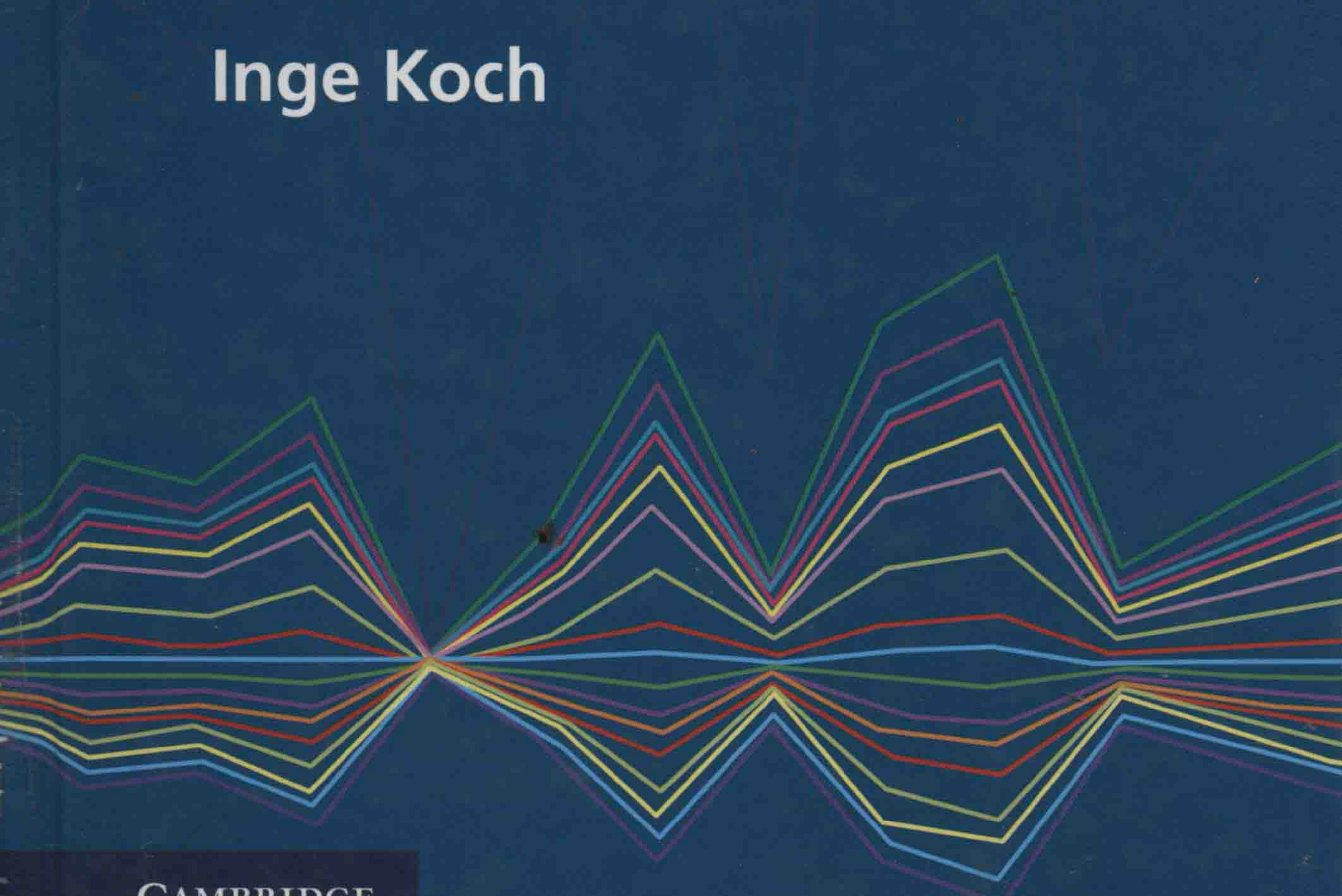
Cambridge Series in Statistical and Probabilistic Mathematics

Analysis of Multivariate and High-Dimensional Data

Inge Koch



CAMBRIDGE

Analysis of Multivariate and High-Dimensional Data

Inge Koch
University of Adelaide, Australia

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9780521887939

First published 2014

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data
Koch, Inge, 1952–
Analysis of multivariate and high-dimensional data / Inge Koch.
pages cm
ISBN 978-0-521-88793-9 (hardback)
1. Multivariate analysis. 2. Big data. I. Title.
QA278.K5935 2013
519.5′35–dc23 2013013351

ISBN 978-0-521-88793-9 Hardback

Additional resources for this publication at www.cambridge.org/9780521887939

Analysis of Multivariate and High-Dimensional Data

'Big data' poses challenges that require both classical multivariate methods and contemporary techniques from machine learning and engineering. This modern text integrates the two strands into a coherent treatment, drawing together theory, data, computation and recent research.

