

*Multivariate
Analysis I*

Alboukadel Kassambara

**Practical Guide To
Cluster Analysis in R**

Unsupervised Machine Learning

Copyright ©2017 by Alboukadel Kassambara. All rights reserved.

Published by STHDA (<http://www.sthda.com>), Alboukadel Kassambara

Contact: Alboukadel Kassambara <alboukadel.kassambara@gmail.com>

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to STHDA (<http://www.sthda.com>).

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials.

Neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

For general information contact Alboukadel Kassambara <alboukadel.kassambara@gmail.com>.

0.1 Preface

Large amounts of data are collected every day from satellite images, bio-medical, security, marketing, web search, geo-spatial or other automatic equipment. Mining knowledge from these big data far exceeds human's abilities.

Clustering is one of the important data mining methods for discovering knowledge in multidimensional data. The goal of clustering is to identify pattern or groups of similar objects within a data set of interest.

In the literature, it is referred as “pattern recognition” or “unsupervised machine learning” - “unsupervised” because we are not guided by a priori ideas of which variables or samples belong in which clusters. “Learning” because the machine algorithm “learns” how to cluster.

Cluster analysis is popular in many fields, including:

- In *cancer research* for classifying patients into subgroups according their gene expression profile. This can be useful for identifying the molecular profile of patients with good or bad prognostic, as well as for understanding the disease.
- In *marketing* for *market segmentation* by identifying subgroups of customers with similar profiles and who might be receptive to a particular form of advertising.
- In *City-planning* for identifying groups of houses according to their type, value and location.

This book provides a practical guide to unsupervised machine learning or cluster analysis using R software. Additionally, we developed an R package named *factoextra* to create, easily, a ggplot2-based elegant plots of cluster analysis results. Factoextra official online documentation: <http://www.sthda.com/english/rpkgs/factoextra>

0.2 About the author

Alboukadel Kassambara is a PhD in Bioinformatics and Cancer Biology. He works since many years on genomic data analysis and visualization. He created a bioinformatics tool named GenomicScape (www.genomicscape.com) which is an easy-to-use web tool for gene expression data analysis and visualization.

He developed also a website called STHDA (Statistical Tools for High-throughput Data Analysis, www.sthda.com/english), which contains many tutorials on data analysis and visualization using R software and packages.

He is the author of the R packages **survminer** (for analyzing and drawing survival curves), **ggcorrplot** (for drawing correlation matrix using ggplot2) and **factoextra** (to easily extract and visualize the results of multivariate analysis such PCA, CA, MCA and clustering). You can learn more about these packages at: <http://www.sthda.com/english/wiki/r-packages>

Recently, he published two books on data visualization:

1. Guide to Create Beautiful Graphics in R (at: <https://goo.gl/vJ0OYb>).
2. Complete Guide to 3D Plots in R (at: <https://goo.gl/v5gw10>).

Contents

0.1	Preface	3
0.2	About the author	4
0.3	Key features of this book	9
0.4	How this book is organized?	10
0.5	Book website	16
0.6	Executing the R codes from the PDF	16
I	Basics	17
1	Introduction to R	18
1.1	Install R and RStudio	18
1.2	Installing and loading R packages	19
1.3	Getting help with functions in R	20
1.4	Importing your data into R	20
1.5	Demo data sets	22
1.6	Close your R/RStudio session	22
2	Data Preparation and R Packages	23
2.1	Data preparation	23
2.2	Required R Packages	24
3	Clustering Distance Measures	25
3.1	Methods for measuring distances	25
3.2	What type of distance measures should we choose?	27
3.3	Data standardization	28
3.4	Distance matrix computation	29
3.5	Visualizing distance matrices	32
3.6	Summary	33

II	Partitioning Clustering	34
4	K-Means Clustering	36
4.1	K-means basic ideas	36
4.2	K-means algorithm	37
4.3	Computing k-means clustering in R	38
4.4	K-means clustering advantages and disadvantages	46
4.5	Alternative to k-means clustering	47
4.6	Summary	47
5	K-Medoids	48
5.1	PAM concept	49
5.2	PAM algorithm	49
5.3	Computing PAM in R	50
5.4	Summary	56
6	CLARA - Clustering Large Applications	57
6.1	CLARA concept	57
6.2	CLARA Algorithm	58
6.3	Computing CLARA in R	58
6.4	Summary	63
III	Hierarchical Clustering	64
7	Agglomerative Clustering	67
7.1	Algorithm	67
7.2	Steps to agglomerative hierarchical clustering	68
7.3	Verify the cluster tree	73
7.4	Cut the dendrogram into different groups	74
7.5	Cluster R package	77
7.6	Application of hierarchical clustering to gene expression data analysis	77
7.7	Summary	78
8	Comparing Dendrograms	79
8.1	Data preparation	79
8.2	Comparing dendrograms	80
9	Visualizing Dendrograms	84
9.1	Visualizing dendrograms	85
9.2	Case of dendrogram with large data sets	90

