Springer Series in Statistics

Advisors:
D. Brillinger, S. Fienberg, J. Gani,
J. Hartigan, K. Krickeberg

Springer Series in Statistics

D.F. Andrews and A.M. Herzberg, Data: A Collection of Problems from Many Fields for the Student and Research Worker. xx, 442 pages, 1985.

F.J. Anscombe, Computing in Statistical Science through APL. xvi, 426 pages, 1981.

J.O. Berger, Statistical Decision Theory and Bayesian Analysis, 2nd edition, xiv, 425 pages, 1985.

P. Brémaud, Point Processes and Queues: Martingale Dynamics. xviii, 354 pages, 1981.

K. Dzhaparidze, Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series. xii, 300 pages, 1985.

R.H. Farrell, Multivariate Calculation. xvi, 367 pages, 1985.

L.A. Goodman and W.H. Kruskal, Measures of Association for Cross Classifications. x, 146 pages, 1979.

J.A. Hartigan, Bayes Theory. xii, 145 pages, 1983.

H. Heyer, Theory of Statistical Experiments. x, 289 pages, 1982.

I.T. Joliffe, Principal Component Analysis. xiii, 272 pages, 1986.

M. Kres, Statistical Tables for Multivariate Analysis. xxii, 504 pages, 1983.

M.R. Leadbetter, G. Lindgren and H. Rootzén, Extremes and Related Properties of Random Sequences and Processes. xii, 336 pages, 1983.

L. LeCam, Asymptotic Methods in Statistical Decision Theory. xii, 700 pages, 1986.

E.B. Manoukian, Modern Concepts and Theorems of Mathematical Statistics. xiv, 156 pages, 1986.

R.G. Miller, Jr., Simultaneous Statistical Inference, 2nd edition. xvi, 299 pages, 1981.

F. Mosteller and D.S. Wallace, Applied Bayesian and Classical Inference: The Case of *The Federalist* Papers. xxxv, 301 pages, 1984.

D. Pollard, Convergence of Stochastic Processes. xiv, 215 pages, 1984.

J.W. Pratt and J.D. Gibbons, Concepts of Nonparametric Theory. xvi, 462 pages, 1981.

L. Sachs, Applied Statistics: A Handbook of Techniques, 2nd edition. xxviii, 706 pages, 1984.

E. Seneta, Non-Negative Matrices and Markov Chains. xv, 279 pages, 1981.

D. Siegmund, Sequential Analysis: Tests and Confidence Intervals. xii, 272 pages, 1985.

V. Vapnik, Estimation of Dependences Based on Empirical Data. xvi, 399 pages, 1982.

K.M. Wolter, Introduction to Variance Estimation. xii, 428 pages, 1985.

A.M. Yaglom, Correlation Theory of Stationary and Related Random Functions. x, 640 pages, 1986.

I. T. Jolliffe

Principal Component Analysis

With 26 Illustrations

Springer Science+Business Media, LLC

I. T. Jolliffe
Mathematical Institute
University of Kent
Canterbury
Kent CT2 7NF
England

AMS Classification: 62H25

Library of Congress Cataloging-in-Publication Data
Jolliffe, I. T.
Principal component analysis.
(Springer series in statistics)
Bibliography: p.
Includes index.
1. Principal components analysis. I. Title. II. Series.
QA278.5.J65 1986 519.5'35 85-27882

Originally published by Springer Verlag New York, Inc. in 1986
Softcover reprint of the hardcover 1st edition 1986

Typeset by Asco Trade Typesetting Ltd., North Point, Hong Kong.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4757-1906-2 ISBN 978-1-4757-1904-8 (eBook)
DOI 10.1007/978-1-4757-1904-8

Preface

Principal component analysis is probably the oldest and best known of the techniques of multivariate analysis. It was first introduced by Pearson (1901), and developed independently by Hotelling (1933). Like many multivariate methods, it was not widely used until the advent of electronic computers, but it is now well entrenched in virtually every statistical computer package.

The central idea of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first *few* retain most of the variation present in *all* of the original variables. Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem for a positive-semidefinite symmetric matrix. Thus, the definition and computation of principal components are straightforward but, as will be seen, this apparently simple technique has a wide variety of different applications, as well as a number of different derivations. Any feelings that principal component analysis is a narrow subject should soon be dispelled by the present book; indeed some quite broad topics which are related to principal component analysis receive no more than a brief mention in the final two chapters.