The theoretical framework includes formal definitions, theorems and proofs which clearly set out the guaranteed 'safe operating zone' for the methods and allow users to assess whether data are in or near the zone. Extensive examples showcase the strengths and limitations of different methods in a range of cases: small classical data; data from medicine, biology, marketing and finance; high-dimensional data from bioinformatics; functional data from proteomics; and simulated data. High-dimension low sample size data get special attention. Several data sets are revisited repeatedly to allow comparison of methods. Generous use of colour, algorithms, MATLAB code and problem sets completes the package. The text is suitable for graduate students in statistics and researchers in data-rich disciplines.

INGE KOCH is Associate Professor of Statistics at the University of Adelaide, Australia.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

A complete list of books in the series can be found at www.cambridge.org/statistics.
Recent titles include the following:

11. *Statistical Models,* by A. C. Davison
12. *Semiparametric Regression,* by David Ruppert, M. P. Wand and R. J. Carroll
13. *Exercises in Probability,* by Loïc Chaumont and Marc Yor
14. *Statistical Analysis of Stochastic Processes in Time,* by J. K. Lindsey
15. *Measure Theory and Filtering,* by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference,* by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory,* by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems,* by Anton Bovier
19. *The Coordinate-Free Approach to Linear Models,* by Michael J. Wichura
20. *Random Graph Dynamics,* by Rick Durrett
21. *Networks,* by Peter Whittle
22. *Saddlepoint Approximations with Applications,* by Ronald W. Butler
23. *Applied Asymptotics,* by A. R. Brazzale, A. C. Davison and N. Reid
24. *Random Networks for Communication,* by Massimo Franceschetti and Ronald Meester
25. *Design of Comparative Experiments,* by R. A. Bailey
26. *Symmetry Studies,* by Marlos A. G. Viana
27. *Model Selection and Model Averaging,* by Gerda Claeskens and Nils Lid Hjort
28. *Bayesian Nonparametrics,* edited by Nils Lid Hjort *et al.*
29. *From Finite Sample to Asymptotic Methods in Statistics,* by Pranab K. Sen, Julio M. Singer and Antonio C. Pedrosa de Lima
30. *Brownian Motion,* by Peter Mörters and Yuval Peres
31. *Probability (Fourth Edition),* by Rick Durrett
33. *Stochastic Processes,* by Richard F. Bass
34. *Regression for Categorical Data,* by Gerhard Tutz
35. *Exercises in Probability (Second Edition),* by Loïc Chaumont and Marc Yor
36. *Statistical Principles for the Design of Experiments,* by R. Mead, S. G. Gilmour and A. Mead

