**	Set the color, the line width and the line type of branches with specific labels, respectively
clear_branches and **clear_leaves**	Clear branches and leaves, respectively

- Examples:

```
library(dendextend)
# 1. Create a customized dendrogram
mycols <- c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07")
dend <-  as.dendrogram(hc) %>%
   set("branches_lwd", 1) %>% # Branches line width
   set("branches_k_color", mycols, k = 4) %>% # Color branches by groups
   set("labels_colors", mycols, k = 4) %>%  # Color labels by groups
   set("labels_cex", 0.5) # Change label size
# 2. Create plot
fviz_dend(dend)
```

Summary

We described functions and packages for visualizing and customizing dendrograms including:

- *fviz_dend*() [in factoextra R package], which provides convenient solutions for plotting easily a beautiful dendrogram. It can be used to create rectangular and circular dendrograms, as well as, a phylogenic tree.
- and the *dendextend* package, which provides a flexible methods to customize dendrograms.

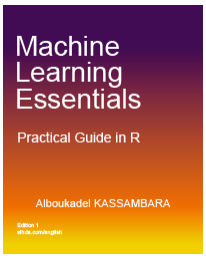
Additionally, we described how to plot a subset of large dendrograms.

Last update : 24/09/2017
□ □ □ □ □ 0 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

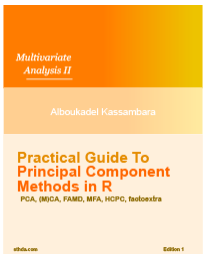
Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!

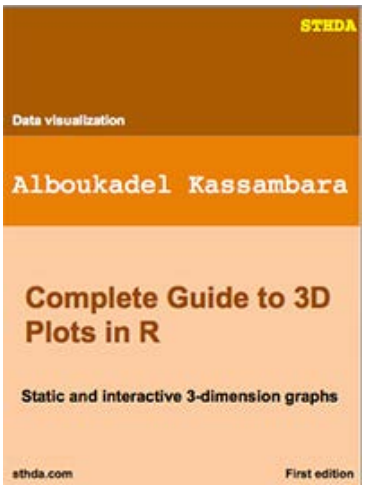


Machine Learning Essentials: Practical Guide in R

Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R



Articles - Hierarchical Clustering Essentials

Comparing Dendrograms: Essentials

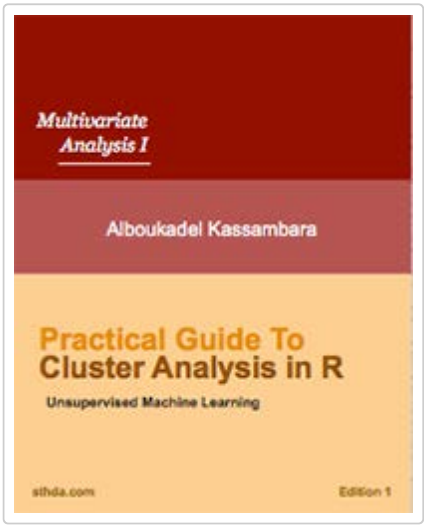
kassambara | 06/09/2017 | 2226 | Comments (2) | Hierarchical Clustering Essentials | Unsupervised machine learning, Multivariate Analysis, Genomic data visualization

After showing how to compute hierarchical clustering (Chapter @ref(agglomerative-clustering)), we describe, here, how to **compare two dendrograms** using the *dendextend* R package.

The *dendextend* package provides several functions for comparing dendrograms. Here, we’ll focus on two functions:

- *tanglegram*() for visual comparison of two dendrograms
- and *cor.dendlist*() for computing a correlation matrix between dendrograms.

Contents:



Practical Guide to Cluster Analysis in R

Data preparation

We'll use the R base USArrests data sets and we start by standardizing the variables using the function *scale*() as follow:

```
df <- scale(USArrests)
```

To make readable the plots, generated in the next sections, we'll work with a small random subset of the data set. Therefore, we'll use the function *sample*() to randomly select 10 observations among the 50 observations contained in the data set:

```
# Subset containing 10 rows
set.seed(123)
ss <- sample(1:50, 10)
df <- df[ss,]
```

Dendrograms comparison

We start by creating a list of two dendrograms by computing hierarchical clustering (HC) using two different linkage methods ("average" and "ward.D2"). Next, we transform the results as dendrograms and create a list to hold the two dendrograms.

```
library(dendextend)
# Compute distance matrix
res.dist <- dist(df, method = "euclidean")
# Compute 2 hierarchical clusterings
hc1 <- hclust(res.dist, method = "average")
hc2 <- hclust(res.dist, method = "ward.D2")
# Create two dendrograms
dend1 <- as.dendrogram (hc1)
dend2 <- as.dendrogram (hc2)
# Create a list to hold dendrograms
```

```
dend_list <- dendlist(dend1, dend2)
```

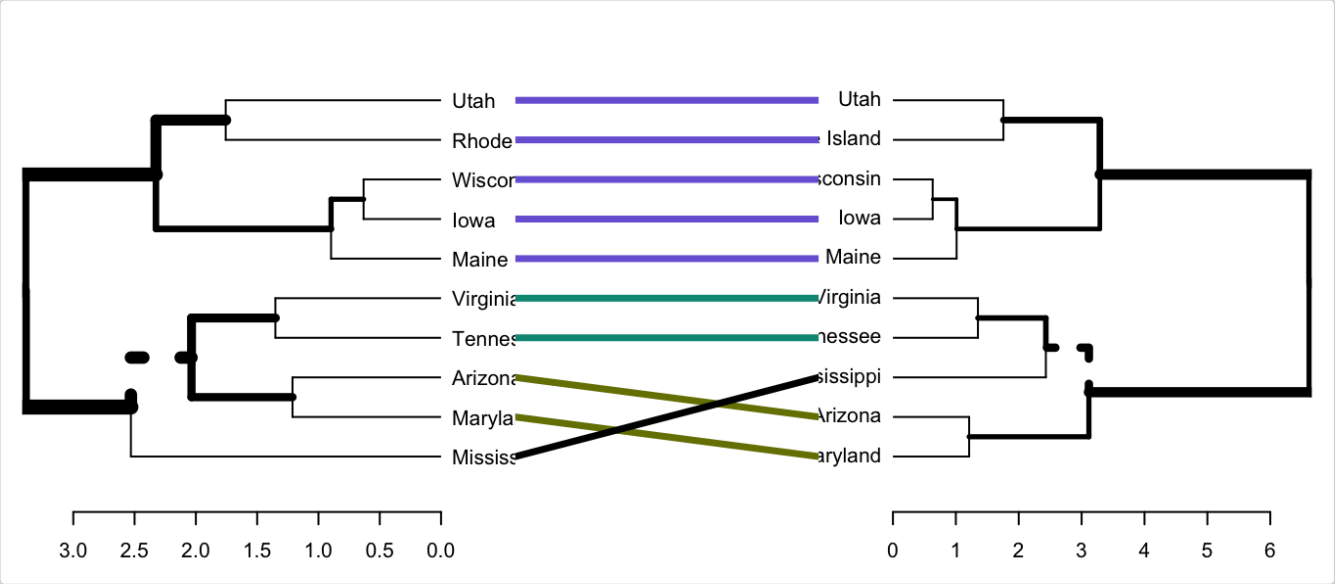
1. **Visual comparison of two dendrograms**

To visually compare two dendrograms, we'll use the following R functions [*dendextend* package]:

- *untangle*(): finds the best layout to align dendrogram lists, using heuristic methods
- *tanglegram*(): plots the two dendrograms, side by side, with their labels connected by lines.
- *entanglement*(): computes the quality of the alignment of the two trees. Entanglement is a measure between 1 (full entanglement) and 0 (no entanglement). A lower entanglement coefficient corresponds to a good alignment.

- Draw a tanglegram:

```
# Align and plot two dendrograms side by side
dendlist(dend1, dend2) %>%
  untangle(method = "step1side") %>% # Find the best alignment layout
  tanglegram()                       # Draw the two dendrograms
```




```
# Compute alignment quality. Lower value = good alignment quality
dendlist(dend1, dend2) %>%
  untangle(method = "step1side") %>% # Find the best alignment layout
  entanglement()                     # Alignment quality
```

```
## [1] 0.0384
```

- Customized the tanglegram using many other options as follow:

```
dendlist(dend1, dend2) %>%
  untangle(method = "step1side") %>%
  tanglegram(
    highlight_distinct_edges = FALSE, # Turn-off dashed lines
    common_subtrees_color_lines = FALSE, # Turn-off line colors
    common_subtrees_color_branches = TRUE # Color common branches
    )
```

Note that, “unique” nodes, with a combination of labels/items not present in the other tree, are highlighted with dashed lines.

Note that, just because we can get two trees to have horizontal connecting lines, it doesn’t mean these trees are identical (or even very similar topologically).

In the following section, we’ll perform correlation analysis to measure the similarity between dendrograms.

2. **Correlation matrix between a list of dendrogams**

The function *cor.dendlist*() is used to compute “*Baker*” or “*Cophenetic*” correlation matrix between a list of trees. The value can range between -1 to 1. With near 0 values meaning that the two trees are not statistically similar.

```
# Cophenetic correlation matrix
cor.dendlist(dend_list, method = "cophenetic")
```

```
##       [,1]  [,2]
## [1,] 1.000 0.965
## [2,] 0.965 1.000
```

```
# Baker correlation matrix
cor.dendlist(dend_list, method = "baker")
```

```
##       [,1]  [,2]
## [1,] 1.000 0.962
## [2,] 0.962 1.000
```

The correlation between two trees can be also computed as follow:

```
# Cophenetic correlation coefficient
cor_cophenetic(dend1, dend2)
```

```
## [1] 0.965
```

```
# Baker correlation coefficient
cor_bakers_gamma(dend1, dend2)
```

```
## [1] 0.962
```

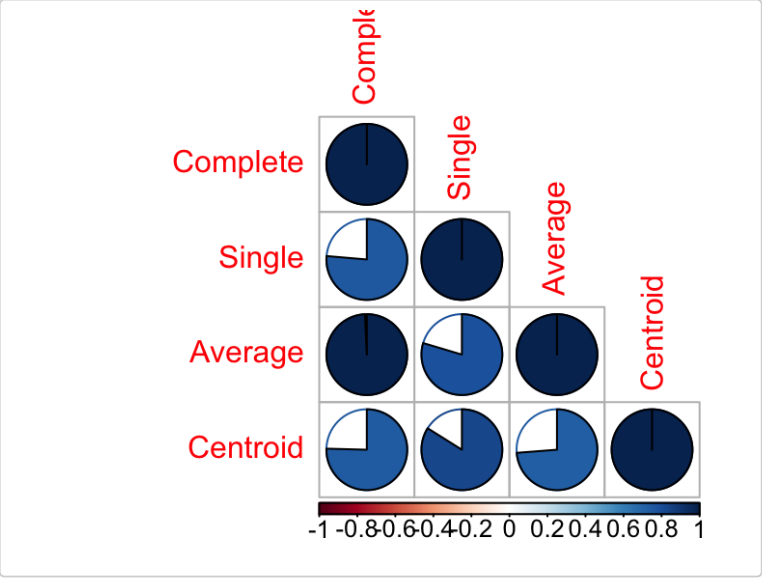
It’s also possible to compare simultaneously multiple dendrograms. A chaining operator *%>%* is used to run multiple function at the same time. It’s useful for simplifying the code:

```
# Create multiple dendrograms by chaining
dend1 <- df %>% dist %>% hclust("complete") %>% as.dendrogram
dend2 <- df %>% dist %>% hclust("single") %>% as.dendrogram
dend3 <- df %>% dist %>% hclust("average") %>% as.dendrogram
dend4 <- df %>% dist %>% hclust("centroid") %>% as.dendrogram
```

```
# Compute correlation matrix
dend_list <- dendlist("Complete" = dend1, "Single" = dend2,
                      "Average" = dend3, "Centroid" = dend4)
cors <- cor.dendlist(dend_list)
# Print correlation matrix
round(cors, 2)
```

```
##          Complete Single Average Centroid
## Complete     1.00   0.76    0.99     0.75
## Single       0.76   1.00    0.80     0.84
## Average      0.99   0.80    1.00     0.74
## Centroid     0.75   0.84    0.74     1.00
```

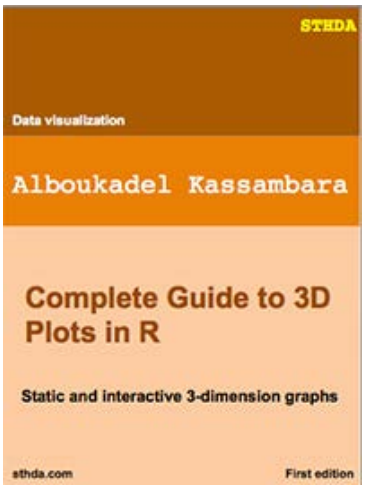
```
# Visualize the correlation matrix using corrplot package
library(corrplot)
corrplot(cors, "pie", "lower")
```



Last update : 19/05/2018
0 Note

Enjoyed this article? Give us 5 stars (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!



Articles - Hierarchical Clustering Essentials

Agglomerative Clustering Essentials

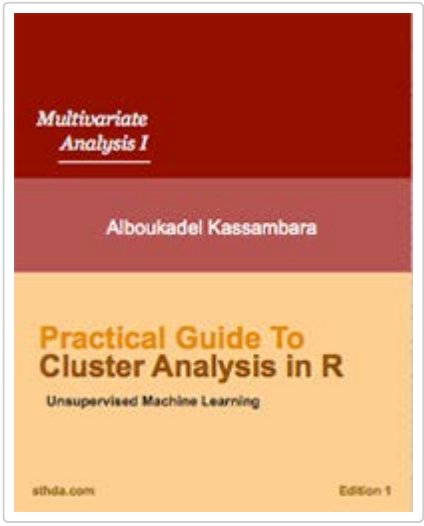
kassambara | 06/09/2017 | 3854 | Comments (6) | Hierarchical Clustering Essentials | Unsupervised machine learning, Multivariate Analysis, Genomic data visualization

The **agglomerative clustering** is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It's also known as *AGNES* (*Agglomerative Nesting*). The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named *dendrogram*.

In this article we start by describing the agglomerative clustering algorithms. Next, we provide R lab sections with many examples for computing and visualizing hierarchical clustering. We continue by explaining how to interpret dendrogram. Finally, we provide R codes for cutting dendrograms into groups.

Contents:

Related Books:



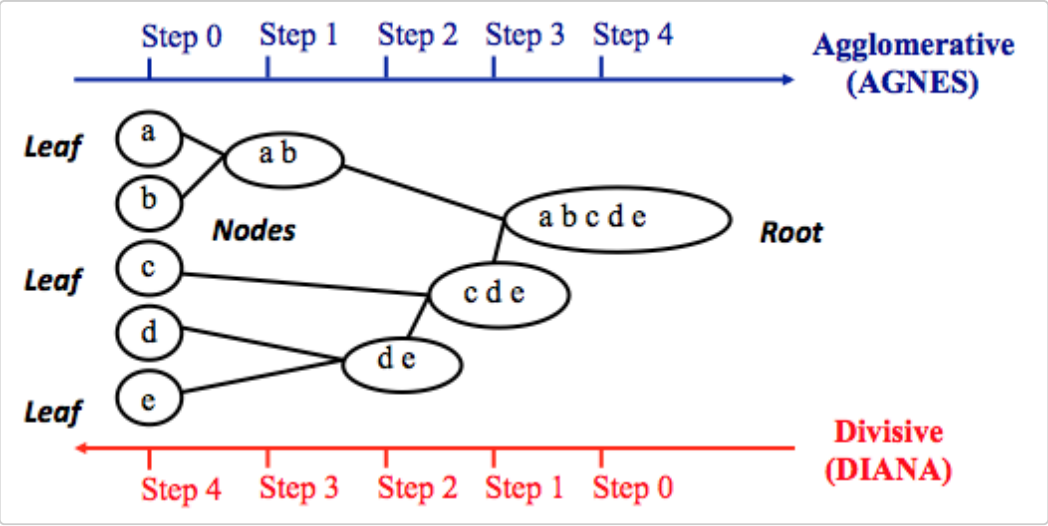
Practical Guide to Cluster Analysis in R

Algorithm

Agglomerative clustering works in a “bottom-up” manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root) (see figure below).

The inverse of agglomerative clustering is *divisive clustering*, which is also known as DIANA (*Divise Analysis*) and it works in a “top-down” manner. It begins with the root, in which all objects are included in a single cluster. At each step of iteration, the most heterogeneous cluster is divided into two. The process is iterated until all objects are in their own cluster (see figure below).

Note that, agglomerative clustering is good at identifying small clusters. Divisive clustering is good at identifying large clusters. In this article, we’ll focus mainly on agglomerative hierarchical clustering.



Steps to agglomerative hierarchical clustering

We’ll follow the steps below to perform agglomerative hierarchical clustering using R software:

1. Preparing the data

2. Computing (dis)similarity information between every pair of objects in the data set.
3. Using linkage function to group objects into hierarchical cluster tree, based on the distance information generated at step 1. Objects/clusters that are in close proximity are linked together using the linkage function.
4. Determining where to cut the hierarchical tree into clusters. This creates a partition of the data.

We'll describe each of these steps in the next section.

Data structure and preparation

The data should be a numeric matrix with:

- rows representing observations (individuals);
- and columns representing variables.

Here, we'll use the R base USArrests data sets.

Note that, it's generally recommended to standardize variables in the data set before performing subsequent analysis. Standardization makes variables comparable, when they are measured in different scales. For example one variable can measure the height in meter and another variable can measure the weight in kg. The R function *scale*() can be used for standardization, See ?scale documentation.

```
# Load the data
data("USArrests")
# Standardize the data
df <- scale(USArrests)
# Show the first 6 rows
head(df, nrow = 6)
```

```
##            Murder Assault UrbanPop     Rape
## Alabama    1.2426   0.783   -0.521 -0.00342
## Alaska     0.5079   1.107   -1.212  2.48420
## Arizona    0.0716   1.479    0.999  1.04288
## Arkansas   0.2323   0.231   -1.074 -0.18492
## California 0.2783   1.263    1.759  2.06782
## Colorado   0.0257   0.399    0.861  1.86497
```

Similarity measures

In order to decide which objects/clusters should be combined or divided, we need methods for measuring the similarity between objects.

