A User's Guide To Principal Components

This Page Intentionally Left Blank

A User's Guide To Principal Components

This Page Intentionally Left Blank

A User's Guide To Principal Components

J. EDWARD JACKSON

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Brisbane • Toronto • Singapore

A NOTE TO THE READER
This book has been electronically reproduced from digital information stored at John Wiley & Sons, Inc. We are pleased that the use of this new technology will enable us to keep works of enduring scholarly value in print as long as there is a reasonable demand for them. The content of this book is identical to previous printings.

In recognition of the importance of preserving what has been written, it is a policy of John Wiley & Sons, Inc., to have books of enduring value published in the United States printed on acid-free paper, and we exert our best efforts to that end.

To order books or for customer service please, call 1(800)-CALL-WILEY (225-5945).

Library of Congress Cataloging in Publication Data:

Jackson, J. Edward.
A user's guide to principal components / J. Edward Jackson.
p. cm. — (Wiley series in probability and mathematical statistics. Applied probability and statistics)
Includes bibliographical references and index.
1. Principal components analysis. I. Title. II. Series.
QA278.5.J27 1991
519.5'354—dc20
ISBN 0-471-62267-2 90-28108
CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2

To my wife,
Suzanne

This Page Intentionally Left Blank

Contents

This Page Intentionally Left Blank

Preface

Principal Component Analysis (PCA) is a multivariate technique in which a number of related variables are transformed to (hopefully, a smaller) set of uncorrelated variables. This book is designed for practitioners of PCA. It is, primarily, a "how-to-do-it" and secondarily a "why-it-works" book. The theoretical aspects of this technique have been adequately dealt with elsewhere and it will suffice to refer to these works where relevant. Similarly, this book will not become overinvolved in computational techniques. These techniques have also been dealt with adequately elsewhere. The user is focusing, primarily, on data reduction and interpretation. Lest one considers the computational aspects of PCA to be a "black box," enough detail will be included in one of the appendices to leave the user with the feeling of being in control of his or her own destiny.

The method of principal components dates back to Karl Pearson in 1901, although the general procedure as we know it today had to wait for Harold Hotelling whose pioneering paper appeared in 1933. The development of the technique has been rather uneven in the ensuing years. There was a great deal of activity in the late 1930s and early 1940s. Things then subsided for a while until computers had been designed that made it possible to apply these techniques to reasonably sized problems. That done, the development activities surged ahead once more. However, this activity has been rather fragmented and it is the purpose of this book to draw all of this information together into a usable guide for practitioners of multivariate data analysis. This book is also designed to be a sourcebook for principal components. Many times a specific technique may be described in detail with references being given to alternate or competing methods. Space considerations preclude describing them all and, in this way, those wishing to investigate a procedure in more detail will know where to find more information. Occasionally, a topic may be presented in what may seem to be less than favorable light. It will be included because it relates to a procedure which is widely used—for better or for worse. In these instances, it would seem better to include the topic with a discussion of the relative pros and cons rather than to ignore it completely.

As PCA forms only one part of multivariate analysis, there are probably few college courses devoted exclusively to this topic. However, if someone did teach a course about PCA, this book could be used because of the detailed development of methodology as well as the many numerical examples. Except for universities with large statistics departments, this book might more likely find use as a supplementary text for multivariate courses. It may also be useful for departments of education, psychology, and business because of the supplementary material dealing with multidimensional scaling and factor analysis. There are no class problems included. Class problems generally consist of either theoretical proofs and identities, which is not a concern of this book, or problems involving data analysis. In the latter case, the instructor would be better off using data sets of his or her own choosing because it would facilitate interpretation and discussion of the problem.