To Alun, Graeme and Reiner

List of Algorithms

Notation

$\mathbf{a} = [a_1 \cdots a_d]^{\mathsf{T}}$	Column vector in $\mathbb{R}^d$
$A,\ A^{\mathsf{T}},\ B$	Matrices, with A^{T} the transpose of A
$A_{p\times q},\ B_{r\times s}$	Matrices A and B of size $p \times q$ and $r \times s$
$A = (a_{ij})$	Matrix A with entries a_{ij}
$A = \begin{bmatrix}\mathbf{a}_1 \cdots \mathbf{a}_p\end{bmatrix} = \begin{bmatrix}\mathbf{a}_{\bullet 1}\\ \vdots \\ \mathbf{a}_{\bullet q}\end{bmatrix}$	Matrix A of size $p \times q$ with columns $\mathbf{a}_i$, rows $\mathbf{a}_{\bullet j}$
A_{diag}	Diagonal matrix consisting of the diagonal entries of A
$\mathbf{0}_{k\times\ell}$	$k \times \ell$ matrix with all entries 0
$\mathbf{1}_k$	Column vector with all entries 1
$\mathbf{I}_{d\times d}$	$d \times d$ identity matrix
$\mathbf{I}_{k\times\ell}$	$\begin{pmatrix}\mathbf{I}_{k\times k} & \mathbf{0}_{k\times(\ell-k)}\end{pmatrix}$ if $k \le \ell$, and $\begin{pmatrix}\mathbf{I}_{\ell\times\ell}\\ \mathbf{0}_{(k-\ell)\times\ell}\end{pmatrix}$ if $k \ge \ell$
$\mathbf{X}, \mathbf{Y}$	d-dimensional random vectors
$\mathbf{X}\vert\mathbf{Y}$	Conditional random vector $\mathbf{X}$ given $\mathbf{Y}$
$\mathbf{X} = \begin{bmatrix}X_1 \ldots X_d\end{bmatrix}^{\mathsf{T}}$	d-dimensional random vector with entries (or variables) X_j
$\mathbb{X} = \begin{bmatrix}\mathbf{X}_1\ \mathbf{X}_2 \cdots \mathbf{X}_n\end{bmatrix}$	$d \times n$ data matrix of random vectors $\mathbf{X}_i,\ i \le n$
$\mathbf{X}_i = \begin{bmatrix}X_{i1} \cdots X_{id}\end{bmatrix}^{\mathsf{T}}$	Random vector from $\mathbb{X}$ with entries X_{ij}
$\mathbf{X}_{\bullet j} = \begin{bmatrix}X_{1j} \cdots X_{nj}\end{bmatrix}$	$1 \times n$ row vector of the jth variable of $\mathbb{X}$
$\boldsymbol{\mu} = \mathbb{E}\mathbf{X}$	Expectation of a random vector $\mathbf{X}$, also denoted by $\boldsymbol{\mu}_X$
$\overline{\mathbf{X}}$	Sample mean
$\overline{\overline{\mathbf{X}}}$	Average sample class mean
$\sigma^2 = \text{var}(X)$	Variance of random variable X
$\Sigma = \text{var}(\mathbf{X})$	Covariance matrix of $\mathbf{X}$ with entries σ_{jk} and $\sigma_{jj} = \sigma_j^2$
S	Sample covariance matrix of $\mathbb{X}$ with entries s_{ij} and $s_{jj} = s_j^2$
$R = (\rho_{ij}),\ R_S = (\widehat{\rho}_{ij})$	Matrix of correlation coefficients for the population and sample
$Q_{(n)},\ Q^{(d)}$	Dual matrices $Q_{(n)} = \mathbb{X}\mathbb{X}^{\mathsf{T}}$ and $Q^{(d)} = \mathbb{X}^{\mathsf{T}}\mathbb{X}$
Σ_{diag}	Diagonal matrix with entries σ_j^2 obtained from Σ
S_{diag}	Diagonal matrix with entries s_j^2 obtained from S
$\Sigma = \Gamma\Lambda\Gamma^{\mathsf{T}}$	Spectral decomposition of Σ