There are many methods to calculate the (dis)similarity information, including Euclidean and manhattan distances (Chapter @ref(clustering-distance-measures)). In R software, you can use the function *dist*() to compute the distance between every pair of object in a data set. The results of this computation is known as a distance or dissimilarity matrix.

By default, the function *dist*() computes the Euclidean distance between objects; however, it's possible to indicate other metrics using the argument method. See ?dist for more information.

For example, consider the R base data set USArrests, you can compute the distance matrix as follow:

```
# Compute the dissimilarity matrix
# df = the standardized data
res.dist <- dist(df, method = "euclidean")
```

Note that, the function () computes the distance between the rows of a data matrix using the specified distance measure method.

To see easily the distance information between objects, we reformat the results of the function *dist*() into a matrix using the *as.matrix*() function. In this matrix, value in the cell formed by the row i, the column j, represents the distance between object i and object j in the original data set. For instance, element 1,1 represents the distance between object 1 and itself (which is zero). Element 1,2 represents the distance between object 1 and object 2, and so on.

The R code below displays the first 6 rows and columns of the distance matrix:

```
as.matrix(res.dist)[1:6, 1:6]
```

```
##            Alabama Alaska Arizona Arkansas California Colorado
## Alabama       0.00   2.70    2.29     1.29       3.26     2.65
## Alaska        2.70   0.00    2.70     2.83       3.01     2.33
## Arizona       2.29   2.70    0.00     2.72       1.31     1.37
## Arkansas      1.29   2.83    2.72     0.00       3.76     2.83
## California    3.26   3.01    1.31     3.76       0.00     1.29
## Colorado      2.65   2.33    1.37     2.83       1.29     0.00
```

Linkage

The linkage function takes the distance information, returned by the function *dist*(), and groups pairs of objects into clusters based on their similarity. Next, these newly formed clusters are linked to each other to create bigger clusters. This process is iterated until all the objects in the original data set are linked together in a hierarchical tree.

For example, given a distance matrix "res.dist" generated by the function *dist*(), the R base function *hclust*() can be used to create the hierarchical tree.

hclust() can be used as follow:

```
res.hc <- hclust(d = res.dist, method = "ward.D2")
```

- **d**: a dissimilarity structure as produced by the **dist()** function.
- **method**: The agglomeration (linkage) method to be used for computing distance between clusters. Allowed values is one of "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median" or "centroid".

There are many cluster agglomeration methods (i.e, linkage methods). The most common linkage methods are described below.

- Maximum or *complete linkage*: The distance between two clusters is defined as the maximum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce more compact clusters.
- Minimum or *single linkage*: The distance between two clusters is defined as the minimum value of all pairwise distances between the elements in cluster 1 and the elements in cluster 2. It tends to produce long, "loose" clusters.
- Mean or *average linkage*: The distance between two clusters is defined as the average distance between the elements in cluster 1 and the elements in cluster 2.
- *Centroid linkage*: The distance between two clusters is defined as the distance between the centroid for cluster 1 (a mean vector of length p variables) and the centroid for cluster 2.
- *Ward's minimum variance method*: It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

Note that, at each stage of the clustering process the two clusters, that have the smallest linkage distance, are linked together.

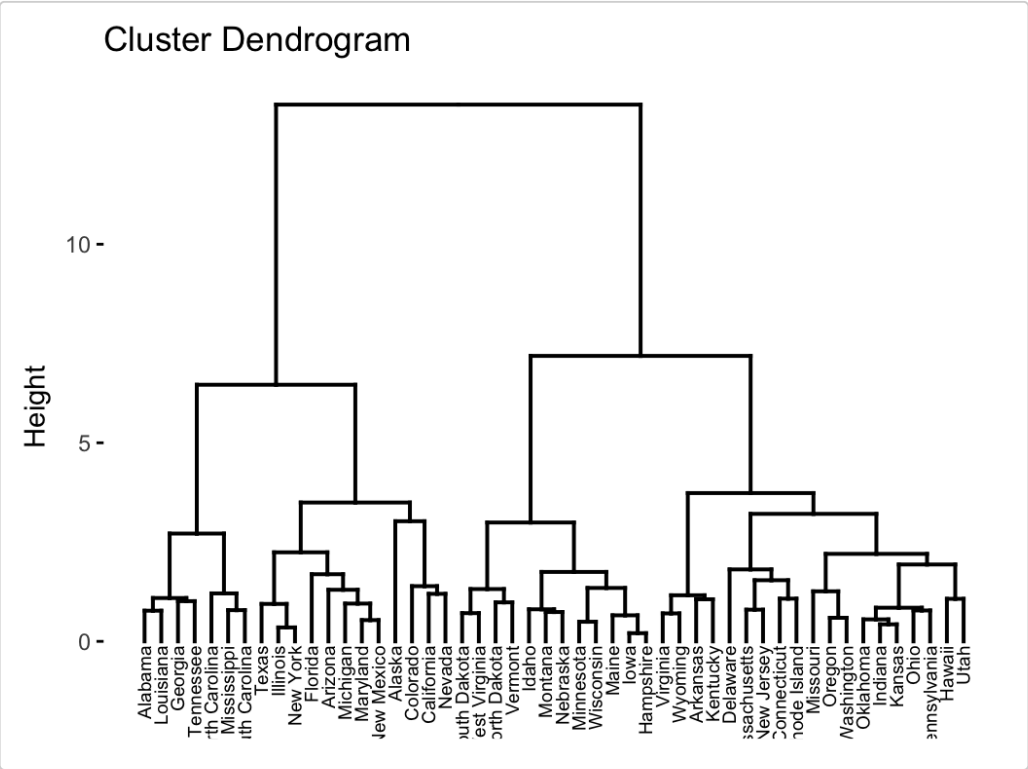
Complete linkage and Ward's method are generally preferred.

Dendrogram

Dendrograms correspond to the graphical representation of the hierarchical tree generated by the function *hclust*(). Dendrogram can be produced in R using the base function *plot*(res.hc), where res.hc is the output of *hclust*(). Here, we'll use the function *fviz_dend*()[in *factoextra* R package] to produce a beautiful dendrogram.

First install factoextra by typing this: install.packages("factoextra"); next visualize the dendrogram as follow:

```
# cex: label size
library("factoextra")
fviz_dend(res.hc, cex = 0.5)
```



In the dendrogram displayed above, each leaf corresponds to one object. As we move up the tree, objects that are similar to each other are combined into branches, which are themselves fused at a higher height.

The height of the fusion, provided on the vertical axis, indicates the (dis)similarity/distance between two objects/clusters. The higher the height of the fusion, the less similar the objects are. This height is known as the *cophenetic distance* between the two objects.

Note that, conclusions about the proximity of two objects can be drawn only based on the height where branches containing those two objects first are fused. We cannot use the proximity of two objects along the horizontal axis as a criteria of their similarity.

In order to identify sub-groups, we can cut the dendrogram at a certain height as described in the next sections.

Verify the cluster tree

After linking the objects in a data set into a hierarchical cluster tree, you might want to assess that the distances (i.e., heights) in the tree reflect the original distances accurately.

One way to measure how well the cluster tree generated by the *hclust*() function reflects your data is to compute the correlation between the *cophenetic* distances and the original distance data generated by the *dist*() function. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the original distance matrix.

The closer the value of the correlation coefficient is to 1, the more accurately the clustering solution reflects your data.

Values above 0.75 are felt to be good. The “average” linkage method appears to produce high values of this statistic. This may be one reason that it is so popular.

The R base function *cophenetic*() can be used to compute the cophenetic distances for hierarchical clustering.

```
# Compute cophentic distance
res.coph <- cophenetic(res.hc)
# Correlation between cophenetic distance and
# the original distance
cor(res.dist, res.coph)
```

```
## [1] 0.698
```

Execute the *hclust*() function again using the average linkage method. Next, call *cophenetic*() to evaluate the clustering solution.

```
res.hc2 <- hclust(res.dist, method = "average")
cor(res.dist, cophenetic(res.hc2))
```

```
## [1] 0.718
```

The correlation coefficient shows that using a different linkage method creates a tree that represents the original distances slightly better.

Cut the dendrogram into different groups

One of the problems with hierarchical clustering is that, it does not tell us how many clusters there are, or where to cut the dendrogram to form clusters.

You can cut the hierarchical tree at a given height in order to partition your data into clusters. The R base function *cutree*() can be used to cut a tree, generated by the *hclust*() function, into several groups either by specifying the desired number of groups or the cut height. It returns a vector containing the cluster number of each observation.

```
# Cut tree into 4 groups
grp <- cutree(res.hc, k = 4)
head(grp, n = 4)
```

```
##  Alabama   Alaska  Arizona Arkansas
##        1        2        2        3
```

```
# Number of members in each cluster
table(grp)
```

```
## grp
##  1  2  3  4
##  7 12 19 12
```

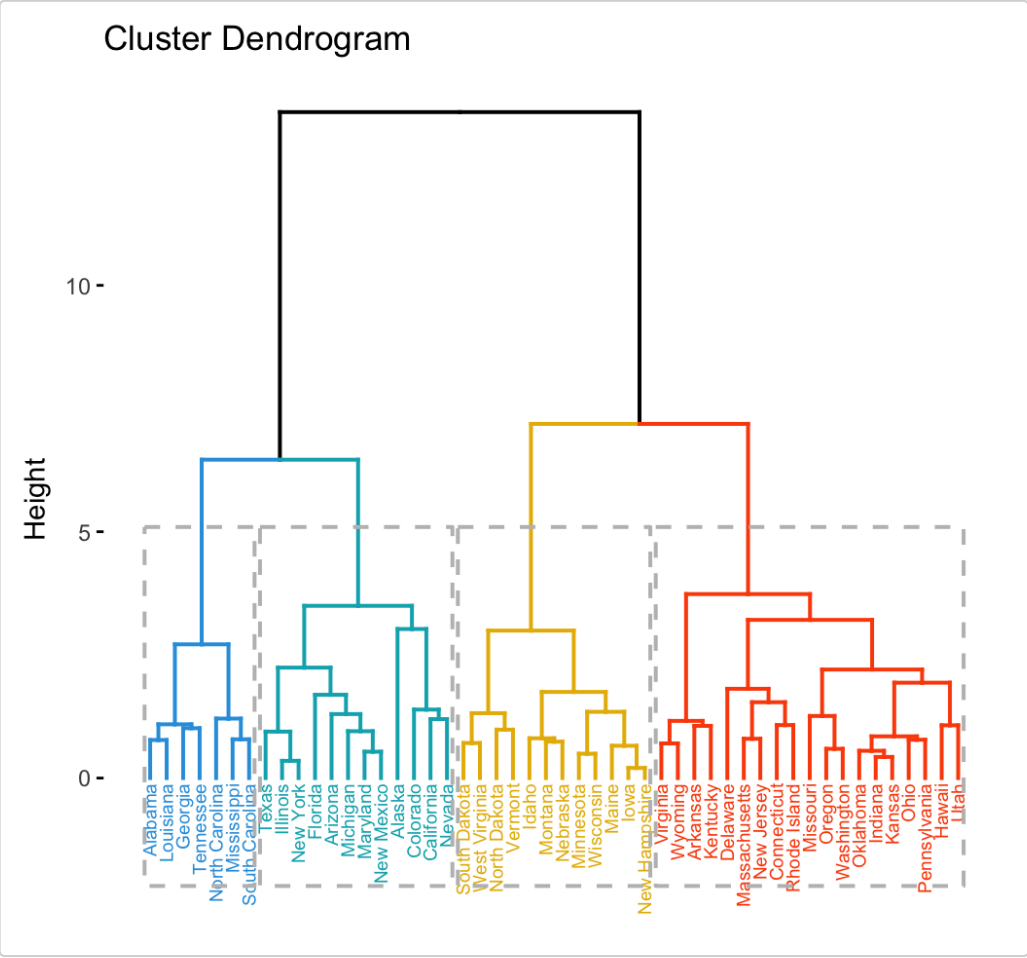
```
# Get the names for the members of cluster 1
```

```
rownames(df)[grp == 1]
```

```
## [1] "Alabama"        "Georgia"        "Louisiana"      "Mississippi"
## [5] "North Carolina" "South Carolina" "Tennessee"
```

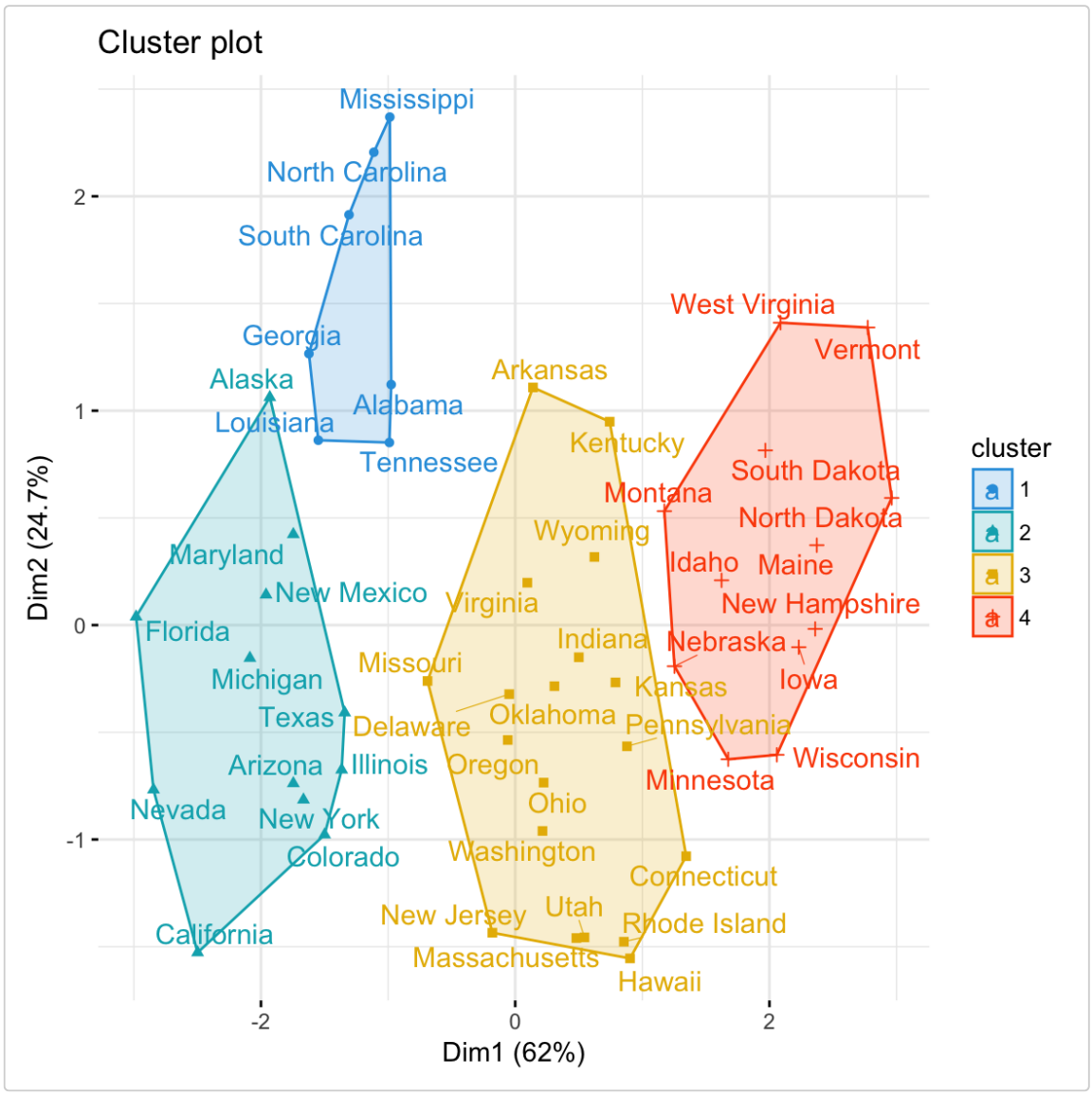
The result of the cuts can be visualized easily using the function *fviz_dend*() [in factoextra]:

```
# Cut in 4 groups and color by groups
fviz_dend(res.hc, k = 4, # Cut in four groups
          cex = 0.5, # label size
          k_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
          color_labels_by_k = TRUE, # color labels by groups
          rect = TRUE # Add rectangle around groups
          )
```



Using the function *fviz_cluster*() [in *factoextra*], we can also visualize the result in a scatter plot. Observations are represented by points in the plot, using principal components. A frame is drawn around each cluster.

```
fviz_cluster(list(data = df, cluster = grp),
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
             ellipse.type = "convex", # Concentration ellipse
             repel = TRUE, # Avoid label overplotting (slow)
             show.clust.cent = FALSE, ggtheme = theme_minimal())
```



Cluster R package

The R package *cluster* makes it easy to perform cluster analysis in R. It provides the function *agnes*() and *diana*() for computing agglomerative and divisive clustering, respectively. These functions perform all the necessary steps for you. You don't need to execute the *scale*(), *dist*() and *hclust*() function separately.

The functions can be executed as follow:

```
library("cluster")
# Agglomerative Nesting (Hierarchical Clustering)
res.agnes <- agnes(x = USArrests, # data matrix
                   stand = TRUE, # Standardize the data
                   metric = "euclidean", # metric for distance matrix
                   method = "ward" # Linkage method
                   )
# DIvisive ANAlysis Clustering
res.diana <- diana(x = USArrests, # data matrix
                   stand = TRUE, # standardize the data
                   metric = "euclidean" # metric for distance matrix
                   )
```

After running *agnes*() and *diana*(), you can use the function *fviz_dend*()[in *factoextra*] to visualize the output:

```
fviz_dend(res.agnes, cex = 0.6, k = 4)
```

Application of hierarchical clustering to gene expression data analysis

In *gene expression data analysis*, *clustering* is generaly used as one of the first step to explore the data. We are interested in whether there are groups of genes or groups of samples that have similar gene expression patterns.

Several distance measures (Chapter @ref(clustering-distance-measures)) have been described for assessing the similarity or the dissimilarity between items, in order to decide which items have to be grouped together or not. These measures can be used to cluster genes or samples that are similar.

For most common clustering softwares, the default distance measure is the Euclidean distance. The most popular methods for gene expression data are to use log2(expression + 0.25), correlation distance and complete linkage clustering agglomerative-clustering.

