BEGINNER'S GUIDE TO

PRINCIPAL COMPONENT ANALYSIS

Applications with Excel

Beginner's Guide to Principal Component Analysis

Applications with Excel

Kilem L. Gwet, Ph.D.

AgreeStat Analytics
P.O. Box 2696
Gaithersburg, MD 20886-2696, USA

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

AgreeStat Analytics
PO BOX 2696,
Gaithersburg, MD 20886-2696
e-mail: gwet@agreestat.com

This publication is designed to provide accurate and authoritative information in regard of the subject matter covered. However, it is sold with the understanding that the publisher assumes no responsibility for errors, inaccuracies or omissions. The publisher is not engaged in rendering any professional services. A competent professional person should be sought for expert assistance.

Publisher's Cataloguing in Publication Data:

Gwet, Kilem Li
Beginner's Guide to Principal Component Analysis
Applications with Excel / 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 variables called the principal components, which have 2 interesting properties: (i) they are uncorrelated, and (ii) they are ordered 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 principal components as surrogates for the original variables:

- Since the first few principal components can explain most of the variation in the original dataset, your analysis can be based on one or two principal components instead of being based on 20 or 30 correlated variables. Therefore, the primary objective of the principal components is dimensionality reduction. It allows for a clearer view of many aspects of a complex dataset built with inter-related variables. The 2 dominant principal components can give you an instructive two-dimensional snapshot of a complex dataset.

- Principal components 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). As a result, the PCA was previously referred to as the Hotelling transformation. 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 did I then decide to add another book to this extensive literature?

While many introductory books on PCA - Dunteman (1989) for example - are good at providing a high-level overview of the concepts and their possible applications, they often do not provide a sufficiently detailed account of how principal components are calculated. The more advanced books such as that of Jolliffe (2002), tend to cover too much ground and to overemphasize the mathematical aspects of the methods. Every technical detail appears to be important and limited effort is devoted to showing what makes this technique work. Moreover, the computational aspects of principal components, which often involve advanced numerical methods are often neglected. I decided to write this book to provide a good balance between the need to describe the computational procedures and that of presenting an adequate mathematical treatment of the methods. The coverage of the different methods for computing principal components is not comprehensive. Instead, I have focused on some of the most widely-used methods.

Even though using an existing PCA algorithm is trivial, the mathematics of principal components can be involved. It is likely the primary cause of them being omitted in many publications. The more you want to understand what is taking place behind the scene, the deeper you need to dig. My goal in this book, is to make the study of principal components as painless as possible.

Mathematical concepts such as the eigenvalue or the eigenvector, previously taught in linear algebra courses only, must now be understood by most researchers and analysts given their growing importance in new fields. The eigenvalues and eigenvectors are widely used today in the field of artificial intelligence in addition to having found numerous applications in the past few decades in the social sciences. Understanding what they are and how to compute them is essential for understanding how large bodies of data are analyzed today.

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, and what makes them work. For those of you who use Microsoft Excel, an Excel template named `pca.xlsm` can be downloaded for free at the address `https://agreestat.com/books/pca/pca.xlsm`.

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 devoted considerable time and energy to the field of inter-rater reliability analysis, in which I published several papers and books (for some of my works in this field see https://www.researchgate.net/profile/Kilem_Gwet). It is after successfully using principal components to analyze multivariate inter-rater reliability data that I first considered writing this book. I wanted a book that shows step by step how principal components are calculated, what they represent, how they must be used and what their limitations are. These are some of the reasons why I wrote this book, which I hope you will find useful.

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

I want to express my 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 and so many long workdays, and busy weekends over the past few years.

Kilem Li Gwet, Ph.D.
Maryland, USA: November 2020