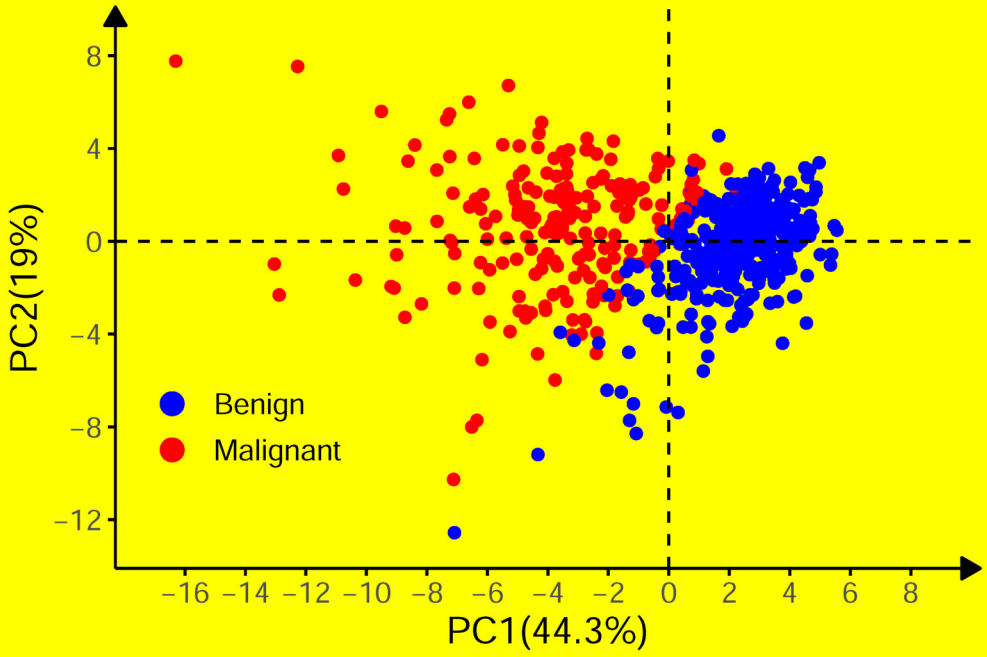
Introductory Principal Component Analysis Using R

A Practical Guide with RStudio



Kilem L. Gwet, Ph.D.

INTRODUCTORY PRINCIPAL COMPONENT ANALYSIS USING R

A Practical Guide with RStudio

Introductory Principal Component Analysis Using R

A Practical Guide with RStudio

Kilem L. Gwet, Ph.D.

AgreeStat Analytics
Gaithersburg, MD 20886-2696, USA

Published by AgreeStat Analytics; in the United States of America.

AgreeStat Analytics
20416 Davencroft Ct,
Gaithersburg, MD 20886-4379
e-mail: gwet@agreestat.com

Publisher's Cataloguing in Publication Data:

Gwet, Kilem Li
Introductory Principal Component Analysis Using R
A Practical Guide with RStudio / By Kilem Li Gwet
p. cm.
Includes bibliographical references and index.
1. Statistics
2. Statistical Analysis
3. Data Analysis
4. Statistics - Study - Learning. I. Title.

Preface

Principal component analysis (PCA) will appeal to you if you have collected a large number of measurements on each of many subjects and find it difficult to extract useful information out of your dataset. Measurements are often associated with correlated variables, making it difficult to evaluate the impact of individual variables on subjects. PCA is a statistical technique that transforms the original correlated variables into a new set of *synthetic variables* also called *score variables* or *PC score variables* where PC stands for Principal Component. These variables have 2 interesting properties: *(i) they are uncorrelated*, and *(ii) they are ranked in such a way that the first few of them can explain most of the variation contained in the original dataset*.

You can see some key advantages of using only a few PC score variables as surrogates for the entire set of original variables:

- Since the first few PC score variables can explain most of the variation in the original dataset, your analysis can be based on one or two of them. The primary objective of PCA is dimensionality reduction. The 2 dominant PC score variables can give you an intuitive two-dimensional snapshot of a complex dataset.

- PC score variables are uncorrelated. Therefore, each can be interpreted independently from the others leading to a more stable statistical analysis.

The literature on principal component analysis (PCA) is abundant. This statistical technique is believed to have been introduced by Pearson (1901) before being rediscovered by Hotelling (1933). While many computer scientists only discover this technique now in the era of machine learning, it is actually a very old statistical technique that has been used by statisticians for more than a century. Why this book?

This book can be used as supplement to many existing introductory books on PCA, which only cover specific aspects of this technique. For example, the book written by Dunteman (1989) provides a good balance between a mathematical treatment of PCA and a gentle introduction to key concepts. Some readers may still find it too mathematical, whereas others may find it too dated

to incorporate modern software solutions, which are essential these days for implementing PCA. The more advanced books such as that of Jolliffe (2002), tend to cover too much ground and to overemphasize the mathematical aspects of PCA.