Single and Complete linkage give the same dendrogram whether you use the raw data, the log of the data or any other transformation of the data that preserves the order because what matters is which ones have the smallest distance. The other methods are sensitive to the measurement scale.

Note that, when the data are scaled, the Euclidean distance of the z-scores is the same as correlation distance.

Pearson's correlation is quite sensitive to outliers. When clustering genes, it is important to be aware of the possible impact of outliers. An alternative option is to use Spearman's correlation instead of Pearson's correlation.

In principle it is possible to cluster all the genes, although visualizing a huge dendrogram might be problematic. Usually, some type of preliminary analysis, such as differential expression analysis is used to select genes for clustering.

Selecting genes based on differential expression analysis removes genes which are likely to have only chance patterns. This should enhance the patterns found in the gene clusters.

Summary

Hierarchical clustering is a cluster analysis method, which produce a tree-based representation (i.e.: dendrogram) of a data. Objects in the dendrogram are linked together based on their similarity.

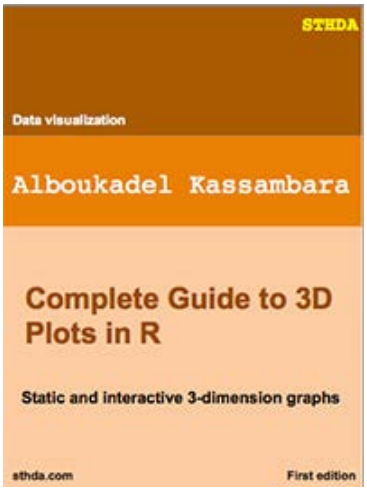
To perform hierarchical cluster analysis in R, the first step is to calculate the pairwise distance matrix using the function *dist*(). Next, the result of this computation is used by the *hclust*() function to produce the hierarchical tree. Finally, you can use the function *fviz_dend*() [in factoextra R package] to plot easily a beautiful dendrogram.

It's also possible to cut the tree at a given height for partitioning the data into multiple groups (R function *cutree*()).

Last update : 24/09/2017
□ □ □ □ □ 0 Note

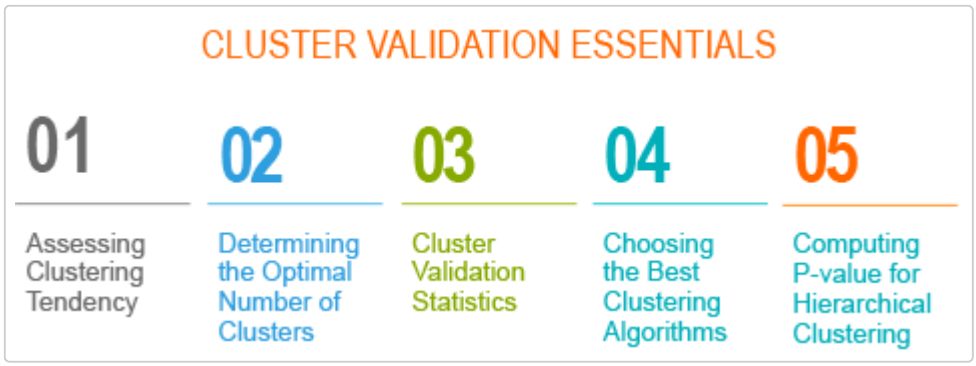
Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!



Articles - Cluster Validation Essentials

The **cluster validation** consists of measuring the goodness of clustering results. Before applying any clustering algorithm to a data set, the first thing to do is to assess the *clustering tendency*. That is, whether applying clustering is suitable for the data. If yes, then how many clusters are there. Next, you can perform hierarchical clustering or partitioning clustering (with a pre-specified number of clusters). Finally, you can use a number of measures, described in this part, to evaluate the goodness of the clustering results.



Contents

- Average silhouette method
- Gap statistic method
- Computing the number of clusters using R
  - Required R packages
  - Data preparation
  - fviz_nbclust() function: Elbow, Silhouhette and Gap statistic methods
  - NbClust() function: 30 indices for choosing the best number of clusters

Cluster Validation Statistics

- Internal measures for cluster validation
  - Silhouette coefficient
  - Dunn index
- External measures for clustering validation
- Computing cluster validation statistics in R
  - Required R packages
  - Data preparation
  - Clustering analysis
  - Cluster validation
  - External clustering validation

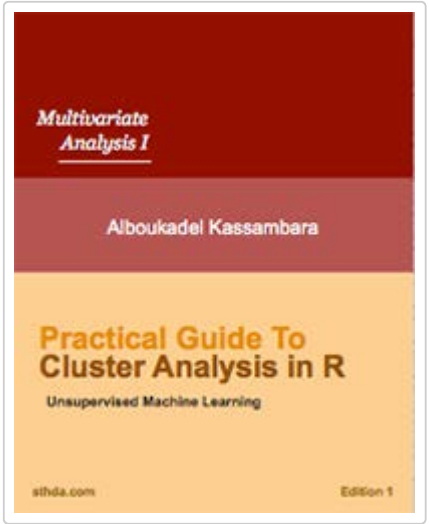
Choosing the Best Clustering Algorithms

- Measures for comparing clustering algorithms
- Compare clustering algorithms in R

Computing P-value for Hierarchical Clustering

- Description of pvclust() function
- Usage of pvclust() function

Related Book:



Practical Guide to Cluster Analysis in R

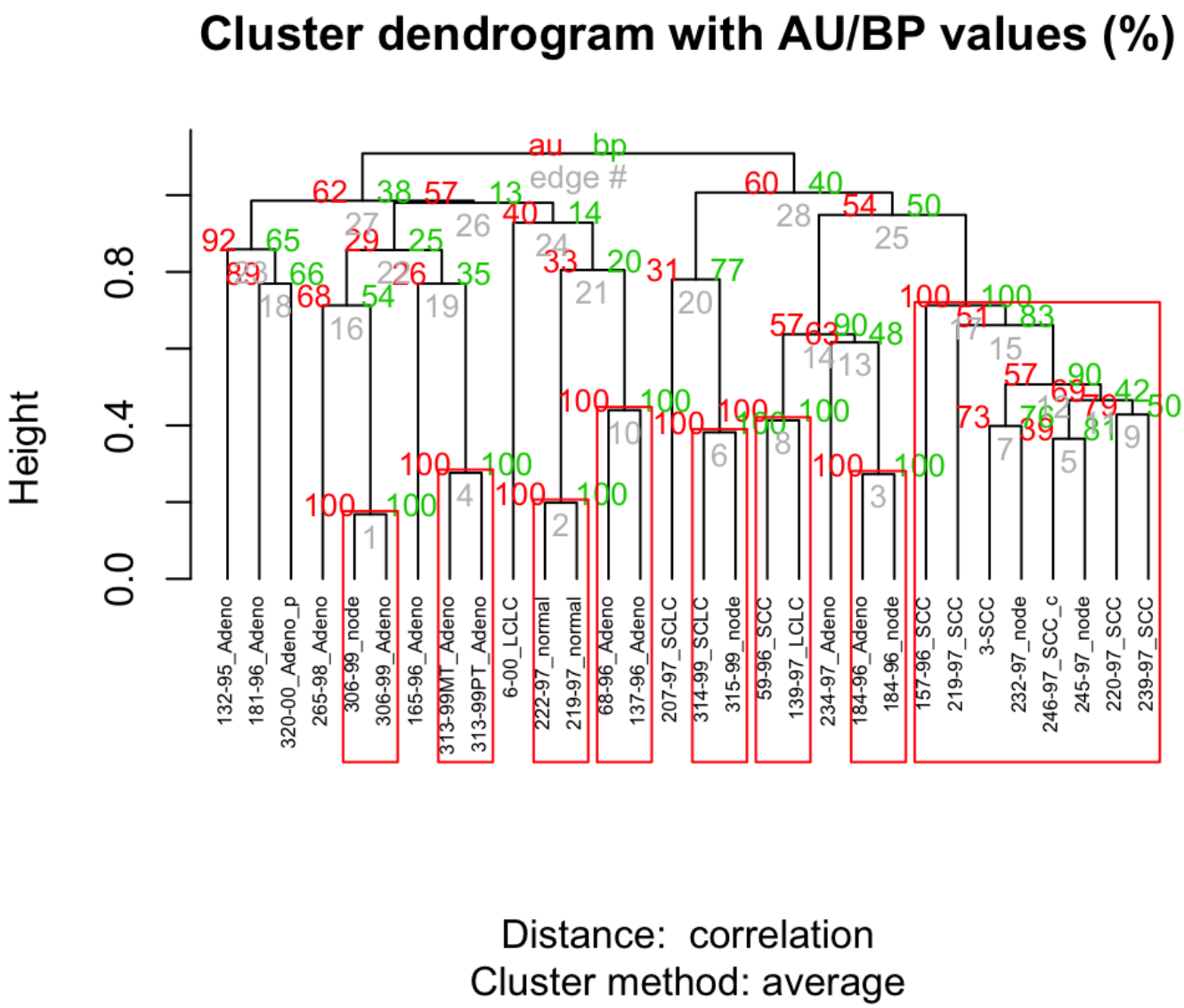
Sort by

Creation date

Descending

Computing P-value for Hierarchical Clustering

By kassambara, The 07/09/2017 in Cluster Validation Essentials



Clusters can be found in a data set by chance due to clustering noise or sampling error. This article describes the R package pvclust (Suzuki and Shimodaira 2015) which uses bootstrap resampling... [Read more]

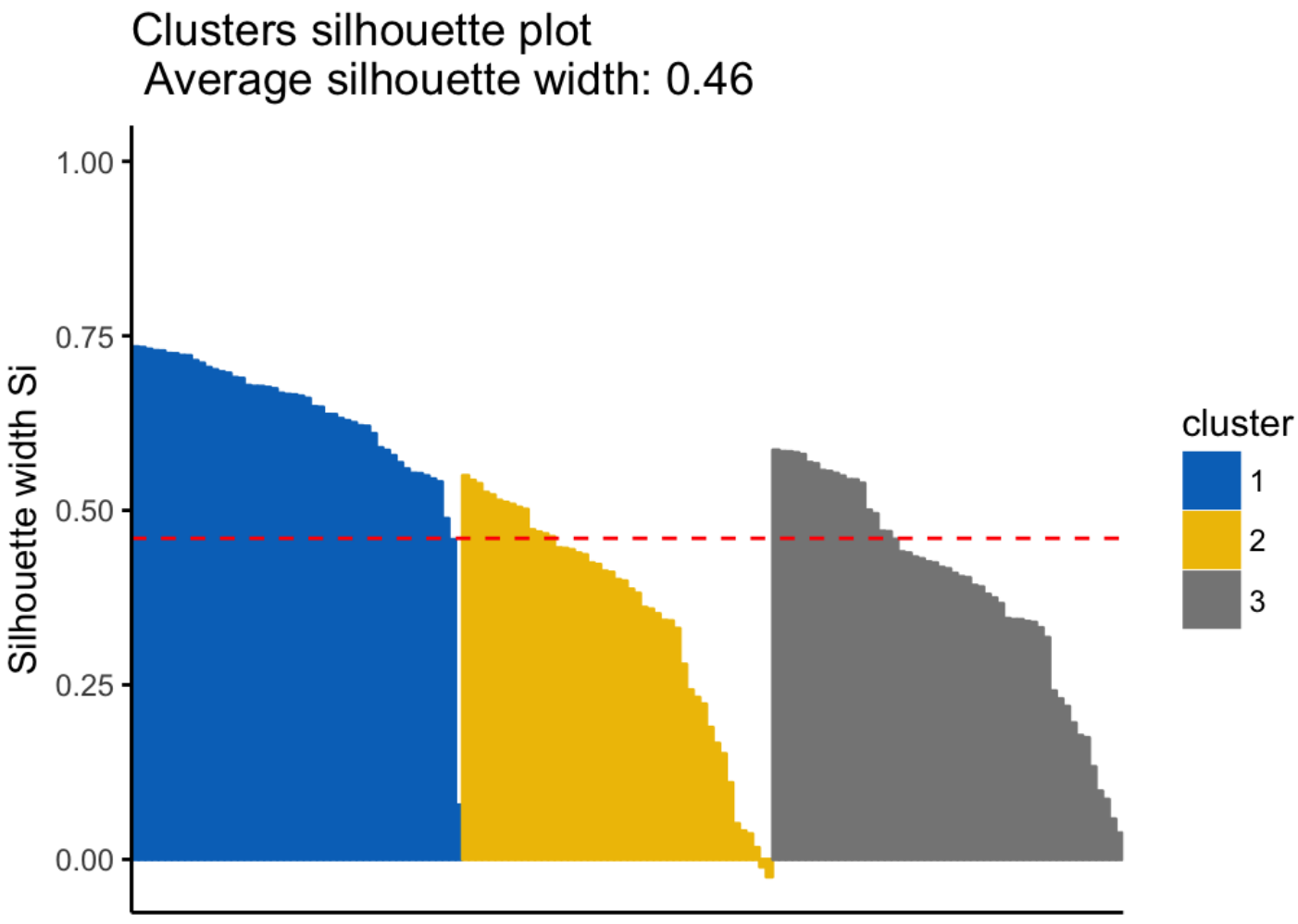
Choosing the Best Clustering Algorithms

By kassambara, The 07/09/2017 in Cluster Validation Essentials

Choosing the best clustering method for a given data can be a hard task for the analyst. This article describes the R package clValid (Brock et al. 2008), which can be used to compare... [Read more]

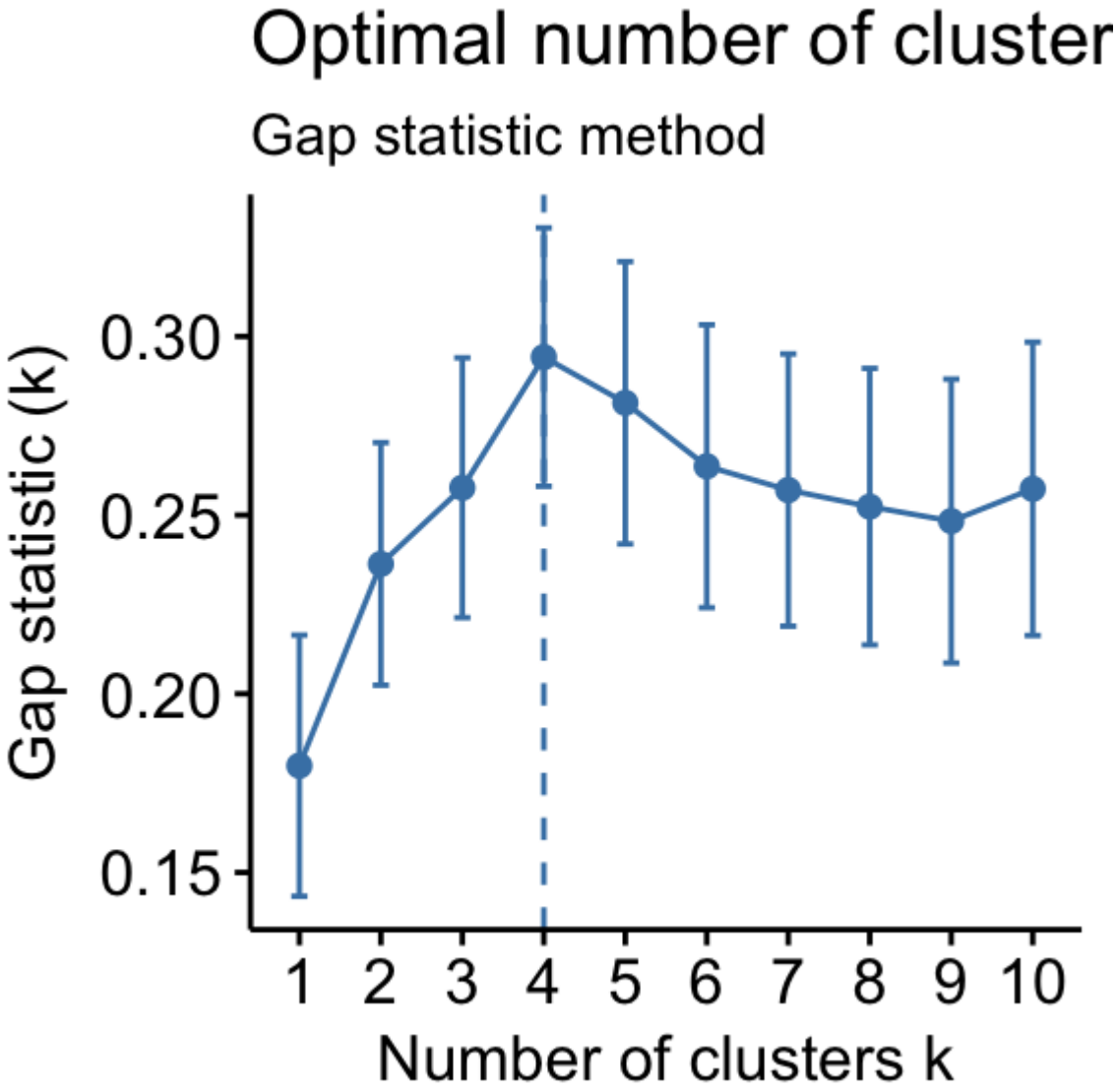
Cluster Validation Statistics: Must Know Methods