$\Gamma_k = [\boldsymbol{\eta}_1 \cdots \boldsymbol{\eta}_k]$	$d \times k$ matrix of (orthogonal) eigenvectors of Σ, $k \le d$
$\Lambda_k = \text{diag}(\lambda_1, \ldots, \lambda_k)$	Diagonal $k \times k$ matrix with diagonal entries the eigenvalues of Σ, $k \le d$
$S = \widehat{\Gamma}\widehat{\Lambda}\widehat{\Gamma}^{\mathsf{T}}$	Spectral decomposition of S with eigenvalues $\widehat{\lambda}_j$ and eigenvectors $\widehat{\boldsymbol{\eta}}_j$
$\beta_3(\mathbf{X})$, $b_3(\mathbb{X})$	Multivariate skewness of $\mathbf{X}$ and sample skewness of $\mathbb{X}$
$\beta_4(\mathbf{X})$, $b_4(\mathbb{X})$	Multivariate kurtosis of $\mathbf{X}$ and sample kurtosis of $\mathbb{X}$
f, F	Multivariate probability density and distribution functions
ϕ, Φ	Standard normal probability density and distribution functions
f, f_G	Multivariate probability density functions; f and f_G have the same mean and covariance matrix, and f_G is Gaussian
$L(\theta)$ or $L(\theta \vert \mathbb{X})$	Likelihood function of the parameter θ, given $\mathbb{X}$
$\mathbf{X} \sim (\boldsymbol{\mu}, \Sigma)$	Random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ
$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$	Random vector from the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ
$\mathbb{X} \sim \mathsf{Sam}(\overline{\mathbf{X}}, S)$	Data – with sample mean $\overline{\mathbf{X}}$ and sample covariance matrix S
$\mathbb{X}_{\text{cent}}$	Centred data $[\mathbf{X}_1 - \overline{\mathbf{X}} \cdots \mathbf{X}_n - \overline{\mathbf{X}}]$, also written as $\mathbb{X} - \overline{\mathbf{X}}$
$\mathbf{X}_\Sigma$, $\mathbb{X}_S$	Sphered vector and data $\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, $S^{-1/2}(\mathbb{X} - \overline{\mathbf{X}})$
$\mathbf{X}_{\text{scale}}$, $\mathbb{X}_{\text{scale}}$	Scaled vector and data $\Sigma_{\text{diag}}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, $S_{\text{diag}}^{-1/2}(\mathbb{X} - \overline{\mathbf{X}})$
$\mathbf{X}^\diamond$, $\mathbb{X}^\diamond$	(Spatially) whitened random vector and data
$\mathbf{W}^{(k)} = [W_1 \cdots W_k]^{\mathsf{T}}$	Vector of first k principal component scores
$\mathbb{W}^{(k)} = [\mathbf{W}_{\bullet 1} \ldots \mathbf{W}_{\bullet k}]^{\mathsf{T}}$	$k \times n$ matrix of first k principal component scores
$\mathbf{P}_k = W_k \boldsymbol{\eta}_k$	Principal component projection vector
$\mathbb{P}_{\bullet k} = \widehat{\boldsymbol{\eta}}_k \mathbf{W}_{\bullet k}$	$d \times n$ matrix of principal component projections
$\mathbf{F}$, $\mathbb{F}$	Common factor for population and data in k-factor model
$\mathbf{S}$, $\mathbb{S}$	Source for population and data in Independent Component Analysis
$\mathcal{O} = \{O_1, \ldots, O_n\}$	Set of objects corresponding to data $\mathbb{X} = [\mathbf{X}_1\ \mathbf{X}_2 \ldots \mathbf{X}_n]$
$\{\mathcal{O}, \varrho\}$	Observed data, consisting of objects O_i and dissimilarities ϱ_{ik} between pairs of objects
$\mathfrak{f}$, $\mathfrak{f}(\mathbf{X})$, $\mathfrak{f}(\mathbb{X})$	Feature map, feature vector and feature data
$\text{cov}(\mathbf{X}, \mathbf{T})$	$d_X \times d_T$ (between) covariance matrix of $\mathbf{X}$ and $\mathbf{T}$
$\Sigma_{12} = \text{cov}(\mathbf{X}^{[1]}, \mathbf{X}^{[2]})$	$d_1 \times d_2$ (between) covariance matrix of $\mathbf{X}^{[1]}$ and $\mathbf{X}^{[2]}$
$S_{12} = \text{cov}(\mathbb{X}^{[1]}, \mathbb{X}^{[2]})$	$d_1 \times d_2$ sample (between) covariance matrix of $d_\ell \times n$ data $\mathbb{X}^{[\ell]}$, for $\ell = 1, 2$
$C = \Sigma_1^{-1/2} \Sigma_{12} \Sigma_2^{-1/2}$	Canonical correlation matrix of $\mathbf{X}^{[\ell]} \sim (\boldsymbol{\mu}_\ell, \Sigma_\ell)$, for $\ell = 1, 2$
$C = P \Upsilon Q^{\mathsf{T}}$	Singular value decomposition of C with singular values υ_j and eigenvectors $\mathbf{p}_j, \mathbf{q}_j$
$\widehat{C} = \widehat{P}\widehat{\Upsilon}\widehat{Q}^{\mathsf{T}}$	Sample canonical correlation matrix and its singular value decomposition
$R^{[C,1]} = CC^{\mathsf{T}}$, $R^{[C,2]} = C^{\mathsf{T}}C$	Matrices of multivariate coefficients of determination, with C the canonical correlation matrix