Although the term 'principal component analysis' is in common usage, and is adopted in this book, other terminology may be encountered for the same technique, particularly outside of the statistical literature. For example, the phrase 'empirical orthogonal functions' is common in meteorology, and in other fields the term 'factor analysis' may be used when 'principal component analysis' is meant. References to 'eigenvector analysis ' or 'latent vector analysis' may also camouflage principal component analysis. Finally,

some authors refer to principal components analysis rather than principal component analysis. To save space, the abbreviations PCA and PC will be used frequently in the present text.

The book should be useful to readers with a wide variety of backgrounds. Some knowledge of probability and statistics, and of matrix algebra, is necessary, but this knowledge need not be extensive for much of the book. It is expected, however, that most readers will have had some exposure to multivariate analysis in general before specializing to PCA. Many textbooks on multivariate analysis have a chapter or appendix on matrix algebra, e.g. Mardia *et al.* (1979, Appendix A), Morrison (1976, Chapter 2), Press (1972, Chapter 2), and knowledge of a similar amount of matrix algebra will be useful in the present book.

After an introductory chapter which gives a definition and derivation of PCA, together with a brief historical review, there are three main parts to the book. The first part, comprising Chapters 2 and 3, is mainly theoretical and some small parts of it require rather more knowledge of matrix algebra and vector spaces than is typically given in standard texts on multivariate analysis. However, it is not necessary to read all of these chapters in order to understand the second, and largest, part of the book. Readers who are mainly interested in applications could omit the more theoretical sections, although Sections 2.3, 2.4, 3.3, 3.4 and 3.8 are likely to be valuable to most readers; some knowledge of the singular value decomposition which is discussed in Section 3.5 will also be useful in some of the subsequent chapters.

This second part of the book is concerned with the various applications of PCA, and consists of Chapters 4 to 10 inclusive. Several chapters in this part refer to other statistical techniques, in particular from multivariate analysis. Familiarity with at least the basic ideas of multivariate analysis will therefore be useful, although each technique is explained briefly when it is introduced.

The third part, comprising Chapters 11 and 12, is a mixture of theory and potential applications. A number of extensions, generalizations and uses of PCA in special circumstances are outlined. Many of the topics covered in these chapters are relatively new, or outside the mainstream of statistics and, for several, their practical usefulness has yet to be fully explored. For these reasons they are covered much more briefly than the topics in earlier chapters.

The book is completed by an Appendix which contains two sections. The first describes some numerical algorithms for finding PCs, and the second describes the current availability of routines for performing PCA and related analyses in five well-known computer packages.

The coverage of individual chapters is now described in a little more detail. A standard definition and derivation of PCs is given in Chapter 1, but there are a number of alternative definitions and derivations, both geometric and algebraic, which also lead to PCs. In particular the PCs are 'optimal' linear functions of $\mathbf{x}$ with respect to several different criteria, and these various optimality criteria are described in Chapter 2. Also included in Chapter

2 are some other mathematical properties of PCs and a discussion of the use of correlation matrices, as opposed to covariance matrices, to derive PCs.

The derivation in Chapter 1, and all of the material of Chapter 2, is in terms of the *population* properties of a random vector $\mathbf{x}$. In practice, a *sample* of data is available, from which to estimate PCs, and Chapter 3 discusses the properties of PCs derived from a sample. Many of these properties correspond to population properties but some, for example those based on the singular value decomposition, are defined only for samples. A certain amount of distribution theory for sample PCs has been derived, almost exclusively asymptotic, and a summary of some of these results, together with related inference procedures, is also included in Chapter 3. Most of the technical details are, however, omitted. In PCA, only the first few PCs are conventionally deemed to be useful. However, some of the properties in Chapters 2 and 3, and an example in Chapter 3, show the potential usefulness of the last few, as well as the first few, PCs. Further uses of the last few PCs will be encountered in Chapters 6, 8 and 10. A final section of Chapter 3 discusses how PCs can sometimes be (approximately) deduced, without calculation, from the patterns of the covariance or correlation matrix.

Although the purpose of PCA, namely to reduce the number of variables from p to m ($\ll p$), is simple, the ways in which the PCs can actually be used are quite varied. At the simplest level, if a few uncorrelated variables (the first few PCs) reproduce most of the variation in all of the original variables, and if, further, these variables are interpretable, then the PCs give an alternative, much simpler, description of the data than the original variables. Examples of this use are given in Chapter 4, while subsequent chapters look at more specialized uses of the PCs.