By kassambara, The 07/09/2017 in Cluster Validation Essentials



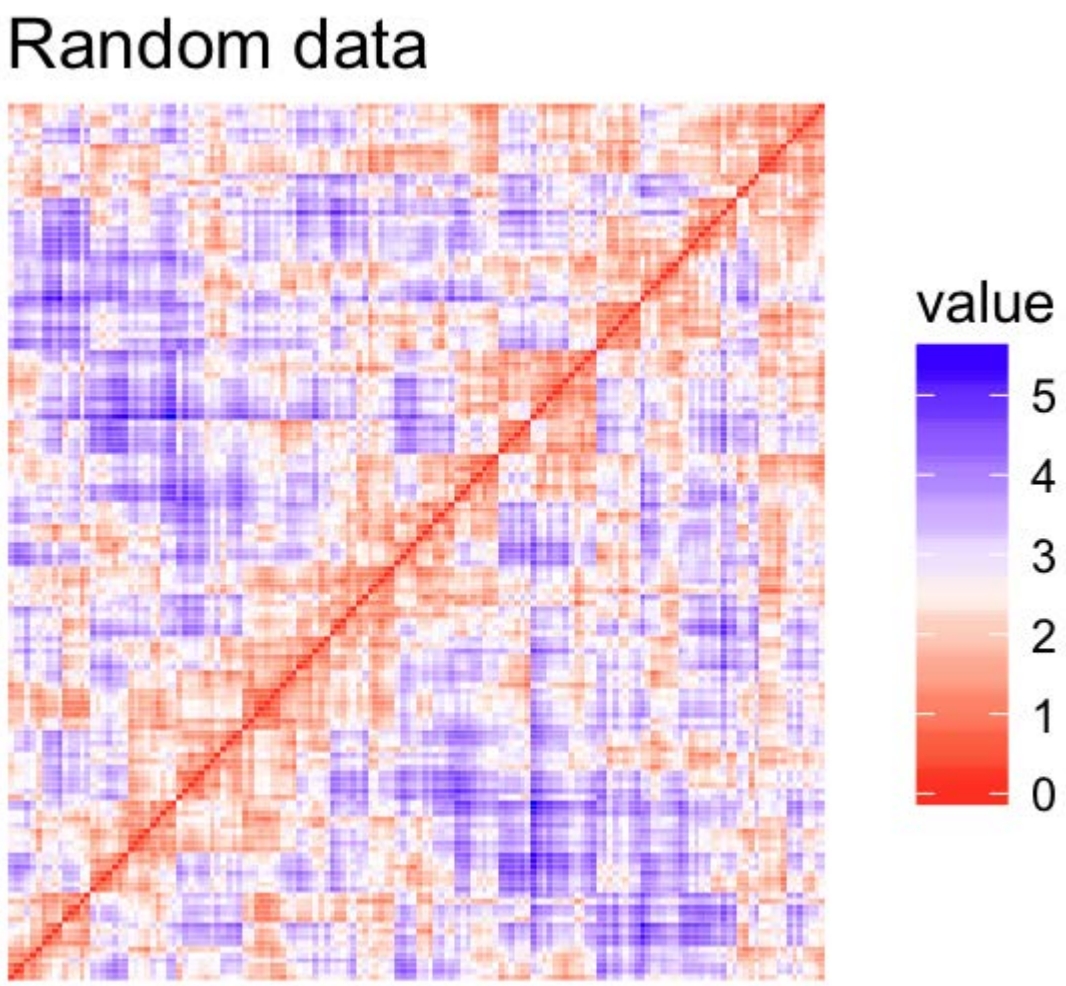
The term cluster validation is used to design the procedure of evaluating the goodness of clustering algorithm results. This is important to avoid finding patterns in a random data, as well as,... [Read more]

Determining The Optimal Number Of Clusters: 3 Must Know Methods
By kassambara, The 07/09/2017 in Cluster Validation Essentials

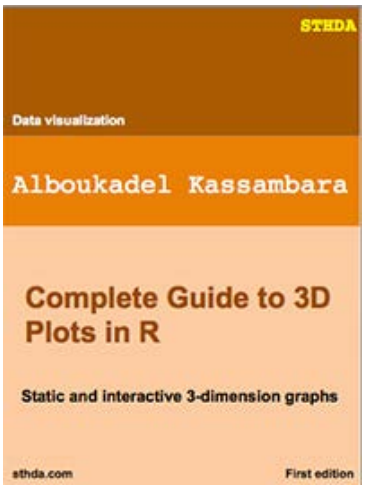


Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering (Chapter @ref(kmeans-clustering)), which requires the user... [Read more]

Assessing Clustering Tendency: Essentials
By kassambara, The 07/09/2017 in Cluster Validation Essentials



Before applying any clustering method on your data, it’s important to evaluate whether the data sets contains meaningful clusters (i.e.: non-random structures) or not. If yes, then how many... [Read more]



Articles - Cluster Validation Essentials

Assessing Clustering Tendency: Essentials

kassambara | 07/09/2017 | 6752 | Comments (6) | Cluster Validation Essentials | Unsupervised machine learning, Cluster evaluation

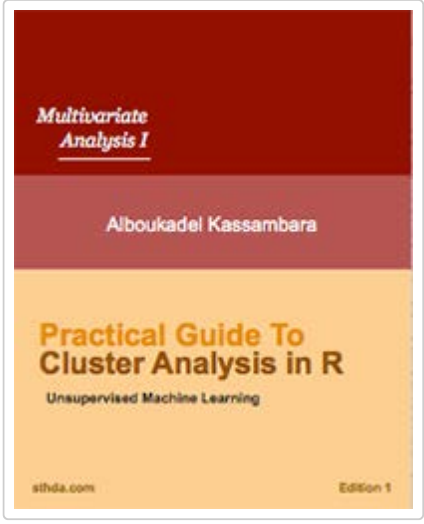
Before applying any clustering method on your data, it’s important to evaluate whether the data sets contains meaningful clusters (i.e.: non-random structures) or not. If yes, then how many clusters are there. This process is defined as the assessing of **clustering tendency** or the feasibility of the clustering analysis.

A big issue, in cluster analysis, is that clustering methods will return clusters even if the data does not contain any clusters. In other words, if you blindly apply a clustering method on a data set, it will divide the data into clusters because that is what it supposed to do.

In this chapter, we start by describing why we should evaluate the clustering tendency before applying any clustering method on a data. Next, we provide statistical and visual methods for assessing the clustering tendency.

Contents:

Related Book:



Practical Guide to Cluster Analysis in R

Required R packages

- *factoextra* for data visualization
- *clustertend* for statistical assessment clustering tendency

To install the two packages, type this:

```
install.packages(c("factoextra", "clustertend"))
```

Data preparation

We'll use two data sets:

- the built-in R data set iris.
- and a random data set generated from the iris data set.

The iris data sets look like this:

```
head(iris, 3)
```

```
##   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
```

We start by excluding the column "Species" at position 5

```
# Iris data set
df <- iris[, -5]
# Random data generated from the iris data set
random_df <- apply(df, 2,
                   function(x){runif(length(x), min(x), (max(x)))})
random_df <- as.data.frame(random_df)
```

```
# Standardize the data sets
df <- iris.scaled <- scale(df)
random_df <- scale(random_df)
```

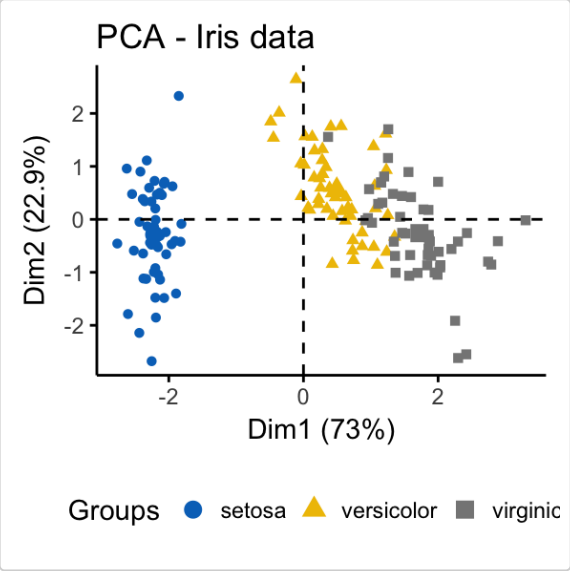
Visual inspection of the data

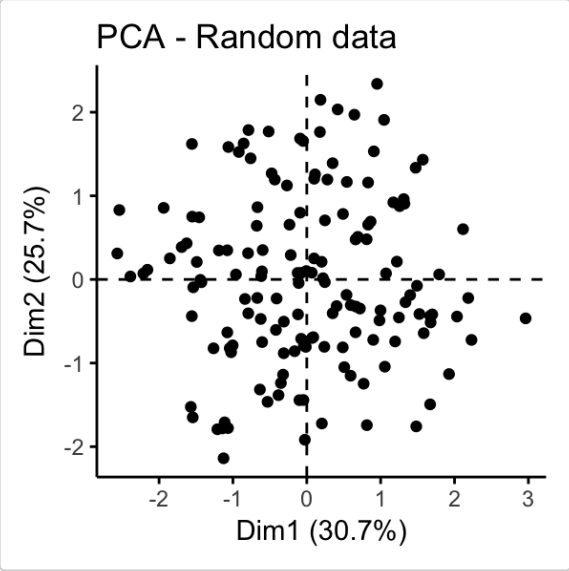
We start by visualizing the data to assess whether they contains any meaningful clusters.

As the data contain more than two variables, we need to reduce the dimensionality in order to plot a scatter plot. This can be done using principal component analysis (PCA) algorithm (R function: *prcomp*()). After performing PCA, we use the function *fviz_pca_ind*() [*factoextra* R package] to visualize the output.

The iris and the random data sets can be illustrated as follow:

```
library("factoextra")
# Plot faithful data set
fviz_pca_ind(prcomp(df), title = "PCA - Iris data",
             habillage = iris$Species,  palette = "jco",
             geom = "point", ggtheme = theme_classic(),
             legend = "bottom")
# Plot the random df
fviz_pca_ind(prcomp(random_df), title = "PCA - Random data",
             geom = "point", ggtheme = theme_classic())
```





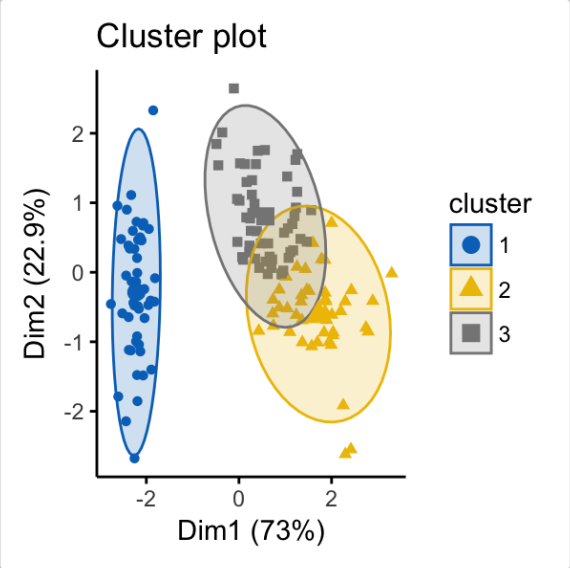
It can be seen that the iris data set contains 3 real clusters. However the randomly generated data set doesn't contain any meaningful clusters.

Why assessing clustering tendency?

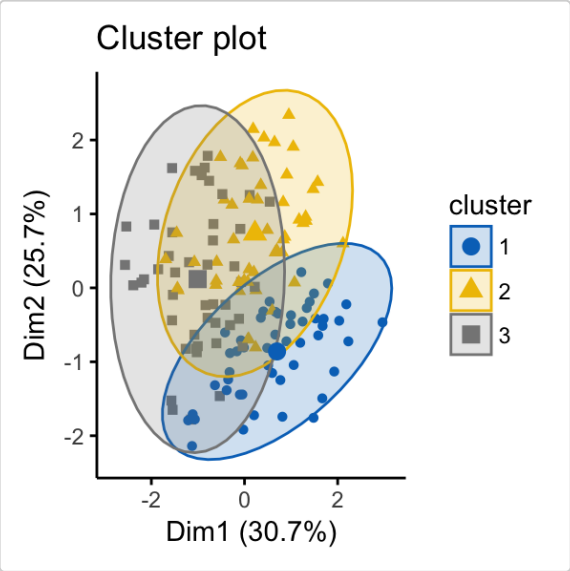
In order to illustrate why it's important to assess cluster tendency, we start by computing k-means clustering (Chapter @ref(kmeans-clustering)) and hierarchical clustering (Chapter @ref(agglomerative-clustering)) on the two data sets (the real and the random data). The function *fviz_cluster*() and *fviz_dend*() [in *factoextra* R package] will be used to visualize the results.

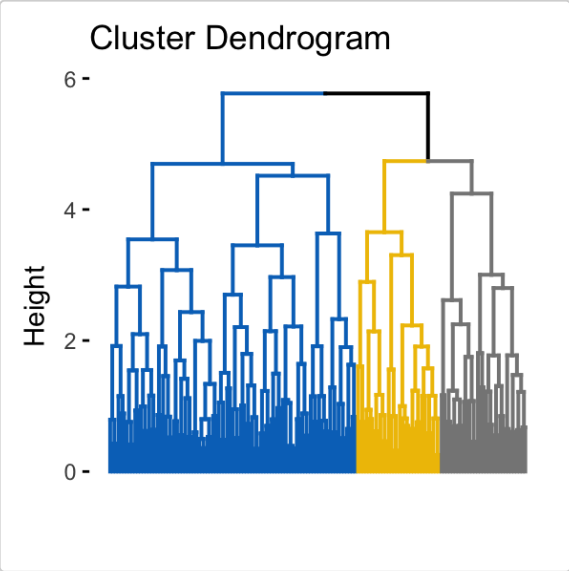
```
library(factoextra)
set.seed(123)
```

```
# K-means on iris dataset
km.res1 <- kmeans(df, 3)
fviz_cluster(list(data = df, cluster = km.res1$cluster),
             ellipse.type = "norm", geom = "point", stand = FALSE,
             palette = "jco", ggtheme = theme_classic())
```




```
# K-means on the random dataset
km.res2 <- kmeans(random_df, 3)
fviz_cluster(list(data = random_df, cluster = km.res2$cluster),
             ellipse.type = "norm", geom = "point", stand = FALSE,
             palette = "jco", ggtheme = theme_classic())
# Hierarchical clustering on the random dataset
fviz_dend(hclust(dist(random_df)), k = 3, k_colors = "jco",
          as.ggplot = TRUE, show_labels = FALSE)
```





It can be seen that the k-means algorithm and the hierarchical clustering impose a classification on the random uniformly distributed data set even if there are no meaningful clusters present in it. This is why, clustering tendency assessment methods should be used to evaluate the validity of clustering analysis. That is, whether a given data set contains meaningful clusters.

Methods for assessing clustering tendency

In this section, we'll describe two methods for evaluating the clustering tendency: i) a statistical (*Hopkins statistic*) and ii) a visual methods (*Visual Assessment of cluster Tendency* (VAT) algorithm).

Statistical methods

The *Hopkins statistic* (Lawson and Jurs 1990) is used to assess the clustering tendency of a data set by measuring the probability that a given data set is generated by a uniform data distribution. In other words, it tests the spatial randomness of the data.

For example, let D be a real data set. The Hopkins statistic can be calculated as follow:

1. Sample uniformly n points (p_1,…, p_n) from D.
2. Compute the distance, x_i, from each real point to each nearest neighbor: For each point $p_i \in D$, find it's nearest neighbor p_j; then compute the distance between p_i and p_j and denote it as $x_i = dist(p_i, p_j)$
3. Generate a simulated data set ($random_D$) drawn from a random uniform distribution with n points (q_1,…, q_n) and the same variation as the original real data set D.
4. Compute the distance, y_i from each artificial point to the nearest real data point: For each point $q_i \in random_D$, find it's nearest neighbor q_j in D; then compute the distance between q_i and q_j and denote it $y_i = dist(q_i, q_j)$
5. Calculate the Hopkins statistic (H) as the mean nearest neighbor distance in the random data set divided by the sum of the mean nearest neighbor distances in the real and across the simulated data set.

The formula is defined as follow:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

How to interpret the Hopkins statistics? If D were uniformly distributed, then $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i$ would be close to each other, and thus H would be about 0.5. However, if clusters are present in D, then the distances for artificial points ($\sum_{i=1}^{n} y_i$) would be substantially larger than for the real ones ($\sum_{i=1}^{n} x_i$) in expectation, and thus the value of H will increase (Han, Kamber, and Pei 2012).

A value for H higher than 0.75 indicates a clustering tendency at the 90% confidence level.

The null and the alternative hypotheses are defined as follow:

- **Null hypothesis**: the data set D is uniformly distributed (i.e., no meaningful clusters)
- **Alternative hypothesis**: the data set D is not uniformly distributed (i.e., contains meaningful clusters)

We can conduct the Hopkins Statistic test iteratively, using 0.5 as the threshold to reject the alternative hypothesis. That is, if H < 0.5, then it is unlikely that D has statistically significant clusters.

Put in other words, If the value of Hopkins statistic is close to 1, then we can reject the null hypothesis and conclude that the dataset D is significantly a clusterable data.

Here, we present two R functions / packages to statistically evaluate clustering tendency by computing the Hopkins statistics:

1. `get_clust_tendency()` function [in factoextra package]. It returns the Hopkins statistics as defined in the formula above. The result is a list containing two elements:
   - hopkins_stat
   - and plot
2. `hopkins()` function [in clustertend package]. It implements 1- the definition of H provided here.