This book had its genesis at the 1973 Fall Technical Conference in Milwaukee, a conference jointly sponsored by the Physical and Engineering Sciences Section of the American Statistical Association and the Chemistry Division of the American Society for Quality Control. That year the program committee wanted two tutorial sessions, one on principal components and the other on factor analysis. When approached to do one of these sessions, I agreed to do either one depending on who else they obtained. Apparently, they ran out of luck at that point because I ended up doing both of them. The end result was a series of papers published in the *Journal of Quality Technology* (Jackson, 1980, 1981a,b). A few years later, my employer offered an early retirement. When I mentioned to Fred Leone that I was considering taking it, he said, "Retire? What are you going to do, write a book?" I ended up not taking it but from that point on, writing a book seemed like a natural thing to do and the topic was obvious.

When I began my career with the Eastman Kodak Company in the late 1940s, most practitioners of multivariate techniques had the dual problem of performing the analysis on the limited computational facilities available at that time and of persuading their clients that multivariate techniques should be given any consideration at all. At Kodak, we were not immune to the first problem but we did have a more sympathetic audience with regard to the second, much of this due to some pioneering efforts on the part of Bob Morris, a chemist with great natural ability in both mathematics and statistics. It was my pleasure to have collaborated with Bob in some of the early development of operational techniques for principal components. Another chemist, Grant Wernimont, and I had adjoining offices when he was advocating the use of principal components in analytical chemistry in the late 1950s and I appreciated his enthusiasm and steady stream of operational "one-liners." Terry Hearne and I worked together for nearly 15 years and collaborated on a number of projects that involved the use of PCA. Often these assignments required some special procedures that called for some ingenuity on our part; Chapter 9 is a typical example of our collaboration.

A large number of people have given me encouragement and assistance in the preparation of this book. In particular, I wish to thank Eastman Kodak's Multivariate Development Committee, including Nancy Farden, Chuck Heckler, Maggie Krier, and John Huber, for their critical appraisal of much of the material in this book as well as some mainframe computational support for some of the multidimensional scaling and factor analysis procedures. Other people from Kodak who performed similar favors include Terry Hearne, Peter Franchuk, Peter Castro, Bill Novik, and John Twist. The format for Chapter 12 was largely the result of some suggestions by Gary Brauer. I received encouragement and assistance with some of the inferential aspects from Govind Mudholkar of the University of Rochester. One of the reviewers provided a number of helpful comments. Any errors that remain are my responsibility.

I also wish to acknowledge the support of my family. My wife Suzanne and my daughter Janice helped me with proofreading. (Our other daughter, Judy, managed to escape by living in Indiana.) My son, Jim, advised me on some of the finer aspects of computing and provided the book from which Table 10.7 was obtained (Leffingwell was a distant cousin.)

I wish to thank the authors, editors, and owners of copyright for permission to reproduce the following figures and tables: Figure 2.4 (Academic Press); Figures 1.1, 1.4, 1.5, 1.6, and 6.1 (American Society for Quality Control and Marcel Dekker); Figure 8.1 and Table 5.9 (American Society for Quality Control); Figures 6.3, 6.4, 6.5, and Table 7.4 (American Statistical Association); Figures 9.1, 9.2, 9.3, and 9.4 (Biometrie-Praximetrie); Figures 18.1 and 18.2 (Marcel Dekker); Figure 11.7 (Psychometrika and D. A. Klahr); Table 8.1 (University of Chicago Press); Table 12.1 (SAS Institute); Appendix G.1 (John Wiley and Sons, Inc.); Appendix G.2 (Biometrika Trustees, the Longman Group Ltd, the Literary Executor of the late Sir Ronald A. Fisher, F.R.S. and Dr. Frank Yates, F.R.S.); Appendices G.3, G.4, and G.6 (Biometrika Trustees); and Appendix G.5 (John Wiley and Sons, Inc., Biometrika Trustees and Marcel Dekker).

Rochester, New York J. EDWARD JACKSON
January 1991

This Page Intentionally Left Blank

A User's Guide To Principal Components

This Page Intentionally Left Blank

Introduction

The method of *principal components* is primarily a data-analytic technique that obtains linear transformations of a group of correlated variables such that certain optimal conditions are achieved. The most important of these conditions is that the transformed variables are uncorrelated. It will be the purpose of this book to show why this technique is useful in statistical analysis and how it is carried out.