$\mathbf{U}^{(k)}, \mathbf{V}^{(k)}$	Pair of vectors of k-dimensional canonical correlations
$\boldsymbol{\varphi}_k, \boldsymbol{\psi}_k$	kth pair of canonical (correlation) transforms
$\mathbb{U}^{(k)}, \mathbb{V}^{(k)}$	$k \times n$ matrices of k-dimensional canonical correlation data
$\mathcal{C}_\nu$	νth class (or cluster)
$\mathfrak{r}$	Discriminant rule or classifier
h, h_β	Decision function for a discriminant rule $\mathfrak{r}$ (h_β depends on β)
$\mathfrak{b}, \mathfrak{w}$	Between-class and within-class variability
$\mathcal{E}$	(Classification) error
$\mathcal{P}(\mathbb{X}, k)$	k-cluster arrangement of $\mathbb{X}$
$A \circ B$	Hadamard or Schur product of matrices A and B
$\operatorname{tr}(A)$	Trace of a matrix A
$\det(A)$	Determinant of a matrix A
$\operatorname{dir}(\mathbf{X})$	Direction (vector) $\mathbf{X}/\ \|\mathbf{X}\|$ of $\mathbf{X}$
$\|\cdot\|$, $\|\cdot\|_p$,	(Euclidean) norm, p-norm of a vector or matrix
$\|\mathbf{X}\|_{tr}$	Trace norm of $\mathbf{X}$ given by $[\operatorname{tr}(\Sigma)]^{1/2}$
$\|A\|_{\text{Frob}}$	Frobenius norm of a matrix A given by $[\operatorname{tr}(AA^{\mathsf{T}})]^{1/2}$
$\Delta(\mathbf{X}, \mathbf{Y})$	Distance between vectors $\mathbf{X}$ and $\mathbf{Y}$
$\varrho(\mathbf{X}, \mathbf{Y})$	Dissimilarity of vectors $\mathbf{X}$ and $\mathbf{Y}$
$\mathfrak{a}(\boldsymbol{\alpha}, \boldsymbol{\beta})$	Angle between directions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$: $\arccos(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\beta})$
$\mathcal{H}$, $\mathcal{I}$, $\mathcal{J}$, $\mathcal{K}$	Entropy, mutual information, negentropy and Kullback-Leibler divergence
$\mathcal{Q}$	Projection index
$\mathbf{n} \succ \mathbf{d}$	Asymptotic domain, d fixed and $n \to \infty$
$\mathbf{n} \succeq \mathbf{d}$	Asymptotic domain, $d, n \to \infty$ and $d = O(n)$
$\mathbf{n} \preceq \mathbf{d}$	Asymptotic domain, $d, n \to \infty$ and $n = o(d)$
$\mathbf{n} \prec \mathbf{d}$	Asymptotic domain, n fixed and $d \to \infty$

Preface

This book is about data in many – and sometimes very many – variables and about analysing such data. The book attempts to integrate classical multivariate methods with contemporary methods suitable for high-dimensional data and to present them in a coherent and transparent framework. Writing about ideas that emerged more than a hundred years ago and that have become increasingly relevant again in the last few decades is exciting and challenging. With hindsight, we can reflect on the achievements of those who paved the way, whose methods we apply to ever bigger and more complex data and who will continue to influence our ideas and guide our research. Renewed interest in the classical methods and their extension has led to analyses that give new insight into data and apply to bigger and more complex problems.

There are two players in this book: *Theory* and *Data*. *Theory* advertises its wares to lure *Data* into revealing its secrets, but *Data* has its own ideas. *Theory* wants to provide elegant solutions which answer many but not all of *Data*'s demands, but these lead *Data* to pose new challenges to *Theory*. Statistics thrives on interactions between theory and data, and we develop better theory when we 'listen' to data. Statisticians often work with experts in other fields and analyse data from many different areas. We, the statisticians, need and benefit from the expertise of our colleagues in the analysis of their data and interpretation of the results of our analysis. At times, existing methods are not adequate, and new methods need to be developed.