In the R code below, we'll use the factoextra R package. Make sure that you have the latest version (or install: `devtools::install_github("kassambara/factoextra")`).

```
library(factoextra)
# Compute Hopkins statistic for iris dataset
res <- get_clust_tendency(df, n = nrow(df)-1, graph = FALSE)
res$hopkins_stat
```

```
## [1] 0.818
```

```
# Compute Hopkins statistic for a random dataset
res <- get_clust_tendency(random_df, n = nrow(random_df)-1,
                          graph = FALSE)
res$hopkins_stat
```

```
## [1] 0.466
```

It can be seen that the iris data set is highly clusterable (the **H** value = 0.82 which is far above the threshold 0.5). However the random_df data set is not clusterable (H = 0.47)

Visual methods

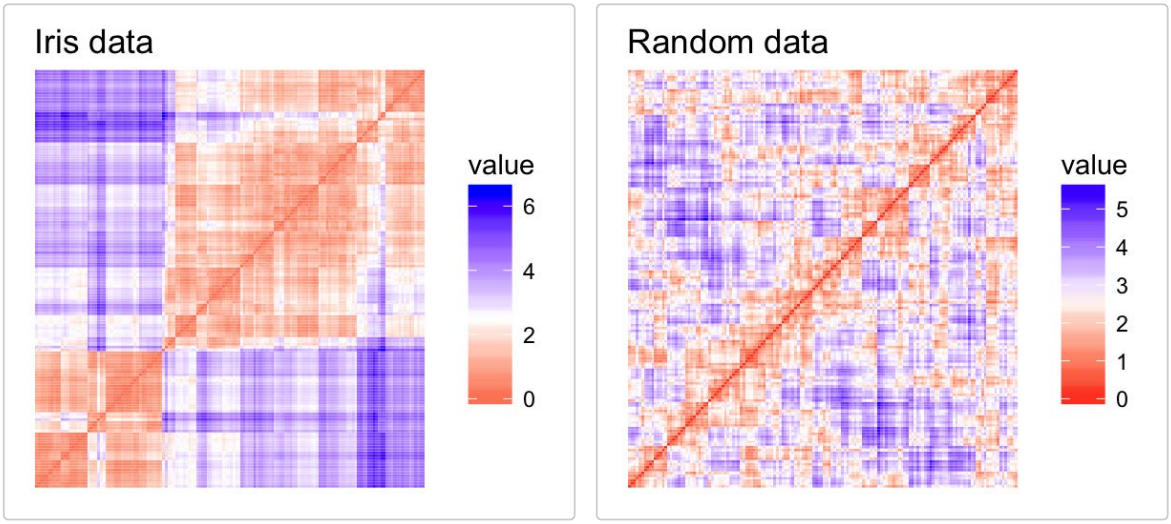
The algorithm of the visual assessment of cluster tendency (VAT) approach (Bezdek and Hathaway, 2002) is as follow:

The algorithm of VAT is as follow:

1. Compute the dissimilarity (DM) matrix between the objects in the data set using the Euclidean distance measure
2. Reorder the DM so that similar objects are close to one another. This process create an ordered dissimilarity matrix (ODM)
3. The ODM is displayed as an ordered dissimilarity image (ODI), which is the visual output of VAT

For the visual assessment of clustering tendency, we start by computing the dissimilarity matrix between observations using the function *dist*(). Next the function *fviz_dist*() [factoextra package] is used to display the dissimilarity matrix.

```
fviz_dist(dist(df), show_labels = FALSE)+
  labs(title = "Iris data")
fviz_dist(dist(random_df), show_labels = FALSE)+
  labs(title = "Random data")
```



- Red: high similarity (ie: low dissimilarity) | Blue: low similarity

The color level is proportional to the value of the dissimilarity between observations: pure red if $dist(x_i, x_j) = 0$ and pure blue if $dist(x_i, x_j) = 1$. Objects belonging to the same cluster are displayed in consecutive order.

The dissimilarity matrix image confirms that there is a cluster structure in the iris data set but not in the random one.

The VAT detects the clustering tendency in a visual form by counting the number of square shaped dark blocks along the diagonal in a VAT image.

Summary

In this article, we described how to assess clustering tendency using the Hopkins statistics and a visual method. After showing that a data is clusterable, the next step is to determine the number of optimal clusters in the data. This will be described in the next chapter.

References

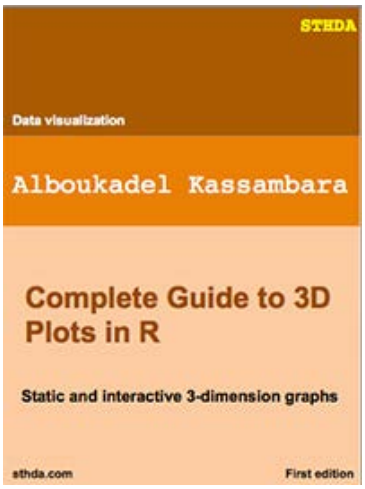
Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Concepts and Techniques*. 3rd ed. Boston: Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00016-2.

Lawson, Richard G., and Peter C. Jurs. 1990. “New Index for Clustering Tendency and Its Application to Chemical Problems.” *Journal of Chemical Information and Computer Sciences* 30 (1): 36–41. http://pubs.acs.org/doi/abs/10.1021/ci00065a010.

Last update : 21/10/2017
□ □ □ □ □ 0 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I’d be very grateful if you’d help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!



Articles - Cluster Validation Essentials

Determining The Optimal Number Of Clusters: 3 Must Know Methods

kassambara | 07/09/2017 | 81586 | Comments (6) | Cluster Validation Essentials | Unsupervised machine learning

Determining the **optimal number of clusters** in a data set is a fundamental issue in partitioning clustering, such as k-means clustering (Chapter @ref(kmeans-clustering)), which requires the user to specify the number of clusters k to be generated.

Unfortunately, there is no definitive answer to this question. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning. A simple and popular solution consists of inspecting the dendrogram produced using hierarchical clustering (Chapter @ref(agglomerative-clustering)) to see if it suggests a particular number of clusters. Unfortunately, this approach is also subjective.

> In this chapter, we'll describe different methods for determining the optimal number of clusters for k-means, k-medoids (PAM) and hierarchical clustering.

These methods include direct methods and statistical testing methods:

1. Direct methods: consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named *elbow* and *silhouette* methods, respectively.
2. Statistical testing methods: consists of comparing evidence against null hypothesis. An example is the *gap statistic*.

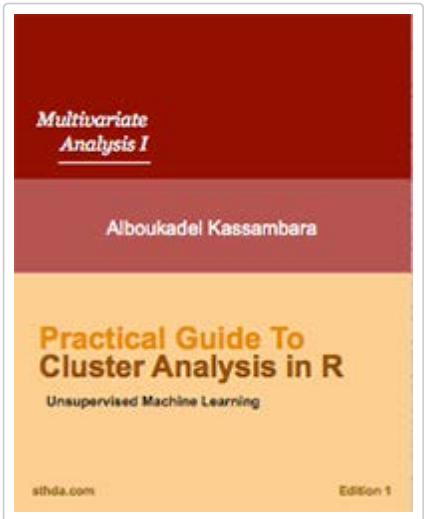
In addition to *elbow*, *silhouette* and *gap statistic* methods, there are more than thirty other indices and methods that have been published for identifying the optimal number of clusters. We'll provide R codes for computing all these 30 indices in order to decide the best number of clusters using the "majority rule".

For each of these methods:

- We'll describe the basic idea and the algorithm
- We'll provide easy-o-use R codes with many examples for determining the optimal number of clusters and visualizing the output.

Contents:

Related Book:



Practical Guide to Cluster Analysis in R

Elbow method

Recall that, the basic idea behind partitioning methods, such as k-means clustering (Chapter @ref(kmeans-clustering)), is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible.

The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

The optimal number of clusters can be defined as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of

clusters.

Note that, the elbow method is sometimes ambiguous. An alternative is the average silhouette method (Kaufman and Rousseeuw [1990]) which can be also used with any clustering approach.

Average silhouette method

The average silhouette approach we'll be described comprehensively in the chapter cluster validation statistics (Chapter @ref(cluster-validation-statistics)). Briefly, it measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990).

The algorithm is similar to the elbow method and can be computed as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the average silhouette of observations (*avg.sil*).
3. Plot the curve of *avg.sil* according to the number of clusters k.
4. The location of the maximum is considered as the appropriate number of clusters.

Gap statistic method

The *gap statistic* has been published by R. Tibshirani, G. Walther, and T. Hastie (Standford University, 2001). The approach can be applied to any clustering method.

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points.

The algorithm works as follow:

1. Cluster the observed data, varying the number of clusters from k = 1, …, k_{max}, and compute the corresponding total within intra-cluster variation W_k.
2. Generate B reference data sets with a random uniform distribution. Cluster each of these reference data sets with varying number of clusters k = 1, …, k_{max}, and compute the corresponding total within intra-cluster variation W_{kb}.
3. Compute the estimated gap statistic as the deviation of the observed W_k value from its expected value W_{kb} under the null hypothesis: $Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}^{*}) - log(W_k)$. Compute also the standard deviation of the statistics.
4. Choose the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at k+1: $Gap(k) \geq Gap(k+1) - s_{k+1}$.

Note that, using B = 500 gives quite precise results so that the gap plot is basically unchanged after an another run.

Computing the number of clusters using R

In this section, we'll describe two functions for determining the optimal number of clusters:

1. *fviz_nbclust*() function [in *factoextra* R package]: It can be used to compute the three different methods [elbow, silhouette and gap statistic] for any partitioning clustering methods [K-means, K-medoids (PAM), CLARA, HCUT]. Note that the *hcut*() function is available only in factoextra package.It computes hierarchical clustering and cut the tree in k pre-specified clusters.
2. *NbClust*() function [in *NbClust* R package] (Charrad et al. 2014): It provides 30 indices for determining the relevant number of clusters and proposes to users the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods. It can simultaneously computes all the indices and determine the number of clusters in a single function call.

Required R packages

We'll use the following R packages:

- *factoextra* to determine the optimal number clusters for a given clustering methods and for data visualization.
- *NbClust* for computing about 30 methods at once, in order to find the optimal number of clusters.

To install the packages, type this:

```
pkgs <- c("factoextra",  "NbClust")
install.packages(pkgs)
```

Load the packages as follow:

```
library(factoextra)
library(NbClust)
```

Data preparation

We'll use the USArrests data as a demo data set. We start by standardizing the data to make variables comparable.

```
# Standardize the data
df <- scale(USArrests)
head(df)
```

```
##            Murder Assault UrbanPop     Rape
## Alabama    1.2426   0.783   -0.521 -0.00342
## Alaska     0.5079   1.107   -1.212  2.48420
## Arizona    0.0716   1.479    0.999  1.04288
## Arkansas   0.2323   0.231   -1.074 -0.18492
## California 0.2783   1.263    1.759  2.06782
## Colorado   0.0257   0.399    0.861  1.86497
```

fviz_nbclust() function: Elbow, Silhouhette and Gap statistic methods

The simplified format is as follow:

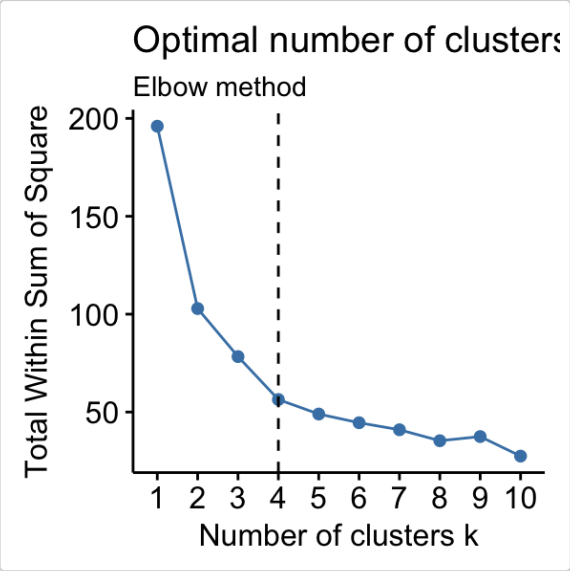
```
fviz_nbclust(x, FUNcluster, method = c("silhouette", "wss", "gap_stat"))
```

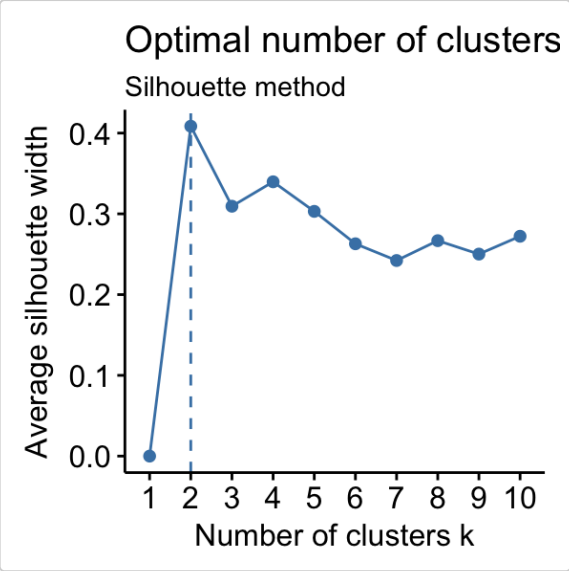
- **x**: numeric matrix or data frame
- **FUNcluster**: a partitioning function. Allowed values include kmeans, pam, clara and hcut (for hierarchical clustering).
- **method**: the method to be used for determining the optimal number of clusters.

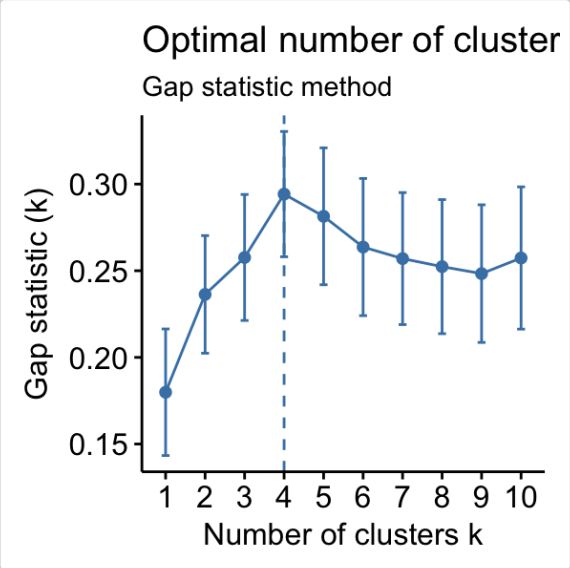
The R code below determine the optimal number of clusters for k-means clustering:

```
# Elbow method
fviz_nbclust(df, kmeans, method = "wss") +
    geom_vline(xintercept = 4, linetype = 2)+
  labs(subtitle = "Elbow method")
# Silhouette method
fviz_nbclust(df, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
# Gap statistic
# nboot = 50 to keep the function speedy.
# recommended value: nboot= 500 for your analysis.
# Use verbose = FALSE to hide computing progression.
set.seed(123)
fviz_nbclust(df, kmeans, nstart = 25,  method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
```

```
## Clustering k = 1,2,..., K.max (= 10): .. done
## Bootstrapping, b = 1,2,..., B (= 50)  [one "." per sample]:
## .................................................. 50
```







- Elbow method: 4 clusters solution suggested
- Silhouette method: 2 clusters solution suggested
- Gap statistic method: 4 clusters solution suggested

According to these observations, it's possible to define k = 4 as the optimal number of clusters in the data.

The disadvantage of elbow and average silhouette methods is that, they measure a global clustering characteristic only. A more sophisticated method is to use the gap statistic which provides a statistical procedure to formalize the elbow/silhouette heuristic in order to estimate the optimal number of clusters.

NbClust() function: 30 indices for choosing the best number of clusters

The simplified format of the function *NbClust*() is:

```
NbClust(data = NULL, diss = NULL, distance = "euclidean",
        min.nc = 2, max.nc = 15, method = NULL)
```

- **data**: matrix
- **diss**: dissimilarity matrix to be used. By default, diss=NULL, but if it is replaced by a dissimilarity matrix, distance should be "NULL"
- **distance**: the distance measure to be used to compute the dissimilarity matrix. Possible values include "euclidean", "manhattan" or "NULL".
- **min.nc, max.nc**: minimal and maximal number of clusters, respectively
- **method**: The cluster analysis method to be used including "ward.D", "ward.D2", "single", "complete", "average", "kmeans" and more.

- To compute *NbClust*() for kmeans, use method = "kmeans".
- To compute *NbClust*() for hierarchical clustering, method should be one of c("ward.D", "ward.D2", "single", "complete", "average").

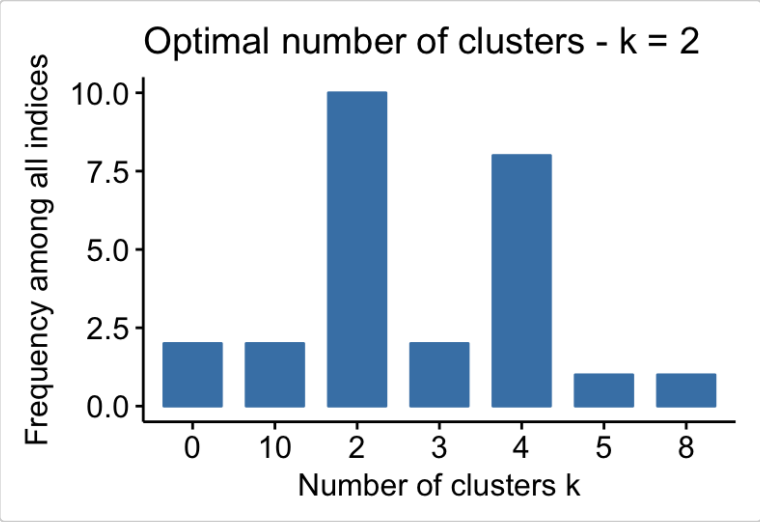
The R code below computes *NbClust*() for k-means:

```
library("NbClust")
nb <- NbClust(df, distance = "euclidean", min.nc = 2,
        max.nc = 10, method = "kmeans")
```

The result of NbClust using the function *fviz_nbclust*() [in *factoextra*], as follow:

```
library("factoextra")
fviz_nbclust(nb)
```

```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 10 proposed  2 as the best number of clusters
## * 2 proposed  3 as the best number of clusters
## * 8 proposed  4 as the best number of clusters
## * 1 proposed  5 as the best number of clusters
## * 1 proposed  8 as the best number of clusters
## * 2 proposed  10 as the best number of clusters
##
## Conclusion
## =========================
## * According to the majority rule, the best number of clusters is  2 .
```



- ….
- 2 proposed 0 as the best number of clusters
- 10 indices proposed 2 as the best number of clusters.
- 2 proposed 3 as the best number of clusters.
- 8 proposed 4 as the best number of clusters.

According to the majority rule, the best number of clusters is 2.

Summary

In this article, we described different methods for choosing the optimal number of clusters in a data set. These methods include the elbow, the silhouette and the gap statistic methods.

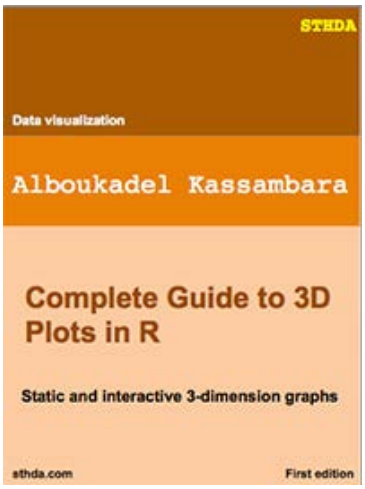
We demonstrated how to compute these methods using the R function *fviz_nbclust*() [in *factoextra* R package]. Additionally, we described the package *NbClust*(), which can be used to compute simultaneously many other indices and methods for determining the number of clusters.