The first three chapters establish the properties and mechanics of principal component analysis (PCA). Chapter 4 considers the various inferential techniques required to conduct PCA and all of this is then put to work in Chapter 5, an example dealing with audiometric testing.

The next three chapters deal with grouped data and with various methods of interpreting the principal components. These tools are then employed in a case history, also dealing with audiometric examinations.

Multidimensional scaling is closely related to PCA, some techniques being common to both. Chapter 10 considers these with relation to *preference*, or *dominance*, scaling and, in so doing, introduces the concept of *singular value decomposition*. Chapter 11 deals with *similarity* scaling.

The application of PCA to linear models is examined in the next two chapters. Chapter 12 considers, primarily, the relationships among the predictor variables and introduces principal component regression along with some competitors. Principal component ANOVA is considered in Chapter 13.

Chapter 14 discusses a number of other applications of PCA, including missing data, data editing, tests for multivariate normality, discriminant and cluster analysis, and time series analysis. There are enough special procedures for the two-dimensional case that it merits Chapter 15 all to itself. Chapter 16 is a "catch-all" that contains a number of extensions of PCA including cross-validation, procedures for two or more samples, and robust estimation.

The reader will notice that several chapters deal with subgrouped data or situations dealing with two or more populations. Rather than devote a separate chapter to this, it seemed better to include these techniques where relevant. Chapter 6 considers the situation where data are subgrouped as one might find in quality control operations. The application of PCA in the analysis of variance is taken up in Chapter 13 where, again, the data may be divided into groups. In both of these chapters, the underlying assumption for these operations is that the variability is homogeneous among groups, as is customary in most ANOVA operations. To the extent that this is not the case, other procedures are called for. In Section 16.6, we will deal with the problem of testing whether or not the characteristic roots and vectors representing two or more populations are, in fact, the same. A similar problem is considered in a case study in Chapter 9 where some ad hoc techniques will be used to functionally relate these quantities to the various populations for which data are available.

There are some competitors for principal component analysis and these are discussed in the last two chapters. The most important of these competitors is *factor analysis*, which is sometimes confused with PCA. Factor analysis will be presented in Chapter 17, which will also contain a comparison of the two methods and a discussion about the confusion existing between them. A number of other techniques that may relevant for particular situations will be given in Chapter 18.

A basic knowledge of matrix algebra is essential for the understanding of this book. The operations commonly employed are given in Appendix A and a brief discussion of computing methods is found in Appendix C. You will find very few theorems in this book and only one proof. Most theorems will appear as statements presented where relevant. It seemed worthwhile, however, to list a number of basic properties of PCA in one place and this will be found in Appendix B. Appendix D deals with symbols and terminology—there being no standards for either in PCA. Appendix E describes a few classic data sets, located elsewhere, that one might wish to use in experimenting with some of the techniques described in this book. For the most part, the original sources contain the raw data. Appendix F summarizes all of the data sets employed in this book and the uses to which they were put. Appendix G contains tables related to the following distributions: normal, t, chi-square, F, the Lawley–Hotelling trace statistic and the extreme characteristic roots of a covariance matrix.

While the bibliography is quite extensive, it is by no means complete. Most of the citations relate to methodology and operations since that is the primary emphasis of the book. References pertaining to the theoretical aspects of PCA form a very small minority. As will be pointed out in Chapter 4, considerable effort has been expended elsewhere on studying the distributions associated with characteristic roots. We shall be content to summarize the results of this work and give some general references to which those interested may turn for more details. A similar policy holds with regard to computational techniques. The references dealing with applications are but a small sample of the large number of uses to which PCA has been put.

This book will follow the general custom of using Greek letters to denote population parameters and Latin letters for their sample estimates. Principal component analysis is employed, for the most part, as an exploratory data analysis technique, so that applications involve sample data sets and sample estimates obtained from them. Most of the presentation in this book will be within that context and for that reason population parameters will appear primarily in connection with inferential techniques, in particular in Chapter 4. It is comforting to know that the general PCA methodology is the same for populations as for samples.