<i>CONTENTS</i>	7
9.3 Manipulating dendrograms using dendextend	94
9.4 Summary	96
10 Heatmap: Static and Interactive	97
10.1 R Packages/functions for drawing heatmaps	97
10.2 Data preparation	98
10.3 R base heatmap: heatmap()	98
10.4 Enhanced heat maps: heatmap.2()	101
10.5 Pretty heat maps: pheatmap()	102
10.6 Interactive heat maps: d3heatmap()	103
10.7 Enhancing heatmaps using dendextend	103
10.8 Complex heatmap	104
10.9 Application to gene expression matrix	114
10.10 Summary	116
IV Cluster Validation	117
11 Assessing Clustering Tendency	119
11.1 Required R packages	119
11.2 Data preparation	120
11.3 Visual inspection of the data	120
11.4 Why assessing clustering tendency?	121
11.5 Methods for assessing clustering tendency	123
11.6 Summary	127
12 Determining the Optimal Number of Clusters	128
12.1 Elbow method	129
12.2 Average silhouette method	130
12.3 Gap statistic method	130
12.4 Computing the number of clusters using R	131
12.5 Summary	137
13 Cluster Validation Statistics	138
13.1 Internal measures for cluster validation	139
13.2 External measures for clustering validation	141
13.3 Computing cluster validation statistics in R	142
13.4 Summary	150
14 Choosing the Best Clustering Algorithms	151
14.1 Measures for comparing clustering algorithms	151

14.2	Compare clustering algorithms in R	152
14.3	Summary	155
15	Computing P-value for Hierarchical Clustering	156
15.1	Algorithm	156
15.2	Required packages	157
15.3	Data preparation	157
15.4	Compute p-value for hierarchical clustering	158
V	Advanced Clustering	161
16	Hierarchical K-Means Clustering	163
16.1	Algorithm	163
16.2	R code	164
16.3	Summary	166
17	Fuzzy Clustering	167
17.1	Required R packages	167
17.2	Computing fuzzy clustering	168
17.3	Summary	170
18	Model-Based Clustering	171
18.1	Concept of model-based clustering	171
18.2	Estimating model parameters	173
18.3	Choosing the best model	173
18.4	Computing model-based clustering in R	173
18.5	Visualizing model-based clustering	175
19	DBSCAN: Density-Based Clustering	177
19.1	Why DBSCAN?	178
19.2	Algorithm	180
19.3	Advantages	181
19.4	Parameter estimation	182
19.5	Computing DBSCAN	182
19.6	Method for determining the optimal eps value	184
19.7	Cluster predictions with DBSCAN algorithm	185
20	References and Further Reading	186

0.3 Key features of this book

Although there are several good books on unsupervised machine learning/clustering and related topics, we felt that many of them are either too high-level, theoretical or too advanced. Our goal was to write a practical guide to cluster analysis, elegant visualization and interpretation.

The main parts of the book include:

- *distance measures*,
- *partitioning clustering*,
- *hierarchical clustering*,
- *cluster validation methods*, as well as,
- *advanced clustering methods* such as fuzzy clustering, density-based clustering and model-based clustering.

The book presents the basic principles of these tasks and provide many examples in R. This book offers solid guidance in data mining for students and researchers.

Key features:

- Covers clustering algorithm and implementation
- Key mathematical concepts are presented
- Short, self-contained chapters with practical examples. This means that, you don't need to read the different chapters in sequence.

At the end of each chapter, we present R lab sections in which we systematically work through applications of the various methods discussed in that chapter.

0.4 How this book is organized?

01	02	03	04	05
Basics	Partitioning Clustering	Hierarchical Clustering	Cluster Validation	Advanced Clustering
<ul style="list-style-type: none"> + <i>Introduction to R</i> + <i>Data Preparation</i> + <i>Required R Packages</i> + <i>Distance Measures</i> 	<ul style="list-style-type: none"> + <i>K-Means</i> + <i>K-Medoids (PAM)</i> + <i>CLARA</i> 	<ul style="list-style-type: none"> + <i>Agglomerative Clustering</i> + <i>Comparing Dendrograms</i> + <i>Visualizing Dendrograms</i> + <i>Heatmap: Static & Interactive</i> 	<ul style="list-style-type: none"> + <i>Clustering Tendency</i> + <i>Optimal Number of Clusters</i> + <i>Validation Statistics</i> + <i>P-value for Hierarchical Clustering</i> 	<ul style="list-style-type: none"> + <i>Hybrid Methods</i> + <i>Fuzzy Clustering</i> + <i>Model-Based Clustering</i> + <i>Density-Based Clustering</i>

This book contains 5 parts. Part I (Chapter 1 - 3) provides a quick introduction to R (chapter 1) and presents required R packages and data format (Chapter 2) for clustering analysis and visualization.

The classification of objects, into clusters, requires some methods for measuring the distance or the (dis)similarity between the objects. Chapter 3 covers the common distance measures used for assessing similarity between observations.

Part II starts with partitioning clustering methods, which include:

- K-means clustering (Chapter 4),
- K-Medoids or PAM (partitioning around medoids) algorithm (Chapter 5) and
- CLARA algorithms (Chapter 6).

Partitioning clustering approaches subdivide the data sets into a set of k groups, where k is the number of groups pre-specified by the analyst.