After choosing the number of clusters k, the next step is to perform partitioning clustering as described at: k-means clustering (Chapter @ref(kmeans-clustering)).

References

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.” *Journal of Statistical Software* 61: 1–36. http://www.jstatsoft.org/v61/i06/paper.

Kaufman, Leonard, and Peter Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*.



Articles - Cluster Validation Essentials

Determining The Optimal Number Of Clusters: 3 Must Know Methods

kassambara | 07/09/2017 | 81586 | Comments (6) | Cluster Validation Essentials | Unsupervised machine learning

Determining the **optimal number of clusters** in a data set is a fundamental issue in partitioning clustering, such as k-means clustering (Chapter @ref(kmeans-clustering)), which requires the user to specify the number of clusters k to be generated.

Unfortunately, there is no definitive answer to this question. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning. A simple and popular solution consists of inspecting the dendrogram produced using hierarchical clustering (Chapter @ref(agglomerative-clustering)) to see if it suggests a particular number of clusters. Unfortunately, this approach is also subjective.

> In this chapter, we'll describe different methods for determining the optimal number of clusters for k-means, k-medoids (PAM) and hierarchical clustering.

These methods include direct methods and statistical testing methods:

1. Direct methods: consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named *elbow* and *silhouette* methods, respectively.
2. Statistical testing methods: consists of comparing evidence against null hypothesis. An example is the *gap statistic*.

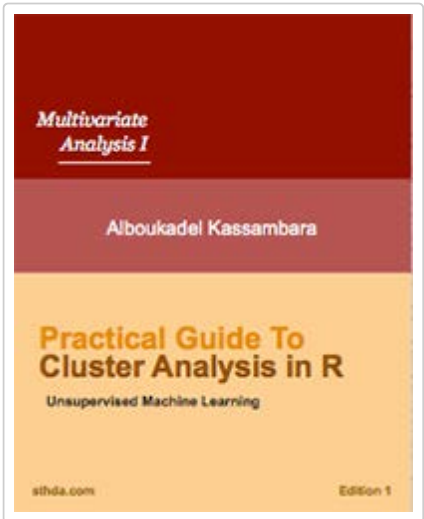
In addition to *elbow*, *silhouette* and *gap statistic* methods, there are more than thirty other indices and methods that have been published for identifying the optimal number of clusters. We'll provide R codes for computing all these 30 indices in order to decide the best number of clusters using the "majority rule".

For each of these methods:

- We'll describe the basic idea and the algorithm
- We'll provide easy-o-use R codes with many examples for determining the optimal number of clusters and visualizing the output.

Contents:

Related Book:



Practical Guide to Cluster Analysis in R

Elbow method

Recall that, the basic idea behind partitioning methods, such as k-means clustering (Chapter @ref(kmeans-clustering)), is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible.

The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.

The optimal number of clusters can be defined as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of

clusters.

Note that, the elbow method is sometimes ambiguous. An alternative is the average silhouette method (Kaufman and Rousseeuw [1990]) which can be also used with any clustering approach.

Average silhouette method

The average silhouette approach we'll be described comprehensively in the chapter cluster validation statistics (Chapter @ref(cluster-validation-statistics)). Briefly, it measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990).

The algorithm is similar to the elbow method and can be computed as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the average silhouette of observations (*avg.sil*).
3. Plot the curve of *avg.sil* according to the number of clusters k.
4. The location of the maximum is considered as the appropriate number of clusters.

Gap statistic method

The *gap statistic* has been published by R. Tibshirani, G. Walther, and T. Hastie (Standford University, 2001). The approach can be applied to any clustering method.

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points.

The algorithm works as follow:

1. Cluster the observed data, varying the number of clusters from k = 1, …, k_{max}, and compute the corresponding total within intra-cluster variation W_k.
2. Generate B reference data sets with a random uniform distribution. Cluster each of these reference data sets with varying number of clusters k = 1, …, k_{max}, and compute the corresponding total within intra-cluster variation W_{kb}.
3. Compute the estimated gap statistic as the deviation of the observed W_k value from its expected value W_{kb} under the null hypothesis: $Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log(W^*_{kb}) - log(W_k)$. Compute also the standard deviation of the statistics.
4. Choose the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at k+1: $Gap(k) \geq Gap(k+1) - s_{k+1}$.

Note that, using B = 500 gives quite precise results so that the gap plot is basically unchanged after an another run.

Computing the number of clusters using R

In this section, we'll describe two functions for determining the optimal number of clusters:

1. *fviz_nbclust*() function [in *factoextra* R package]: It can be used to compute the three different methods [elbow, silhouette and gap statistic] for any partitioning clustering methods [K-means, K-medoids (PAM), CLARA, HCUT]. Note that the *hcut*() function is available only in factoextra package.It computes hierarchical clustering and cut the tree in k pre-specified clusters.
2. *NbClust*() function [in *NbClust* R package] (Charrad et al. 2014): It provides 30 indices for determining the relevant number of clusters and proposes to users the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods. It can simultaneously computes all the indices and determine the number of clusters in a single function call.

Required R packages

We'll use the following R packages:

- *factoextra* to determine the optimal number clusters for a given clustering methods and for data visualization.
- *NbClust* for computing about 30 methods at once, in order to find the optimal number of clusters.

To install the packages, type this:

```
pkgs <- c("factoextra",  "NbClust")
install.packages(pkgs)
```

Load the packages as follow:

```
library(factoextra)
library(NbClust)
```

Data preparation

We'll use the USArrests data as a demo data set. We start by standardizing the data to make variables comparable.

```
# Standardize the data
df <- scale(USArrests)
head(df)
```

```
##            Murder Assault UrbanPop     Rape
## Alabama    1.2426   0.783   -0.521 -0.00342
## Alaska     0.5079   1.107   -1.212  2.48420
## Arizona    0.0716   1.479    0.999  1.04288
## Arkansas   0.2323   0.231   -1.074 -0.18492
## California 0.2783   1.263    1.759  2.06782
## Colorado   0.0257   0.399    0.861  1.86497
```

fviz_nbclust() function: Elbow, Silhouhette and Gap statistic methods

The simplified format is as follow:

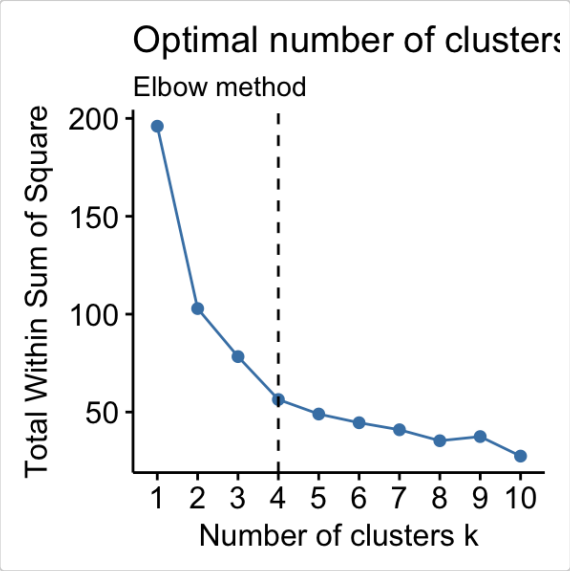
```
fviz_nbclust(x, FUNcluster, method = c("silhouette", "wss", "gap_stat"))
```

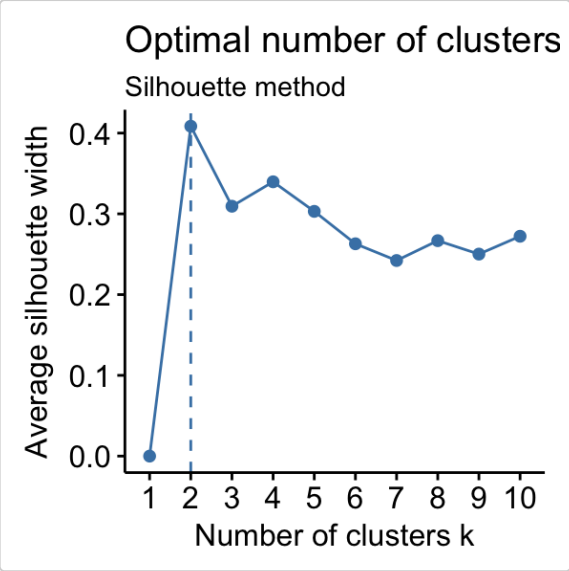
- **x**: numeric matrix or data frame
- **FUNcluster**: a partitioning function. Allowed values include kmeans, pam, clara and hcut (for hierarchical clustering).
- **method**: the method to be used for determining the optimal number of clusters.

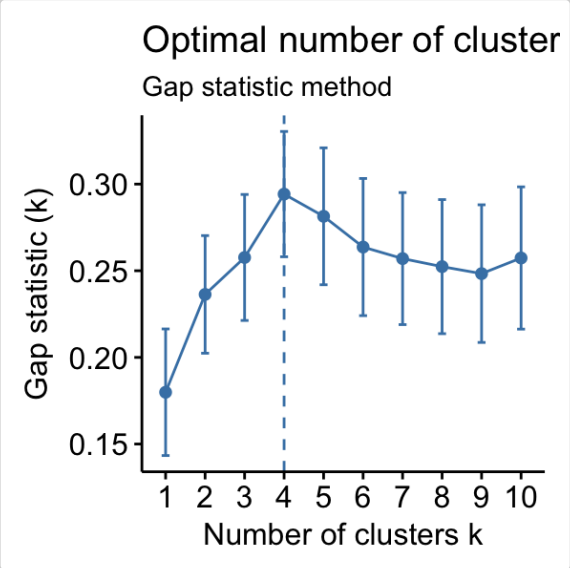
The R code below determine the optimal number of clusters for k-means clustering:

```
# Elbow method
fviz_nbclust(df, kmeans, method = "wss") +
    geom_vline(xintercept = 4, linetype = 2)+
  labs(subtitle = "Elbow method")
# Silhouette method
fviz_nbclust(df, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
# Gap statistic
# nboot = 50 to keep the function speedy.
# recommended value: nboot= 500 for your analysis.
# Use verbose = FALSE to hide computing progression.
set.seed(123)
fviz_nbclust(df, kmeans, nstart = 25,  method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
```

```
## Clustering k = 1,2,..., K.max (= 10): .. done
## Bootstrapping, b = 1,2,..., B (= 50)  [one "." per sample]:
## .................................................. 50
```







- Elbow method: 4 clusters solution suggested
- Silhouette method: 2 clusters solution suggested
- Gap statistic method: 4 clusters solution suggested

According to these observations, it's possible to define k = 4 as the optimal number of clusters in the data.

The disadvantage of elbow and average silhouette methods is that, they measure a global clustering characteristic only. A more sophisticated method is to use the gap statistic which provides a statistical procedure to formalize the elbow/silhouette heuristic in order to estimate the optimal number of clusters.

NbClust() function: 30 indices for choosing the best number of clusters

The simplified format of the function *NbClust*() is:

```
NbClust(data = NULL, diss = NULL, distance = "euclidean",
        min.nc = 2, max.nc = 15, method = NULL)
```

- **data**: matrix
- **diss**: dissimilarity matrix to be used. By default, diss=NULL, but if it is replaced by a dissimilarity matrix, distance should be "NULL"
- **distance**: the distance measure to be used to compute the dissimilarity matrix. Possible values include "euclidean", "manhattan" or "NULL".
- **min.nc, max.nc**: minimal and maximal number of clusters, respectively
- **method**: The cluster analysis method to be used including "ward.D", "ward.D2", "single", "complete", "average", "kmeans" and more.

- To compute *NbClust*() for kmeans, use method = "kmeans".
- To compute *NbClust*() for hierarchical clustering, method should be one of c("ward.D", "ward.D2", "single", "complete", "average").

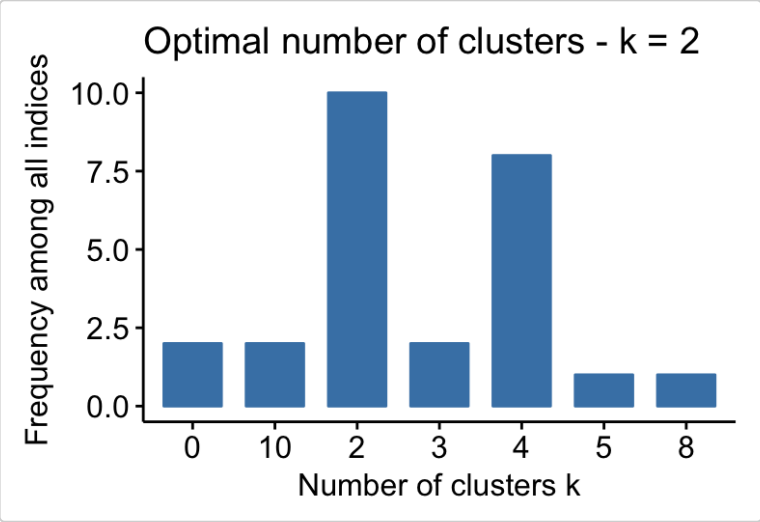
The R code below computes *NbClust*() for k-means:

```
library("NbClust")
nb <- NbClust(df, distance = "euclidean", min.nc = 2,
        max.nc = 10, method = "kmeans")
```

The result of NbClust using the function *fviz_nbclust*() [in *factoextra*], as follow:

```
library("factoextra")
fviz_nbclust(nb)
```

```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 10 proposed  2 as the best number of clusters
## * 2 proposed  3 as the best number of clusters
## * 8 proposed  4 as the best number of clusters
## * 1 proposed  5 as the best number of clusters
## * 1 proposed  8 as the best number of clusters
## * 2 proposed  10 as the best number of clusters
##
## Conclusion
## =========================
## * According to the majority rule, the best number of clusters is  2 .
```



- ….
- 2 proposed 0 as the best number of clusters
- 10 indices proposed 2 as the best number of clusters.
- 2 proposed 3 as the best number of clusters.
- 8 proposed 4 as the best number of clusters.

According to the majority rule, the best number of clusters is 2.

Summary

In this article, we described different methods for choosing the optimal number of clusters in a data set. These methods include the elbow, the silhouette and the gap statistic methods.

We demonstrated how to compute these methods using the R function *fviz_nbclust*() [in *factoextra* R package]. Additionally, we described the package *NbClust*(), which can be used to compute simultaneously many other indices and methods for determining the number of clusters.

After choosing the number of clusters k, the next step is to perform partitioning clustering as described at: k-means clustering (Chapter @ref(kmeans-clustering)).

References

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.” *Journal of Statistical Software* 61: 1–36. http://www.jstatsoft.org/v61/i06/paper.

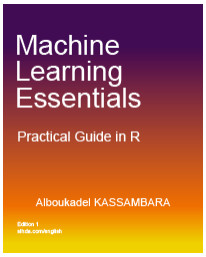
Kaufman, Leonard, and Peter Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*.

Last update : 24/09/2017
□ □ □ □ □ 0 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

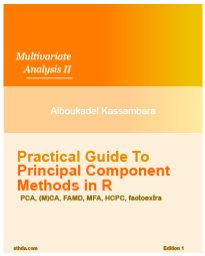
Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!



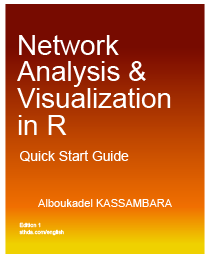
Machine Learning Essentials: Practical Guide in R

Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R

R Graphics Essentials for Great Data Visualization



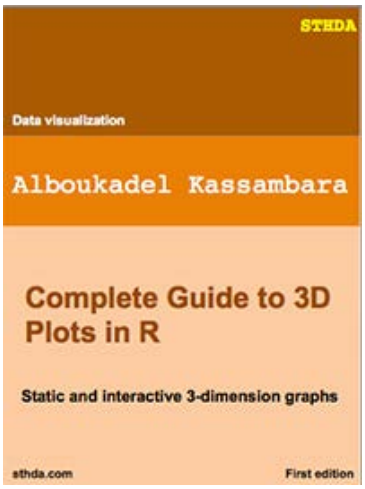
Network Analysis and Visualization in R

More books on R and data science

Comments

The fields marked with a * are required !

Add a comment
Name
Visitor
Message



Articles - Cluster Validation Essentials

Cluster Validation Statistics: Must Know Methods

kassambara | 07/09/2017 | 15728 | Post a comment | Cluster Validation Essentials | Unsupervised machine learning, Multivariate Analysis

The term **cluster validation** is used to design the procedure of evaluating the goodness of clustering algorithm results. This is important to avoid finding patterns in a random data, as well as, in the situation where you want to compare two clustering algorithms.

Generally, clustering validation statistics can be categorized into 3 classes (Charrad et al. 2014,Brock et al. (2008), Theodoridis and Koutroumbas (2008)):

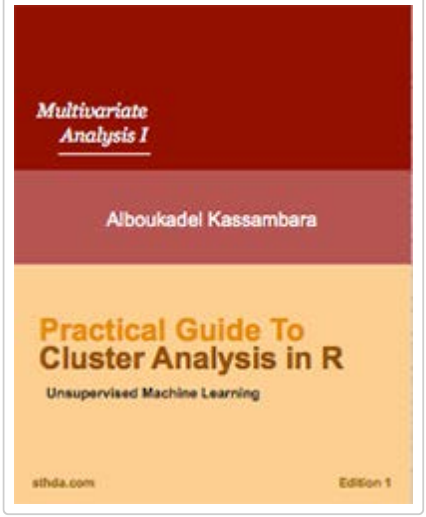
1. **Internal cluster validation**, which uses the internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. It can be also used for estimating the number of clusters and the appropriate clustering algorithm without any external data.
2. **External cluster validation**, which consists in comparing the results of a cluster analysis to an externally known result, such as externally provided class labels. It measures the extent to which cluster labels match externally supplied class labels. Since we know the “true” cluster number in advance, this approach is mainly used for selecting the right clustering algorithm for a specific data set.
3. **Relative cluster validation**, which evaluates the clustering structure by varying different parameter values for the same algorithm (e.g.,: varying the number of clusters k). It’s generally used for determining the optimal number of clusters.

In this chapter, we start by describing the different methods for clustering validation. Next, we’ll demonstrate how to compare the quality of clustering results obtained with different clustering algorithms. Finally, we’ll provide R scripts for validating clustering results.