Fortunately, many of the operations associated with PCA estimation are distribution free. When inferential procedures are employed, we shall generally assume that the population or populations from which the data were obtained have multivariate normal distributions. The problems associated with non-normality will be discussed where relevant.

Widespread development and application of PCA techniques had to wait for the advent of the high-speed electronic computer and hence one usually thinks of PCA and other multivariate techniques in this vein. It is worth pointing out, however, that with the exception of a few examples where specific mainframe programs were used, the computations in this book were all performed on a 128K microcomputer. No one should be intimidated by PCA computations.

Many statistical computer packages contain a PCA procedure. However, these procedures, in general, cover some, but not all, of the first three chapters, in addition to some parts of Chapters 8 and 17 and in some cases parts of Chapters 10, 11, and 12. For the remaining techniques, the user will have to provide his or her own software. Generally, these techniques are relatively easy to program and one of the reasons for the many examples is to provide the reader some sample data with which to work. Do not be surprised if your answers do not agree to the last digit with those in the book. In addition to the usual problems of computational accuracy, the number of digits has often been reduced in presentation, either in this book or the original sources, to two or three digits for reason of space of clarity. If these results are then used in other computations, an additional amount of precision may be lost. The signs for the characteristic vectors may be reversed from the ones you obtain. This is either because of the algorithm employed or because someone reversed the signs deliberately for presentation. The interpretation will be the same either way.

CHAPTER 1

Getting Started

1.1 INTRODUCTION

The field of *multivariate analysis* consists of those statistical techniques that consider two or more related random variables as a single entity and attempts to produce an overall result taking the relationship among the variables into account. A simple example of this is the correlation coefficient. Most inferential multivariate techniques are generalizations of classical univariate procedures. Corresponding to the univariate t-test is the multivariate T^2-test and there are multivariate analogs of such techniques as regression and the analysis of variance. The majority of most multivariate texts are devoted to such techniques and the multivariate distributions that support them.

There is, however, another class of techniques that is unique to the multivariate arena. The correlation coefficient is a case in point. Although these techniques may also be employed in statistical inference, the majority of their applications are as data-analytic techniques, in particular, techniques that seek to describe the multivariate *structure* of the data. *Principal Component Analysis* or PCA, the topic of this book, is just such a technique and while its main use is as a descriptive technique, we shall see that it may also be used in many inferential procedures as well.

In this chapter, the method of principal components will be illustrated by means of a small hypothetical two-variable example, allowing us to introduce the mechanics of PCA. In subsequent chapters, the method will be extended to the general case of p variables, some larger examples will be introduced, and we shall see where PCA fits into the realm of multivariate analysis.

1.2 A HYPOTHETICAL EXAMPLE

Suppose, for instance, one had a process in which a quality control test for the concentration of a chemical component in a solution was carried out by two different methods. It may be that one of the methods, say Method 1, was the standard procedure and that Method 2 was a proposed alternative, a procedure that was used as a back-up test or was employed for some other reason. It was assumed that the two methods were interchangeable and in order to check that assumption a series of 15 production samples was obtained, each of which was measured by both methods. These 15 pairs of observations are displayed in Table 1.1. (The choice of $n = 15$ pairs is merely for convenience in keeping the size of this example small; most quality control techniques would require more than this.)

What can one do with these data? The choices are almost endless. One possibility would be to compute the differences in the observed concentrations and test that the mean difference was zero, using the paired difference t-test based on the variability of the 15 differences. The analysis of variance technique would treat these data as a two-way ANOVA with methods and runs as factors. This would probably be a mixed model with *methods* being a fixed factor and *runs* generally assumed to be random. One would get the by-product of a run component of variability as well as an overall measure of inherent variability if the inherent variability of the two methods were the same. This assumption could be checked by a techniques such as the one due to Grubbs (1948, 1973) or that of Russell and Bradley (1958), which deal with heterogeneity of variance in two-way data arrays. Another complication could arise if the variability of the analyses was a function of level but a glance at the scattergram of the data shown in Figure 1.1 would seem to indicate that this is not the case.