This book attempts to combine theoretical ideas and advances with their application to data, in particular, to interesting and real data. I do not shy away from stating theorems as they are an integral part of the ideas and methods. Theorems are important because they summarise what we know and the conditions under which we know it. They tell us when methods may work with particular data; the hypotheses may not always be satisfied exactly, but a method may work nevertheless. The precise details do matter sometimes, and theorems capture this information in a concise way.

Yet a balance between theoretical ideas and data analysis is vital. An important aspect of any data analysis is its interpretation, and one might ask questions like: What does the analysis tell us about the data? What new insights have we gained from a particular analysis? How suitable is my method for my data? What are the limitations of a particular method, and what other methods would produce more appropriate analyses? In my attempts to answer such questions, I endeavour to be objective and emphasise the strengths and weaknesses of different approaches.

Who Should Read This Book?

This book is suitable for readers with various backgrounds and interests and can be read at different levels. It is appropriate as a graduate-level course – two course outlines are suggested in the section 'Teaching from This Book'. A second or more advanced course could make use of the more advanced sections in the early chapters and include some of the later chapters. The book is equally appropriate for working statisticians who need to find and apply a relevant method for analysis of their multivariate or high-dimensional data and who want to understand how the chosen method deals with the data, what its limitations might be and what alternatives are worth considering.

Depending on the expectation and aims of the reader, different types of backgrounds are needed. Experience in the analysis of data combined with some basic knowledge of statistics and statistical inference will suffice if the main aim involves applying the methods of this book. To understand the underlying theoretical ideas, the reader should have a solid background in the theory and application of statistical inference and multivariate regression methods and should be able to apply confidently ideas from linear algebra and real analysis.

Readers interested in **statistical ideas and their application to data** may benefit from the theorems and their illustrations in the examples. These readers may, in a first journey through the book, want to focus on the basic ideas and properties of each method and leave out the last few more advanced sections of each chapter. For possible paths, see the models for a one-semester course later in this preface.

Researchers and graduate students with a good background in statistics and mathematics who are primarily interested in the **theoretical developments of the different topics** will benefit from the formal setting of definitions, theorems and proofs and the careful distinction of the population and the sample case. This setting makes it easy to understand what each method requires and which ideas can be adapted. Some of these readers may want to refer to the recent literature and the references I provide for theorems that I do not prove.

Yet another broad group of readers may want to focus on **applying the methods of this book to particular data**, with an emphasis on the results of the data analysis and the new insight they gain into their data. For these readers, the interpretation of their results is of prime interest, and they can benefit from the many examples and discussions of the analysis for the different data sets. These readers could concentrate on the descriptive parts of each method and the interpretative remarks which follow many theorems and need not delve into the theorem/proof framework of the book.

Outline

This book consists of **three parts**. Typically, each method corresponds to a single chapter, and because the methods have different origins and varied aims, it is convenient to group the chapters into parts. The methods focus on two main themes:

1. *Component Analysis*, which aims to simplify the data by summarising them in a smaller number of more relevant or more interesting components

2. *Statistical Learning*, which aims to group, classify or regress the (component) data by partitioning the data appropriately or by constructing rules and applying these rules to new data.

The two themes are related, and each method I describe addresses at least one of the themes.

The first chapter in each part presents notation and summarises results required in the following chapters. I give references to background material and to proofs of results, which may help readers not acquainted with some topics. Readers who are familiar with the topics of the three first chapters in their part may only want to refer to the notation. Properties or theorems in these three chapters are called *Results* and are stated without proof. Each of the main chapters in the three parts is dedicated to a specific method or topic and illustrates its ideas on data.

Part I deals with the classical methods *Principal Component Analysis, Canonical Correlation Analysis* and *Discriminant Analysis*, which are 'musts' in multivariate analysis as they capture essential aspects of analysing multivariate data. The later sections of each of these chapters contain more advanced or more recent ideas and results, such as Principal Component Analysis for high-dimension low sample size data and Principal Component Regression. These sections can be left out in a first reading of the book without greatly affecting the understanding of the rest of Parts I and II.