In all the examples presented here, we’ll apply k-means, PAM and hierarchical clustering. Note that, the functions used in this article can be applied to evaluate the validity of any other clustering methods.

Contents:

Related Book:



Practical Guide to Cluster Analysis in R

Internal measures for cluster validation

In this section, we describe the most widely used clustering validation indices. Recall that the goal of partitioning clustering algorithms (Part @ref(partitioning-clustering)) is to split the data set into clusters of objects, such that:

- the objects in the same cluster are similar as much as possible,
- and the objects in different clusters are highly distinct

That is, we want the average distance within cluster to be as small as possible; and the average distance between clusters to be as large as possible.

Internal validation measures reflect often the **compactness**, the **connectedness** and the **separation** of the cluster partitions.

1. **Compactness** or cluster cohesion: Measures how close are the objects within the same cluster. A lower **within-cluster variation** is an indicator of a good compactness (i.e., a good clustering). The different indices for evaluating the compactness of clusters are base on distance measures such as the cluster-wise within average/median distances between observations.
2. **Separation**: Measures how well-separated a cluster is from other clusters. The indices used as separation measures include:

distances between cluster centers
- the pairwise minimum distances between objects in different clusters

3. **Connectivity**: corresponds to what extent items are placed in the same cluster as their nearest neighbors in the data space. The connectivity has a value between 0 and infinity and should be minimized.

Generally most of the indices used for internal clustering validation combine compactness and separation measures as follow:

$$Index = \frac{(\alpha \times Separation)}{(\beta \times Compactness)}$$

Where α and β are weights.

In this section, we'll describe the two commonly used indices for assessing the goodness of clustering: the **silhouette width** and the **Dunn index**. These internal measure can be used also to determine the optimal number of clusters in the data.

Silhouette coefficient

The silhouette analysis measures how well an observation is clustered and it estimates the **average distance between clusters**. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters.

For each observation i, the silhouette width s_i is calculated as follows:

1. For each observation i, calculate the average dissimilarity a_i between i and all other points of the cluster to which i belongs.
2. For all other clusters C, to which i does not belong, calculate the average dissimilarity $d(i, C)$ of i to all observations of C. The smallest of these $d(i, C)$ is defined as $b_i = \min_C d(i, C)$. The value of b_i can be seen as the dissimilarity between i and its "neighbor" cluster, i.e., the nearest one to which it does not belong.
3. Finally the silhouette width of the observation i is defined by the formula: $S_i = (b_i - a_i)/max(a_i, b_i)$.

Silhouette width can be interpreted as follow:

- Observations with a large S_i (almost 1) are very well clustered.
- A small S_i (around 0) means that the observation lies between two clusters.
- Observations with a negative S_i are probably placed in the wrong cluster.

Dunn index

The **Dunn index** is another internal clustering validation measure which can be computed as follow:

1. For each cluster, compute the distance between each of the objects in the cluster and the objects in the other clusters
2. Use the minimum of this pairwise distance as the inter-cluster separation (*min.separation*)
3. For each cluster, compute the distance between the objects in the same cluster.
4. Use the maximal intra-cluster distance (i.e maximum diameter) as the intra-cluster compactness
5. Calculate the *Dunn index* (D) as follow:

$$D = \frac{min.separation}{max.diameter}$$

If the data set contains compact and well-separated clusters, the diameter of the clusters is expected to be small and the distance between the clusters is expected to be large. Thus, Dunn index should be maximized.

External measures for clustering validation

The aim is to compare the identified clusters (by k-means, pam or hierarchical clustering) to an external reference.

It's possible to quantify the agreement between partitioning clusters and external reference using either the corrected *Rand index* and *Meila's variation index VI*, which are implemented in the R function *cluster.stats*()[*fpc* package].

The corrected *Rand index* varies from -1 (no agreement) to 1 (perfect agreement).

External clustering validation, can be used to select suitable clustering algorithm for a given data set.

Computing cluster validation statistics in R

Required R packages

The following R packages are required in this chapter:

- *factoextra* for data visualization
- *fpc* for computing clustering validation statistics
- *NbClust* for determining the optimal number of clusters in the data set.

- Install the packages:

```
install.packages(c("factoextra", "fpc", "NbClust"))
```

- Load the packages:

```
library(factoextra)
library(fpc)
library(NbClust)
```

Data preparation

We'll use the built-in R data set iris:

```
# Excluding the column "Species" at position 5
df <- iris[, -5]
# Standardize
df <- scale(df)
```

Clustering analysis

We'll use the function *eclust*() [enhanced clustering, in *factoextra*] which provides several advantages:

- It simplifies the workflow of clustering analysis
- It can be used to compute hierarchical clustering and partitioning clustering in a single line function call
- Compared to the standard partitioning functions (kmeans, pam, clara and fanny) which requires the user to specify the optimal number of clusters, the function *eclust*() computes automatically the gap statistic for estimating the right number of clusters.

- It provides silhouette information for all partitioning methods and hierarchical clustering
- It draws beautiful graphs using ggplot2

The simplified format the *eclust*() function is as follow:

```
eclust(x, FUNcluster = "kmeans", hc_metric = "euclidean", ...)
```

- **x**: numeric vector, data matrix or data frame
- **FUNcluster**: a clustering function including “kmeans”, “pam”, “clara”, “fanny”, “hclust”, “agnes” and “diana”. Abbreviation is allowed.
- **hc_metric**: character string specifying the metric to be used for calculating dissimilarities between observations. Allowed values are those accepted by the function *dist*() [including “euclidean”, “manhattan”, “maximum”, “canberra”, “binary”, “minkowski”] and correlation based distance measures [“pearson”, “spearman” or “kendall”]. Used only when FUNcluster is a hierarchical clustering function such as one of “hclust”, “agnes” or “diana”.
- **…**: other arguments to be passed to FUNcluster.

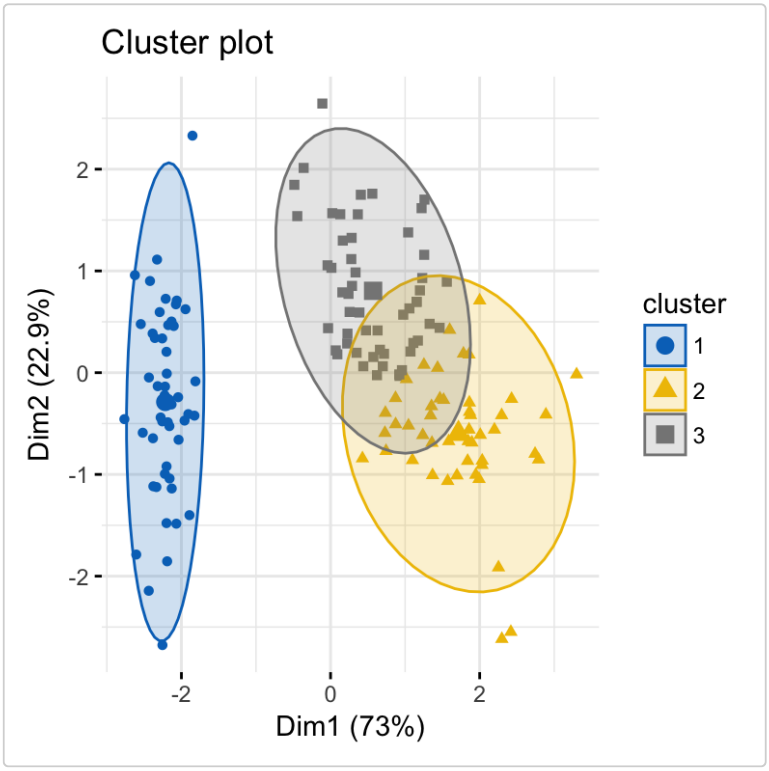
The function **eclust()** returns an object of class **eclust** containing the result of the standard function used (e.g., kmeans, pam, hclust, agnes, diana, etc.).

It includes also:

- **cluster**: the cluster assignment of observations after cutting the tree
- **nbclust**: the number of clusters
- **silinfo**: the silhouette information of observations
- **size**: the size of clusters
- **data**: a matrix containing the original or the standardized data (if stand = TRUE)
- **gap_stat**: containing gap statistics

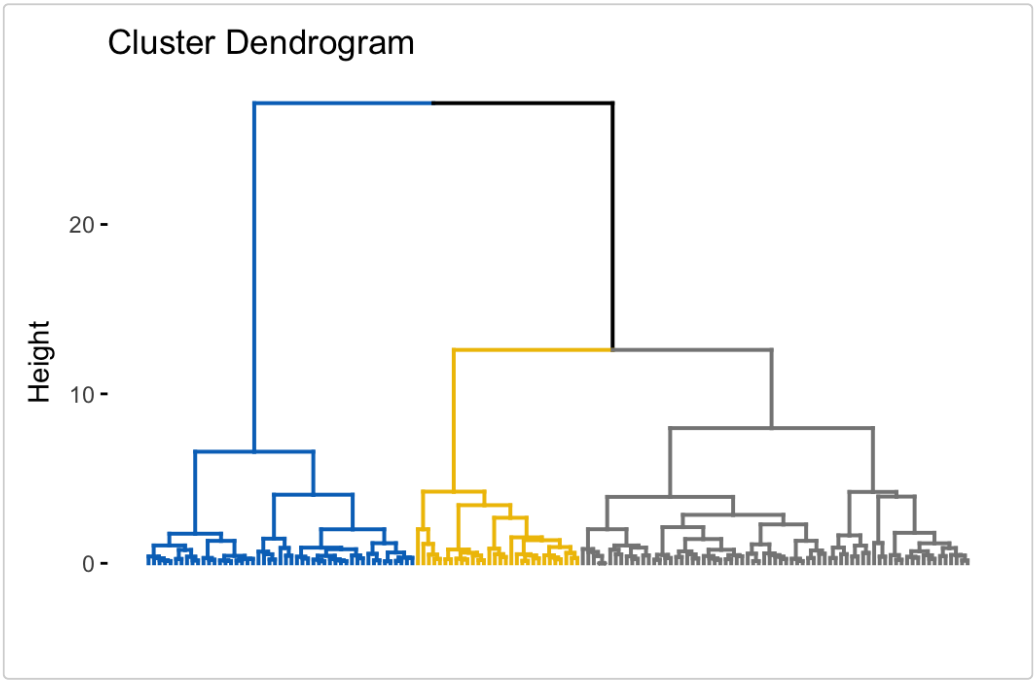
To compute a partitioning clustering, such as k-means clustering with k = 3, type this:

```
# K-means clustering
km.res <- eclust(df, "kmeans", k = 3, nstart = 25, graph = FALSE)
# Visualize k-means clusters
fviz_cluster(km.res, geom = "point", ellipse.type = "norm",
             palette = "jco", ggtheme = theme_minimal())
```



To compute a hierarchical clustering, use this:

```
# Hierarchical clustering
hc.res <- eclust(df, "hclust", k = 3, hc_metric = "euclidean",
                 hc_method = "ward.D2", graph = FALSE)
# Visualize dendrograms
fviz_dend(hc.res, show_labels = FALSE,
          palette = "jco", as.ggplot = TRUE)
```



Cluster validation

Silhouette plot

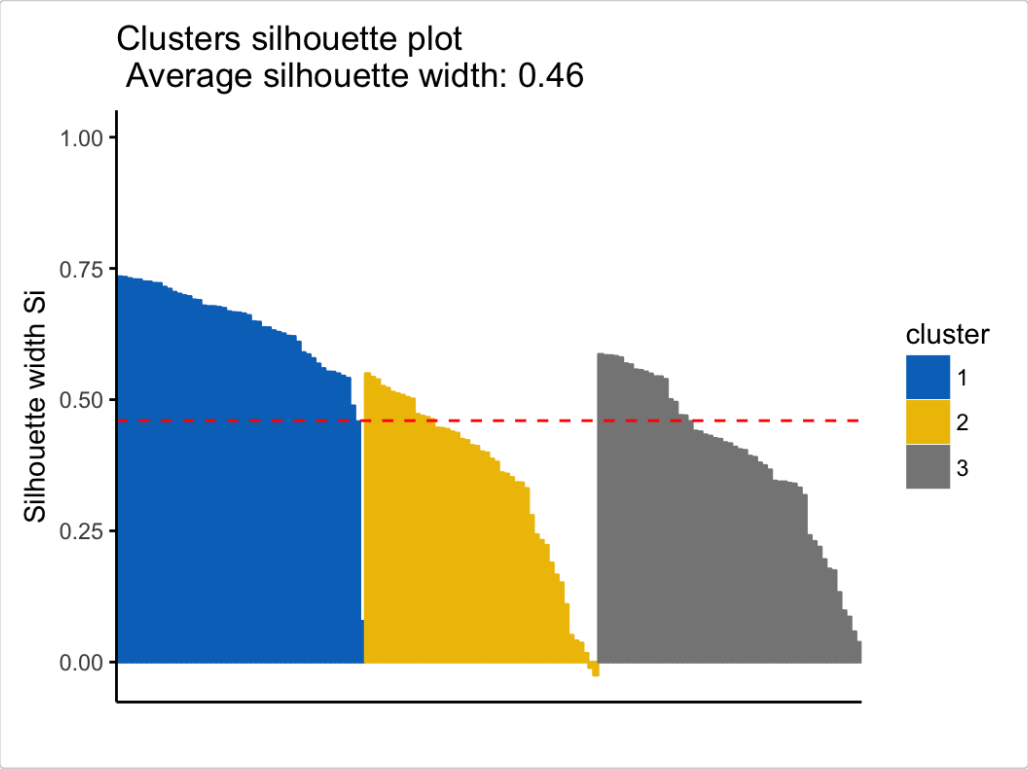
Recall that the silhouette coefficient (S_i) measures how similar an object i is to the the other objects in its own cluster versus those in the neighbor cluster. S_i values range from 1 to - 1:

- A value of S_i close to 1 indicates that the object is well clustered. In the other words, the object i is similar to the other objects in its group.
- A value of S_i close to -1 indicates that the object is poorly clustered, and that assignment to some other cluster would probably improve the overall results.

It's possible to draw silhouette coefficients of observations using the function *fviz_silhouette*() [*factoextra* package], which will also print a summary of the silhouette analysis output. To avoid this, you can use the option *print.summary = FALSE*.

```
fviz_silhouette(km.res, palette = "jco",
                ggtheme = theme_classic())
```

```
##   cluster size ave.sil.width
## 1       1   50          0.64
## 2       2   47          0.35
## 3       3   53          0.39
```



Silhouette information can be extracted as follow:

```
# Silhouette information
silinfo <- km.res$silinfo
names(silinfo)
# Silhouette widths of each observation
head(silinfo$widths[, 1:3], 10)
# Average silhouette width of each cluster
silinfo$clus.avg.widths
# The total average (mean of all individual silhouette widths)
silinfo$avg.width
```

```
# The size of each clusters
km.res$size
```

It can be seen that several samples, in cluster 2, have a negative silhouette coefficient. This means that they are not in the right cluster. We can find the name of these samples and determine the clusters they are closer (neighbor cluster), as follow:

```
# Silhouette width of observation
sil <- km.res$silinfo$widths[, 1:3]
# Objects with negative silhouette
neg_sil_index <- which(sil[, 'sil_width'] < 0)
sil[neg_sil_index, , drop = FALSE]
```

```
##     cluster neighbor sil_width
## 112       2        3   -0.0106
## 128       2        3   -0.0249
```

Computing Dunn index and other cluster validation statistics

The function *cluster.stats*() [*fpc* package] and the function *NbClust()* [in *NbClust* package] can be used to compute *Dunn index* and many other indices.

The simplified format is:

```
cluster.stats(d = NULL, clustering, al.clustering = NULL)
```

- **d**: a distance object between cases as generated by the **dist()** function
- **clustering**: vector containing the cluster number of each observation
- **alt.clustering**: vector such as for clustering, indicating an alternative clustering

The function *cluster.stats*() returns a list containing many components useful for analyzing the intrinsic characteristics of a clustering:

- **cluster.number**: number of clusters
- **cluster.size**: vector containing the number of points in each cluster
- **average.distance**, **median.distance**: vector containing the cluster-wise within average/median distances
- **average.between**: average distance between clusters. We want it to be as large as possible
- **average.within**: average distance within clusters. We want it to be as small as possible
- **clus.avg.silwidths**: vector of cluster average silhouette widths. Recall that, the **silhouette width** is also an estimate of the average distance between clusters. Its value is comprised between 1 and -1 with a value of 1 indicating a very good cluster.
- **within.cluster.ss**: a generalization of the within clusters sum of squares (k-means objective function), which is obtained if d is a Euclidean distance matrix.
- **dunn, dunn2**: Dunn index
- **corrected.rand, vi**: Two indexes to assess the similarity of two clustering: the corrected Rand index and Meila's VI

All the above elements can be used to evaluate the internal quality of clustering.

In the following sections, we'll compute the clustering quality statistics for k-means. Look at the **within.cluster.ss** (within clusters sum of squares), the **average.within** (average distance within clusters) and **clus.avg.silwidths** (vector of cluster average silhouette widths).

```
library(fpc)
# Statistics for k-means clustering
km_stats <- cluster.stats(dist(df),  km.res$cluster)
# Dun index
km_stats$dunn
```

```
## [1] 0.0265
```

To display all statistics, type this:

```
km_stats
```

Read the documentation of *cluster.stats*() for details about all the available indices.

External clustering validation

Among the values returned by the function **cluster.stats**(), there are two indexes to assess the similarity of two clustering, namely the corrected Rand index and Meila's VI.

We know that the iris data contains exactly 3 groups of species.

Does the K-means clustering matches with the true structure of the data?

We can use the function **cluster.stats**() to answer to this question.

Let start by computing a cross-tabulation between k-means clusters and the reference Species labels:

```
table(iris$Species, km.res$cluster)
```

```
##
##                1  2  3
##   setosa      50  0  0
##   versicolor   0 11 39
##   virginica    0 36 14
```

It can be seen that:

- All setosa species (n = 50) has been classified in cluster 1
- A large number of versicor species (n = 39) has been classified in cluster 3. Some of them (n = 11) have been classified in cluster 2.
- A large number of virginica species (n = 36) has been classified in cluster 2. Some of them (n = 14) have been classified in cluster 3.