Certainly, the preparation of Figure 1.1 is one of the first things to be considered, because in an example this small it would easily indicate any outliers or other aberrations in the data as well as provide a quick indication of the relationship between the two methods. Second, it would suggest the use of

Table 1.1. Data for Chemical Example

Obs. No.	Method 1	Method 2
1	10.0	10.7
2	10.4	9.8
3	9.7	10.0
4	9.7	10.1
5	11.7	11.5
6	11.0	10.8
7	8.7	8.8
8	9.5	9.3
9	10.1	9.4
10	9.6	9.6
11	10.5	10.4
12	9.2	9.0
13	11.3	11.6
14	10.1	9.8
15	8.5	9.2

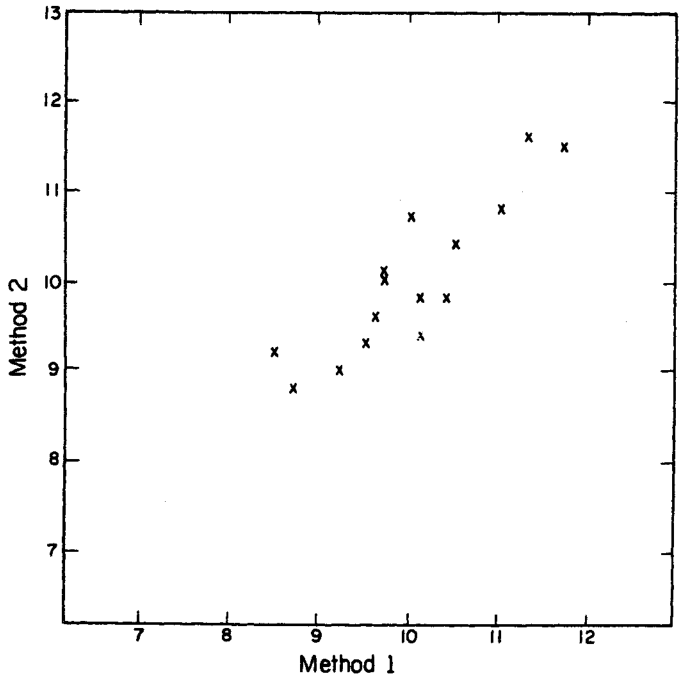

FIGURE 1.1. Chemical example: original data. Reproduced from Jackson (1980) with permission of the American Society for Quality Control and Jackson (1985) with permission of Marcel Dekker.

regression to determine to what extent it is possible to predict the results of one method from the other. However, the requirement that these two methods should be interchangeable means being able to predict in either direction, which (by using ordinary least-squares) would result in two different equations. The least-squares equation for predicting Method 1 from Method 2 minimizes the variability in Method 1 given a specific level of Method 2, while the equation for predicting Method 2 from Method 1 minimizes the variability in Method 2 given a specific level of Method 1.

A single prediction equation is required that could be used in either direction. One could invert either of the two regression equations, but which one and what about the theoretical consequences of doing this? The line that will perform this role directly is called the *orthogonal* regression line which minimizes the deviations *perpendicular to the line itself.* This line is obtained by the method of principal components and, in fact, was the first application of PCA, going back to Karl Pearson (1901). We shall obtain this line in the next section and in so doing will find that PCA will furnish us with a great deal of other information as well. Although many of these properties may seem superfluous for this small two-variable example, its size will allow us to easily understand these properties and the operations required to use PCA. This will be helpful when we then go on to larger problems.