The main objective of this book is two-fold:

→ First, this book will show you how to perform PCA using the R software. Although this software comes with a documentation that shows you how to perform PCA, R outputs are not always clearly explained. Moreover, documentation files always assume that you have the appropriate background in PCA. However, the background information you obtain from textbooks will still not explain R outputs. This book will close that gap.

→ Secondly, this book will provide the reader with a non-mathematical and yet rigorous presentation of the nature of Principal Component Analysis. The use of numerical examples is extensive and is the main approach adopted for introducing new concepts. You will learn as you do, or as you see me do.

After you complete the reading this book, I expect you to have an in-depth understanding of what principal components are, how to compute them using the R software, and how you can use them to improve the performance of other statistical techniques.

About the Author

I received my PhD in Mathematics from Carleton University's School of Mathematics and Statistics, Ottawa (Canada) in 1997. My specialization was the design and analysis of statistical surveys. In the past few years however, I wrote books and papers in the field of inter-rater reliability analysis (for some of my works in this field see `https://www.researchgate.net/profile/Kilem_Gwet`). After applying principal components to analyze multivariate inter-rater reliability data, I decided to write my first book on PCA. That book is more mathematical and uses MS Excel to implement the methods (see Gwet, 2020).

If you have comments or questions, please contact me at gwet@agreestat.com. I will respond to your inquiry as early as I possibly can.

Kilem Li Gwet, Ph.D.

Contents

Acknowledgment ix

1 Basics of Principal Components 1
- 1.1 *Introduction* . . . 2
- 1.2 *A Glimpse into Principal Component Analysis* . . . 3
- 1.3 *Principal Components and Associated Scores* . . . 6
- 1.4 *A More Realistic PCA Problem* . . . 8
  - 1.4.1 *The Principal Component Scores* . . . 9
  - 1.4.2 *Visualizing Key Principal Component Scores* . . . 11
  - 1.4.3 *Interpreting PC Loadings* . . . 13
- 1.5 *Datasets* . . . 14
- 1.6 *Concluding Remarks* . . . 17

2 Overview of R and RStudio 19
- 2.1 *Introduction* . . . 20
- 2.2 *RStudio Cloud* . . . 20
- 2.3 *RStudio Desktop* . . . 25
- 2.4 *Using RStudio* . . . 28
  - 2.4.1 *R Commands in RStudio* . . . 28
  - 2.4.2 *RStudio Workflow* . . . 29
- 2.5 *Solving Problems with R* . . . 32
  - 2.5.1 *R Objects* . . . 34
  - 2.5.2 *Built-In Functions* . . . 49
- 2.6 *Concluding Remarks* . . . 53

3 Computing Principal Components with R 55
- 3.1 *Introduction* . . . 56
- 3.2 *Calculating the Principal Components* . . . 57
  - 3.2.1 *Problem: Analysis of the MTcars Dataset* . . . 58

3.2.2 *Exploring the PCA Results* . . . 63
3.3 *Comparing prcomp() and princomp() Functions* . . . 68
3.4 *Importance of Data Centering & Standardization* . . . 73
3.5 *Concluding Remarks* . . . 80

4 Visualization of Principal Components **83**
4.1 *Introduction* . . . 84
4.2 *Useful Packages* . . . 87
4.3 *Scree & Cumulative Variance Plots* . . . 89
4.4 *PCA Score Plot* . . . 98
4.5 *Loadings and Variable Contribution Plots* . . . 103
4.6 *Concluding Remarks* . . . 112

5 Basic Use of Principal Components **115**
5.1 *Introduction* . . . 116
5.2 *Optimal Number of Principal Components* . . . 116
5.3 *Predicting Scores for New Data* . . . 122
5.4 *Reconstructing Original Data from Limited PCs* . . . 129
5.4.1 *Anomaly Detection Using Principal Components* . . . 129
5.4.2 *Data Reconstruction in Matrix Notations* . . . 138
5.5 *Concluding Remarks* . . . 140

6 Statistical Analysis Based on Principal Components **143**
6.1 *Introduction* . . . 144
6.2 *Identifying Relevant Variables with Principal Components* . . . 145
6.3 *Identifying Outliers with Principal Components* . . . 151
6.4 *Principal Component Regression (PCR)* . . . 159
6.5 *Improving Classification with PCA* . . . 188
6.6 *Concluding Remarks* . . . 197

A Performing Linear Discriminant Analysis in R **199**

Bibliography **207**

Author Index **209**

Subject Index **210**

Acknowledgment

I want to express gratitude to my family for their support while working on this book. Most of all, I thank my wife Suzy and our three girls Mata, Lelna, and Addia. They have all graciously put up with my insatiable computer habits, my long workdays, and busy weekends over the past few years.

Kilem Li Gwet, Ph.D.
Maryland, USA: November 2023