It's possible to quantify the agreement between Species and k-means clusters using either the corrected Rand index and Meila's VI provided as follow:

```
library("fpc")
# Compute cluster stats
species <- as.numeric(iris$Species)
clust_stats <- cluster.stats(d = dist(df),
                             species, km.res$cluster)
# Corrected Rand index
clust_stats$corrected.rand
```

```
## [1] 0.62
```

```
# VI
clust_stats$vi
```

```
## [1] 0.748
```

The corrected **Rand index** provides a measure for assessing the similarity between two partitions, adjusted for chance. Its range is -1 (no agreement) to 1 (perfect agreement). Agreement between the specie types and the cluster solution is 0.62 using **Rand index** and 0.748 using Meila's VI.

The same analysis can be computed for both PAM and hierarchical clustering:

```
# Agreement between species and pam clusters
pam.res <- eclust(df, "pam", k = 3, graph = FALSE)
table(iris$Species, pam.res$cluster)
cluster.stats(d = dist(iris.scaled),
              species, pam.res$cluster)$vi
# Agreement between species and HC clusters
res.hc <- eclust(df, "hclust", k = 3, graph = FALSE)
table(iris$Species, res.hc$cluster)
cluster.stats(d = dist(iris.scaled),
              species, res.hc$cluster)$vi
```

External clustering validation, can be used to select suitable clustering algorithm for a given data set.

Summary

We described how to validate clustering results using the silhouette method and the Dunn index. This task is facilitated using the combination of two R functions: *eclust*() and *fviz_silhouette* in the factoextra package We also demonstrated how to assess the agreement between a clustering result and an external reference. In the next chapters, we'll show how to i) choose the appropriate clustering algorithm for your data; and ii) computing p-values for hierarchical clustering.

References

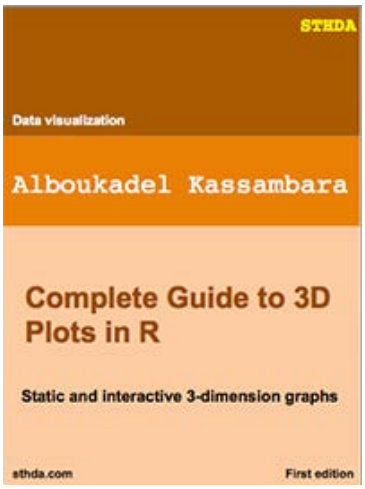
Brock, Guy, Vasyl Pihur, Susmita Datta, and Somnath Datta. 2008. "ClValid: An R Package for Cluster Validation." *Journal of Statistical Software* 25 (4): 1–22. https://www.jstatsoft.org/v025/i04.

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." *Journal of Statistical Software* 61: 1–36. http://www.jstatsoft.org/v61/i06/paper.

Theodoridis, Sergios, and Konstantinos Koutroumbas. 2008. *Pattern Recognition*. 2nd ed. Academic Press.

Last update : 24/09/2017
□ □ □ □ □ 1 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.



Articles - Cluster Validation Essentials

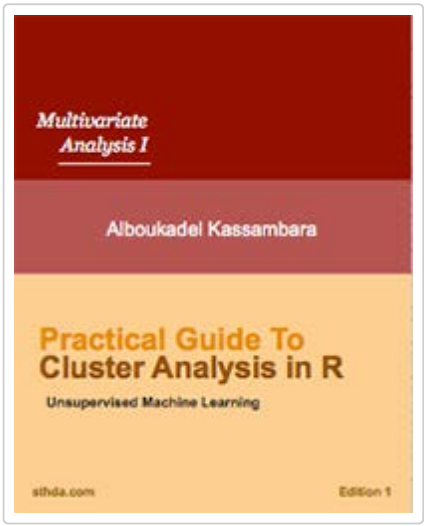
Choosing the Best Clustering Algorithms

kassambara | 07/09/2017 | 5753 | Comment (1) | Cluster Validation Essentials | Unsupervised machine learning

Choosing the best clustering method for a given data can be a hard task for the analyst. This article describes the R package **clValid** (Brock et al. 2008), which can be used to compare simultaneously multiple clustering algorithms in a single function call for identifying the best clustering approach and the optimal number of clusters.

We'll start by describing the different measures in the clValid package for comparing clustering algorithms. Next, we'll present the function *clValid*(). Finally, we'll provide R scripts for validating clustering results and comparing clustering algorithms.

Contents:



Practical Guide to Cluster Analysis in R

Measures for comparing clustering algorithms

The clValid package compares clustering algorithms using two cluster validation measures:

1. *Internal measures*, which uses intrinsic information in the data to assess the quality of the clustering. Internal measures include the connectivity, the silhouette coefficient and the Dunn index as described in Chapter @ref(cluster-validation-statistics) (Cluster Validation Statistics).
2. *Stability measures*, a special version of internal measures, which evaluates the consistency of a clustering result by comparing it with the clusters obtained after each column is removed, one at a time.

Cluster stability measures include:

- The average proportion of non-overlap (APN)
- The average distance (AD)
- The average distance between means (ADM)
- The figure of merit (FOM)

The APN, AD, and ADM are all based on the cross-classification table of the original clustering on the full data with the clustering based on the removal of one column.

- The APN measures the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed.
- The AD measures the average distance between observations placed in the same cluster under both cases (full data set and removal of one column).
- The ADM measures the average distance between cluster centers for observations placed in the same cluster under both cases.
- The FOM measures the average intra-cluster variance of the deleted column, where the clustering is based on the remaining (undeleted) columns.

The values of APN, ADM and FOM ranges from 0 to 1, with smaller value corresponding with highly consistent clustering results. AD has a value between 0 and infinity, and smaller values are also preferred.

Note that, the clValid package provides also biological validation measures, which evaluates the ability of a clustering algorithm to produce biologically meaningful clusters. An application is microarray or RNAseq data where observations corresponds to genes.

Compare clustering algorithms in R

We'll use the function **clValid**() [in the *clValid* package], which simplified format is as follow:

```
clValid(obj, nClust, clMethods = "hierarchical",
        validation = "stability", maxitems = 600,
        metric = "euclidean", method = "average")
```

- **obj**: A numeric matrix or data frame. Rows are the items to be clustered and columns are samples.
- **nClust**: A numeric vector specifying the numbers of clusters to be evaluated. e.g., 2:10
- **clMethods**: The clustering method to be used. Available options are “hierarchical”, “kmeans”, “diana”, “fanny”, “som”, “model”, “sota”, “pam”, “clara”, and “agnes”, with multiple choices allowed.
- **validation**: The type of validation measures to be used. Allowed values are “internal”, “stability”, and “biological”, with multiple choices allowed.
- **maxitems**: The maximum number of items (rows in matrix) which can be clustered.
- **metric**: The metric used to determine the distance matrix. Possible choices are “euclidean”, “correlation”, and “manhattan”.
- **method**: For hierarchical clustering (hclust and agnes), the agglomeration method to be used. Available choices are “ward”, “single”, “complete” and “average”.

For example, consider the iris data set, the *clValid*() function can be used as follow.

We start by cluster internal measures, which include the connectivity, silhouette width and Dunn index. It’s possible to compute simultaneously these internal measures for multiple clustering algorithms in combination with a range of cluster numbers.

```
library(clValid)
# Iris data set:
# - Remove Species column and scale
df <- scale(iris[, -5])
# Compute clValid
clmethods <- c("hierarchical","kmeans","pam")
intern <- clValid(df, nClust = 2:6,
              clMethods = clmethods, validation = "internal")
# Summary
summary(intern)
```

```
##  Length   Class    Mode
##       1 clValid      S4
```

It can be seen that hierarchical clustering with two clusters performs the best in each case (i.e., for connectivity, Dunn and Silhouette measures). Regardless of the clustering algorithm, the optimal number of clusters seems to be two using the three measures.

The stability measures can be computed as follow:

```
# Stability measures
clmethods <- c("hierarchical","kmeans","pam")
stab <- clValid(df, nClust = 2:6, clMethods = clmethods,
                validation = "stability")
# Display only optimal Scores
optimalScores(stab)
```

```
##       Score       Method Clusters
## APN 0.00327 hierarchical        2
```

```
## AD  1.00429          pam          6
## ADM 0.01609 hierarchical          2
## FOM 0.45575          pam          6
```

For the APN and ADM measures, hierarchical clustering with two clusters again gives the best score. For the other measures, PAM with six clusters has the best score.

Summary

Here, we described how to compare clustering algorithms using the *clValid* R package.

References

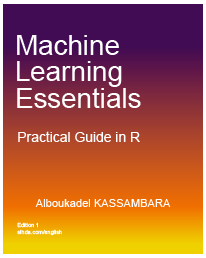
Brock, Guy, Vasyl Pihur, Susmita Datta, and Somnath Datta. 2008. “ClValid: An R Package for Cluster Validation.” *Journal of Statistical Software* 25 (4): 1–22. https://www.jstatsoft.org/v025/i04.

Last update : 24/09/2017
□ □ □ □ □ 1 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I’d be very grateful if you’d help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

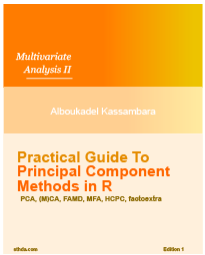
Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!

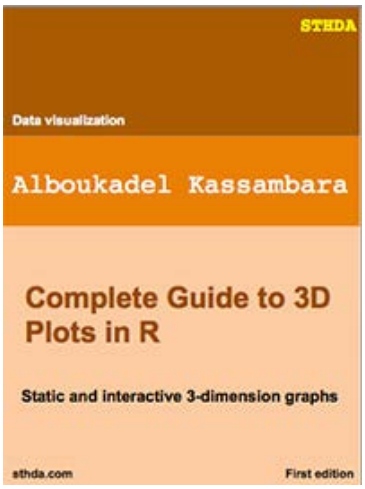


Machine Learning Essentials: Practical Guide in R

Practical Guide to Cluster Analysis in R



Practical Guide to Principal Component Methods in R



Articles - Cluster Validation Essentials

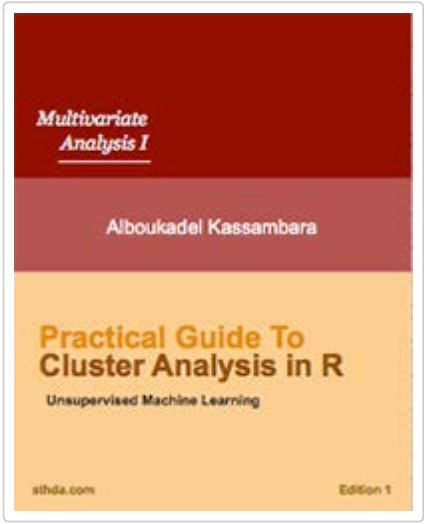
Computing P-value for Hierarchical Clustering

kassambara | 07/09/2017 | 3064 | Post a comment | Cluster Validation Essentials | Unsupervised machine learning

Clusters can be found in a data set by chance due to clustering noise or sampling error. This article describes the R package **pvclust** (Suzuki and Shimodaira 2015) which uses bootstrap resampling techniques to **compute p-value** for each **hierarchical clusters**.

Contents:

Related Book:



Practical Guide to Cluster Analysis in R

Algorithm

1. Generated thousands of bootstrap samples by randomly sampling elements of the data
2. Compute hierarchical clustering on each bootstrap copy
3. For each cluster:
   - compute the *bootstrap probability* (*BP*) value which corresponds to the frequency that the cluster is identified in bootstrap copies.
   - Compute the *approximately unbiased* (AU) probability values (p-values) by multiscale bootstrap resampling

Clusters with AU >= 95% are considered to be strongly supported by data.

Required packages

1. Install **pvclust**:

```
install.packages("pvclust")
```

2. Load **pvclust**:

```
library(pvclust)
```

Data preparation

We'll use *lung* data set [in *pvclust* package]. It contains the gene expression profile of 916 genes of 73 lung tissues including 67 tumors. Columns are samples and rows are genes.

```
library(pvclust)
# Load the data
data("lung")
head(lung[, 1:4])
```

```
##              fetal_lung 232-97_SCC 232-97_node 68-96_Adeno
```

```
## IMAGE:196992       -0.40       4.28        3.68       -1.35
## IMAGE:587847       -2.22       5.21        4.75       -0.91
## IMAGE:1049185      -1.35      -0.84       -2.88        3.35
## IMAGE:135221        0.68       0.56       -0.45       -0.20
## IMAGE:298560          NA       4.14        3.58       -0.40
## IMAGE:119882       -3.23      -2.84       -2.72       -0.83
```

```
# Dimension of the data
dim(lung)
```

```
## [1] 916  73
```

We'll use only a subset of the data set for the clustering analysis. The R function *sample*() can be used to extract a random subset of 30 samples:

```
set.seed(123)
ss <- sample(1:73, 30) # extract 20 samples out of
df <- lung[, ss]
```

Compute p-value for hierarchical clustering

Description of pvclust() function

The function *pvclust*() can be used as follow:

```
pvclust(data, method.hclust = "average",
        method.dist = "correlation", nboot = 1000)
```

Note that, the computation time can be strongly decreased using parallel computation version called *parPvclust*(). (Read ?parPvclust() for more information.)

```
parPvclust(cl=NULL, data, method.hclust = "average",
           method.dist = "correlation", nboot = 1000,
           iseed = NULL)
```

- **data**: numeric data matrix or data frame.
- **method.hclust**: the agglomerative method used in hierarchical clustering. Possible values are one of "average", "ward", "single", "complete", "mcquitty", "median" or "centroid". The default is "average". See method argument in **?hclust**.
- **method.dist**: the distance measure to be used. Possible values are one of "correlation", "uncentered", "abscor" or those which are allowed for **method** argument in **dist()** function, such "euclidean" and "manhattan".
- **nboot**: the number of bootstrap replications. The default is 1000.
- **iseed**: an integrer for random seeds. Use iseed argument to achieve reproducible results.

The function *pvclust*() returns an object of class *pvclust* containing many elements including *hclust* which contains hierarchical clustering result for the original data generated by the function *hclust*().

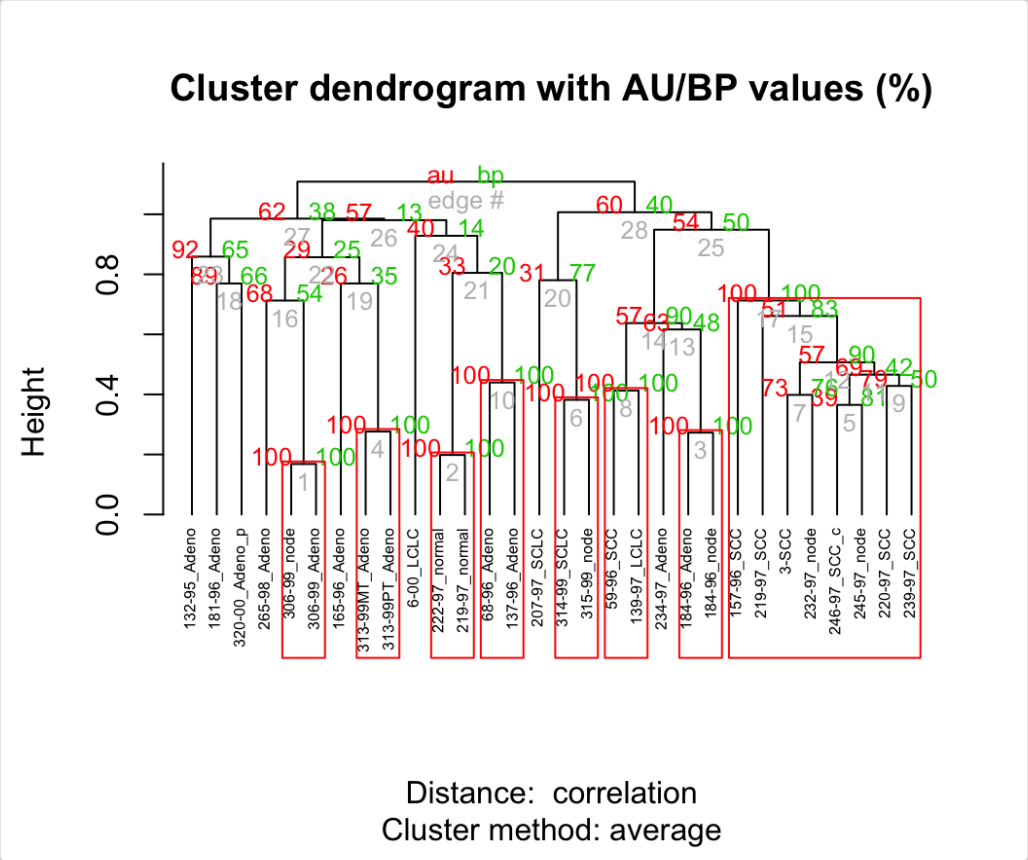
Usage of pvclust() function

pvclust() performs clustering on the columns of the data set, which correspond to samples in our case. If you want to perform the clustering on the variables (here, genes) you have to transpose the data set using the function *t*().

The R code below computes *pvclust*() using 10 as the number of bootstrap replications (for speed):

```
library(pvclust)
set.seed(123)
res.pv <- pvclust(df, method.dist="cor",
                  method.hclust="average", nboot = 10)
```

```
# Default plot
plot(res.pv, hang = -1, cex = 0.5)
pvrect(res.pv)
```



Values on the dendrogram are *AU p-values* (Red, left), *BP values* (green, right), and *clusterlabels* (grey, bottom). Clusters with AU > = 95% are indicated by the rectangles and are considered to be strongly supported by data.

To extract the objects from the significant clusters, use the function *pvpick*():

```
clusters <- pvpick(res.pv)
clusters
```

Parallel computation can be applied as follow:

```
# Create a parallel socket cluster
library(parallel)
cl <- makeCluster(2, type = "PSOCK")
# parallel version of pvclust
res.pv <- parPvclust(cl, df, nboot=1000)
stopCluster(cl)
```

References

Suzuki, Ryota, and Hidetoshi Shimodaira. 2015. *Pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling*. https://CRAN.R-project.org/package=pvclust.

Last update : 24/09/2017
□ □ □ □ □ 0 Note

Enjoyed this article? Give us 5 stars □ □ □□□ (just above this text block)! Reader needs to be STHDA member for voting. I'd be very grateful if you'd help it spread by emailing it to a friend, or sharing it on Twitter, Facebook or Linked In.

Show me some love with the like buttons below... Thank you and please don't forget to share and comment below!!

Recommended for You!