Chapter 5 describes how PCs may be used to look at data graphically, Other graphical representations based on principal co-ordinate analysis, biplots and correspondence analysis, each of which have connections with PCA, are also discussed.

A common question in PCA is how many PCs are needed to account for 'most' of the variation in the original variables. A large number of rules has been proposed to answer this question, and Chapter 6 describes many of them. When PCA replaces a large set of variables by a much smaller set, the smaller set are new variables (the PCs) rather than a subset of the original variables. However, if a subset of the original variables is preferred, then the PCs can also be used to suggest suitable subsets. How this can be done is also discussed in Chapter 6.

In many texts on multivariate analysis, especially those written by non-statisticians, PCA is treated as though it is part of the factor analysis. Similarly, many computer packages give PCA as one of the options in a factor analysis subroutine. Chapter 7 explains that, although factor analysis and PCA have similar aims, they are, in fact, quite different techniques. There are, however, some ways in which PCA can be used in factor analysis and these are briefly described.

The use of PCA to 'orthogonalize' a regression problem, by replacing a set of highly correlated regressor variables by their PCs, is fairly well known. This technique, and several other related ways of using PCs in regression are discussed in Chapter 8.

Principal component analysis is sometimes used as a preliminary to, or in conjunction with, other statistical techniques, the obvious example being in regression, as described in Chapter 8. Chapter 9 discusses the possible uses of PCA in conjunction with three well-known multivariate techniques, namely discriminant analysis, cluster analysis and canonical correlation analysis.

It has been suggested that PCs, especially the last few, can be useful in the detection of outliers in a data set. This idea is discussed in Chapter 10, together with two different, but related, topics. One of these topics is the robust estimation of PCs when it is suspected that outliers may be present in the data, and the other is the evaluation, using influence functions, of which individual observations have the greatest effect on the PCs.

The last two chapters, 11 and 12, are mostly concerned with modifications or generalizations of PCA. The implications for PCA of special types of data are discussed in Chapter 11, with sections on discrete data, non-independent and time series data, compositional data, data from designed experiments, data with group structure, missing data and goodness-of-fit statistics. Most of these topics are covered rather briefly, as are a number of possible generalizations and adaptations of PCA which are described in Chapter 12.

Throughout the monograph various other multivariate techniques are introduced. For example, principal co-ordinate analysis and correspondence analysis appear in Chapter 5, factor analysis in Chapter 7, cluster analysis, discriminant analysis and canonical correlation analysis in Chapter 9, and multivariate analysis of variance in Chapter 11. However, it has not been the intention to give full coverage of multivariate methods or even to cover all those methods which reduce to eigenvalue problems. The various techniques have been introduced only where they are relevant to PCA and its application, and the relatively large number of techniques which have been mentioned is a direct result of the widely varied ways in which PCA can be used.

Throughout the book, a substantial number of examples are given, using data from a wide variety of areas of applications. However, no exercises have been included, since most potential exercises would fall into two narrow categories. One type would ask for proofs or extensions of the theoy given, in particular, in Chapters 2, 3 and 12, and would be exercises mainly in algebra rather than statistics. The second type would require PCAs to be performed and interpreted for various data sets. This is certainly a useful type of exercise, but many readers will find it most fruitful to analyse their own data sets. Furthermore, although the numerous examples given in the book should provide some guidance, there may not be a single 'correct' interpretation of a PCA.

Acknowledgements

My interest in principal component analysis was initiated, nearly 20 years ago, by John Scott, so he is, in one way, responsible for this book being written.

A number of friends and colleagues have commented on earlier drafts of parts of the book, or helped in other ways. I am grateful to Patricia Calder, Chris Folland, Nick Garnham, Tim Hopkins, Byron Jones, Wojtek Krzanowski, Philip North and Barry Vowden for their assistance and encouragement. Particular thanks are due to John Jeffers and Byron Morgan, who each read the entire text of an earlier version of the book, and made many constructive comments which substantially improved the final product. Any remaining errors and omissions are, of course, my responsibility, and I shall be glad to have them brought to my attention.

I have never ceased to be amazed by the patience and efficiency of Mavis Swain, who has expertly typed virtually all of the text, in its various drafts. I am extremely grateful to her, and also to my wife, Jean, who took over my rôle in the household during the last few hectic weeks of preparation. Finally, thanks to Anna, Jean and Nils for help with indexing and proof-reading.

Contents