

## 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

---

*Editorial Board*

- Z. Ghahramani (Department of Engineering, University of Cambridge)  
 R. Gill (Mathematical Institute, Leiden University)  
 F. P. Kelly (Department of Pure Mathematics and Mathematical Statistics,  
 University of Cambridge)  
 B. D. Ripley (Department of Statistics, University of Oxford)  
 S. Ross (Department of Industrial and Systems Engineering,  
 University of Southern California)  
 M. Stein (Department of Statistics, University of Chicago)

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](http://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

Cambridge University Press  
978-0-521-88793-9 - Analysis of Multivariate and High-Dimensional Data  
Inge Koch  
Frontmatter  
[More information](#)

---

# Analysis of Multivariate and High-Dimensional Data

Inge Koch  
*University of Adelaide, Australia*



Cambridge University Press  
978-0-521-88793-9 - Analysis of Multivariate and High-Dimensional Data  
Inge Koch  
Frontmatter  
[More information](#)

**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](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9780521887939](http://www.cambridge.org/9780521887939)

© Inge Koch 2014

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

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](http://www.cambridge.org/9780521887939)

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Cambridge University Press  
978-0-521-88793-9 - Analysis of Multivariate and High-Dimensional Data  
Inge Koch  
Frontmatter  
[More information](#)

---

*To Alun, Graeme and Reiner*

Cambridge University Press

978-0-521-88793-9 - Analysis of Multivariate and High-Dimensional Data

Inge Koch

Frontmatter

[More information](#)

---

---

## Contents

<i>List of Algorithms</i>	<i>page</i> xiii
<i>Notation</i>	xv
<i>Preface</i>	xix
<b>I CLASSICAL METHODS</b>	
<b>1 Multidimensional Data</b>	3
1.1 Multivariate and High-Dimensional Problems	3
1.2 Visualisation	4
1.2.1 Three-Dimensional Visualisation	4
1.2.2 Parallel Coordinate Plots	6
1.3 Multivariate Random Vectors and Data	8
1.3.1 The Population Case	8
1.3.2 The Random Sample Case	9
1.4 Gaussian Random Vectors	11
1.4.1 The Multivariate Normal Distribution and the Maximum Likelihood Estimator	11
1.4.2 Marginal and Conditional Normal Distributions	13
1.5 Similarity, Spectral and Singular Value Decomposition	14
1.5.1 Similar Matrices	14
1.5.2 Spectral Decomposition for the Population Case	14
1.5.3 Decompositions for the Sample Case	16
<b>2 Principal Component Analysis</b>	18
2.1 Introduction	18
2.2 Population Principal Components	19
2.3 Sample Principal Components	22
2.4 Visualising Principal Components	27
2.4.1 Scree, Eigenvalue and Variance Plots	27
2.4.2 Two- and Three-Dimensional PC Score Plots	30
2.4.3 Projection Plots and Estimates of the Density of the Scores	31
2.5 Properties of Principal Components	34
2.5.1 Correlation Structure of $\mathbf{X}$ and Its PCs	34
2.5.2 Optimality Properties of PCs	37

viii	<i>Contents</i>	
2.6	Standardised Data and High-Dimensional Data	42
2.6.1	Scaled and Sphered Data	42
2.6.2	High-Dimensional Data	47
2.7	Asymptotic Results	55
2.7.1	Classical Theory: Fixed Dimension $d$	55
2.7.2	Asymptotic Results when $d$ Grows	57
2.8	Principal Component Analysis, the Number of Components and Regression	62
2.8.1	Number of Principal Components Based on the Likelihood	62
2.8.2	Principal Component Regression	65
<b>3</b>	<b>Canonical Correlation Analysis</b>	70
3.1	Introduction	70
3.2	Population Canonical Correlations	71
3.3	Sample Canonical Correlations	76
3.4	Properties of Canonical Correlations	82
3.5	Canonical Correlations and Transformed Data	88
3.5.1	Linear Transformations and Canonical Correlations	88
3.5.2	Transforms with Non-Singular Matrices	90
3.5.3	Canonical Correlations for Scaled Data	98
3.5.4	Maximum Covariance Analysis	100
3.6	Asymptotic Considerations and Tests for Correlation	100
3.7	Canonical Correlations and Regression	104
3.7.1	The Canonical Correlation Matrix in Regression	105
3.7.2	Canonical Correlation Regression	108
3.7.3	Partial Least Squares	109
3.7.4	The Generalised Eigenvalue Problem	113
<b>4</b>	<b>Discriminant Analysis</b>	116
4.1	Introduction	116
4.2	Classes, Labels, Rules and Decision Functions	118
4.3	Linear Discriminant Rules	120
4.3.1	Fisher's Discriminant Rule for the Population	120
4.3.2	Fisher's Discriminant Rule for the Sample	123
4.3.3	Linear Discrimination for Two Normal Populations or Classes	127
4.4	Evaluation of Rules and Probability of Misclassification	129
4.4.1	Boundaries and Discriminant Regions	129
4.4.2	Evaluation of Discriminant Rules	131
4.5	Discrimination under Gaussian Assumptions	136
4.5.1	Two and More Normal Classes	136
4.5.2	Gaussian Quadratic Discriminant Analysis	140
4.6	Bayesian Discrimination	143
4.6.1	Bayes Discriminant Rule	143
4.6.2	Loss and Bayes Risk	146
4.7	Non-Linear, Non-Parametric and Regularised Rules	148
4.7.1	Nearest-Neighbour Discrimination	149