In order to illustrate the method of PCA, we shall need to obtain the sample means, variances and the covariance between the two methods for the data in Table 1.1. Let x_{1k} be the test result for Method 1 for the kth run and the corresponding result for Method 2 be denoted by x_{2k}. The vector of sample means is

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 10.00 \\ 10.00 \end{bmatrix}$$

and the sample covariance matrix is

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix} = \begin{bmatrix} .7986 & .6793 \\ .6793 & .7343 \end{bmatrix}$$

where s_i^2 is the variance and the covariance is

$$s_{ij} = \frac{n\sum x_{ik}x_{jk} - \sum x_{ik}\sum x_{jk}}{[n(n-1)]}$$

with the index of summation, k, going over the entire sample of $n = 15$. Although the correlation between x_1 and x_2 is not required, it may be of interest to estimate this quantity, which is

$$r = \frac{s_{12}}{(s_1 s_2)} = .887$$

1.3 CHARACTERISTIC ROOTS AND VECTORS

The method of principal components is based on a key result from matrix algebra: A $p \times p$ symmetric, nonsingular matrix, such as the covariance matrix $\mathbf{S}$, may be reduced to a diagonal matrix $\mathbf{L}$ by premultiplying and postmultiplying it by a particular orthonormal matrix $\mathbf{U}$ such that

$$\mathbf{U}'\mathbf{S}\mathbf{U} = \mathbf{L} \tag{1.3.1}$$

The diagonal elements of $\mathbf{L}$, $l_1, l_2, \ldots, l_p$ are called the *characteristic roots, latent roots* or *eigenvalues* of $\mathbf{S}$. The columns of $\mathbf{U}$, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ are called the *characteristic vectors* or *eigenvectors* of $\mathbf{S}$. (Although the term *latent* vector is also correct, it often has a specialized meaning and it will not be used in this book except in that context.) The characteristic roots may be obtained from the solution of the following determinental equation, called the *characteristic equation*:

$$|\mathbf{S} - l\mathbf{I}| = 0 \tag{1.3.2}$$

where $\mathbf{I}$ is the identity matrix. This equation produces a pth degree polynomial in l from which the values $l_1, l_2, \ldots, l_p$ are obtained.

For this example, there are $p = 2$ variables and hence,

$$|\mathbf{S} - l\mathbf{I}| = \begin{vmatrix} .7986 - l & .6793 \\ .6793 & .7343 - l \end{vmatrix} = .124963 - 1.53291 + l^2 = 0$$

The values of l that satisfy this equation are $l_1 = 1.4465$ and $l_2 = .0864$.

The characteristic vectors may then be obtained by the solution of the equations

$$[\mathbf{S} - l\mathbf{I}]\mathbf{t}_i = 0 \tag{1.3.3}$$

and

$$\mathbf{u}_i = \frac{\mathbf{t}_i}{\sqrt{\mathbf{t}_i'\mathbf{t}_i}} \tag{1.3.4}$$

for $i = 1, 2, \ldots, p$. For this example, for $i = 1$,

$$[\mathbf{S} - l_1\mathbf{I}]\mathbf{t}_1 = \begin{bmatrix} .7986 - 1.4465 & .6793 \\ .6793 & .7343 - 1.4465 \end{bmatrix}\begin{bmatrix} t_{11} \\ t_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

These are two homogeneous linear equations in two unknowns. To solve, let $t_{11} = 1$ and use just the first equation:

$$-.6478 + .6793t_{21} = 0$$

The solution is $t_{21} = .9538$. These values are then placed in the normalizing equation (1.3.4) to obtain the first characteristic vector:

$$\mathbf{u}_1 = \frac{\mathbf{t}_1}{\sqrt{\mathbf{t}_1'\mathbf{t}_2}} = \frac{1}{\sqrt{1.9097}}\begin{bmatrix} 1.0 \\ .9538 \end{bmatrix} = \begin{bmatrix} .7236 \\ .6902 \end{bmatrix}$$

Similarly, using $l_2 = .0864$ and letting $t_{22} = 1$, the second characteristic vector is

$$\mathbf{u}_2 = \begin{bmatrix} -.6902 \\ .7236 \end{bmatrix}$$

These characteristic vectors make up the matrix

$$\mathbf{U} = [\mathbf{u}_1 \mid \mathbf{u}_2] = \left[\begin{array}{c:c} .7236 & -.6902 \\ .6902 & .7236 \end{array}\right]$$