Part II complements Part I and is still classical in its origin: *Cluster Analysis* is similar to Discriminant Analysis and partitions data but without the advantage of known classes. *Factor Analysis* and Principal Component Analysis enrich and complement each other, yet the two methods pursue distinct goals and differ in important ways. Classical *Multidimensional Scaling* may seem to be different from Principal Component Analysis, but Multidimensional Scaling, which ventures into non-linear component analysis, can be regarded as a generalisation of Principal Component Analysis. The three methods, Principal Component Analysis, Factor Analysis and Multidimensional Scaling, paved the way for non-Gaussian component analysis and in particular for Independent Component Analysis and Projection Pursuit.

Part III gives an overview of more recent and current ideas and developments in component analysis methods and links these to statistical learning ideas and research directions for high-dimensional data. A natural starting point are the twins, *Independent Component Analysis* and *Projection Pursuit*, which stem from the signal-processing and statistics communities, respectively. Because of their similarities as well as their resulting analysis of data, we may regard both as non-Gaussian component analysis methods. Since the early 1980s, when Independent Component Analysis and Projection Pursuit emerged, the concept of independence has been explored by many authors. Chapter 12 showcases *Independent Component Methods* which have been developed since about 2000. There are many different approaches; I have chosen some, including *Kernel Independent Component Analysis*, which have a more statistical rather than heuristic basis. The final chapter returns to the beginning – *Principal Component Analysis* – but focuses on current ideas and research directions: feature selection, component analysis of high-dimension low sample size data, decision rules for such data, asymptotics and consistency results when the dimension increases faster than the sample size. This last chapter includes *inconsistency* results and concludes with a new and general asymptotic framework for Principal Component Analysis which covers the different asymptotic domains of sample size and dimension of the data.

Data and Examples

This book uses many contrasting data sets: small classical data sets such as Fisher's four-dimensional iris data; data sets of moderate dimension (up to about thirty) from medical, biological, marketing and financial areas; and big and complex data sets. The data sets vary in the number of observations from fewer than fifty to about one million. We will also generate data and work with simulated data because such data can demonstrate the performance of a particular method, and the strengths and weaknesses of particular approaches more clearly. We will meet high-dimensional data with more than 1,000 variables and high-dimension low sample size (HDLSS) data, including data from genomics with dimensions in the tens of thousands, and typically fewer than 100 observations. In addition, we will encounter functional data from proteomics, for which each observation is a curve or profile.

Visualising data is important and is typically part of a first exploratory step in data analysis. If appropriate, I show the results of an analysis in graphical form.

I describe the analysis of data in *Examples* which illustrate the different tools and methodologies. In the examples I provide relevant information about the data, describe each analysis and give an interpretation of the outcomes of the analysis. As we travel through the book, we frequently return to data we previously met. The *Data Index* shows, for each data set in this book, which chapter contains examples pertaining to these data. Continuing with the same data throughout the book gives a more comprehensive picture of how we can study a particular data set and what methods and analyses are suitable for specific aims and data sets. Typically, the data sets I use are available on the Cambridge University Press website www.cambridge.org/9780521887939.

Use of Software and Algorithms

I use MATLAB for most of the examples and make generic MATLAB code available on the Cambridge University Press website. Readers and data analysts who prefer R could use that software instead of MATLAB. There are, however, some differences in implementation between MATLAB and R, in particular, in the Independent Component Analysis algorithms. I have included some comments about these differences in Section 11.4.

Many of the methods in Parts I and II have a standard one-line implementation in MATLAB and R. For example, to carry out a Principal Component Analysis or a likelihood-based Factor Analysis, all that is needed is a single command which includes the data to be analysed. These stand-alone routines in MATLAB and R avoid the need for writing one's own code. The MATLAB code I provide typically includes the initial visualisation of the data and, where appropriate, code for a graphical presentation of the results.

This book contains algorithms, that is, descriptions of the mathematical or computational steps that are needed to carry out particular analyses. Algorithm 4.2, for example, details the steps that are required to carry out classification for principal component data. A list of all algorithms in this book follows the Contents.

Theoretical Framework

I have chosen the conventional format with *Definitions, Theorems, Propositions* and *Corollaries*. This framework allows me to state assumptions and conclusions precisely and in an