*Contents*

ix

4.7.2	Logistic Regression and Discrimination	153
4.7.3	Regularised Discriminant Rules	154
4.7.4	Support Vector Machines	155
4.8	Principal Component Analysis, Discrimination and Regression	157
4.8.1	Discriminant Analysis and Linear Regression	157
4.8.2	Principal Component Discriminant Analysis	158
4.8.3	Variable Ranking for Discriminant Analysis	159
	<i>Problems for Part I</i>	165
<b>II FACTORS AND GROUPINGS</b>		
<b>5</b>	<b>Norms, Proximities, Features and Dualities</b>	175
5.1	Introduction	175
5.2	Vector and Matrix Norms	176
5.3	Measures of Proximity	176
5.3.1	Distances	176
5.3.2	Dissimilarities	178
5.3.3	Similarities	179
5.4	Features and Feature Maps	180
5.5	Dualities for $\mathbb{X}$ and $\mathbb{X}^T$	181
<b>6</b>	<b>Cluster Analysis</b>	183
6.1	Introduction	183
6.2	Hierarchical Agglomerative Clustering	185
6.3	$k$ -Means Clustering	191
6.4	Second-Order Polynomial Histogram Estimators	199
6.5	Principal Components and Cluster Analysis	207
6.5.1	$k$ -Means Clustering for Principal Component Data	207
6.5.2	Binary Clustering of Principal Component Scores and Variables	210
6.5.3	Clustering High-Dimensional Binary Data	212
6.6	Number of Clusters	216
6.6.1	Quotients of Variability Measures	216
6.6.2	The Gap Statistic	217
6.6.3	The Prediction Strength Approach	219
6.6.4	Comparison of $\hat{k}$ -Statistics	220
<b>7</b>	<b>Factor Analysis</b>	223
7.1	Introduction	223
7.2	Population $k$ -Factor Model	224
7.3	Sample $k$ -Factor Model	227
7.4	Factor Loadings	228
7.4.1	Principal Components and Factor Analysis	228
7.4.2	Maximum Likelihood and Gaussian Factors	233
7.5	Asymptotic Results and the Number of Factors	236
7.6	Factor Scores and Regression	239

7.6.1	Principal Component Factor Scores	239
7.6.2	Bartlett and Thompson Factor Scores	240
7.6.3	Canonical Correlations and Factor Scores	241
7.6.4	Regression-Based Factor Scores	242
7.6.5	Factor Scores in Practice	244
7.7	Principal Components, Factor Analysis and Beyond	245
<b>8</b>	<b>Multidimensional Scaling</b>	248
8.1	Introduction	248
8.2	Classical Scaling	249
8.2.1	Classical Scaling and Principal Coordinates	251
8.2.2	Classical Scaling with Strain	254
8.3	Metric Scaling	257
8.3.1	Metric Dissimilarities and Metric Stresses	258
8.3.2	Metric Strain	261
8.4	Non-Metric Scaling	263
8.4.1	Non-Metric Stress and the Shepard Diagram	263
8.4.2	Non-Metric Strain	268
8.5	Data and Their Configurations	268
8.5.1	HDLSS Data and the $\mathbb{X}$ and $\mathbb{X}^T$ Duality	269
8.5.2	Procrustes Rotations	271
8.5.3	Individual Differences Scaling	273
8.6	Scaling for Grouped and Count Data	274
8.6.1	Correspondence Analysis	274
8.6.2	Analysis of Distance	279
8.6.3	Low-Dimensional Embeddings	282
	<i>Problems for Part II</i>	286
<b>III</b>	<b>NON-GAUSSIAN ANALYSIS</b>	
<b>9</b>	<b>Towards Non-Gaussianity</b>	295
9.1	Introduction	295
9.2	Gaussianity and Independence	296
9.3	Skewness, Kurtosis and Cumulants	297
9.4	Entropy and Mutual Information	299
9.5	Training, Testing and Cross-Validation	301
9.5.1	Rules and Prediction	302
9.5.2	Evaluating Rules with the Cross-Validation Error	302
<b>10</b>	<b>Independent Component Analysis</b>	305
10.1	Introduction	305
10.2	Sources and Signals	307
10.2.1	Population Independent Components	307
10.2.2	Sample Independent Components	308
10.3	Identification of the Sources	310
10.4	Mutual Information and Gaussianity	314

*Contents*

xi

10.4.1	Independence, Uncorrelatedness and Non-Gaussianity	314
10.4.2	Approximations to the Mutual Information	317
10.5	Estimation of the Mixing Matrix	320
10.5.1	An Estimating Function Approach	321
10.5.2	Properties of Estimating Functions	322
10.6	Non-Gaussianity and Independence in Practice	324
10.6.1	Independent Component Scores and Solutions	324
10.6.2	Independent Component Solutions for Real Data	326
10.6.3	Performance of $\hat{\mathcal{J}}$ for Simulated Data	331
10.7	Low-Dimensional Projections of High-Dimensional Data	335
10.7.1	Dimension Reduction and Independent Component Scores	335
10.7.2	Properties of Low-Dimensional Projections	339
10.8	Dimension Selection with Independent Components	343
<b>11</b>	<b>Projection Pursuit</b>	<b>349</b>
11.1	Introduction	349
11.2	One-Dimensional Projections and Their Indices	350
11.2.1	Population Projection Pursuit	350
11.2.2	Sample Projection Pursuit	356
11.3	Projection Pursuit with Two- and Three-Dimensional Projections	359
11.3.1	Two-Dimensional Indices: $\mathcal{Q}_E$ , $\mathcal{Q}_C$ and $\mathcal{Q}_U$	359
11.3.2	Bivariate Extension by Removal of Structure	361
11.3.3	A Three-Dimensional Cumulant Index	363
11.4	Projection Pursuit in Practice	363
11.4.1	Comparison of Projection Pursuit and Independent Component Analysis	364
11.4.2	From a Cumulant-Based Index to FastICA Scores	365
11.4.3	The Removal of Structure and FastICA	366
11.4.4	Projection Pursuit: A Continuing Pursuit	371
11.5	Theoretical Developments	373
11.5.1	Theory Relating to $\mathcal{Q}_R$	373
11.5.2	Theory Relating to $\mathcal{Q}_U$ and $\mathcal{Q}_D$	374
11.6	Projection Pursuit Density Estimation and Regression	376
11.6.1	Projection Pursuit Density Estimation	376
11.6.2	Projection Pursuit Regression	378
<b>12</b>	<b>Kernel and More Independent Component Methods</b>	<b>381</b>
12.1	Introduction	381
12.2	Kernel Component Analysis	382
12.2.1	Feature Spaces and Kernels	383
12.2.2	Kernel Principal Component Analysis	385
12.2.3	Kernel Canonical Correlation Analysis	389
12.3	Kernel Independent Component Analysis	392
12.3.1	The $\mathcal{F}$ -Correlation and Independence	392
12.3.2	Estimating the $\mathcal{F}$ -Correlation	394

12.3.3	Comparison of Non-Gaussian and Kernel Independent Components Approaches	396
12.4	Independent Components from Scatter Matrices (aka Invariant Coordinate Selection)	402
12.4.1	Scatter Matrices	403
12.4.2	Population Independent Components from Scatter Matrices	404
12.4.3	Sample Independent Components from Scatter Matrices	407
12.5	Non-Parametric Estimation of Independence Criteria	413
12.5.1	A Characteristic Function View of Independence	413
12.5.2	An Entropy Estimator Based on Order Statistics	416
12.5.3	Kernel Density Estimation of the Unmixing Matrix	417
<b>13</b>	<b>Feature Selection and Principal Component Analysis Revisited</b>	<b>421</b>
13.1	Introduction	421
13.2	Independent Components and Feature Selection	423
13.2.1	Feature Selection in Supervised Learning	423
13.2.2	Best Features and Unsupervised Decisions	426
13.2.3	Test of Gaussianity	429
13.3	Variable Ranking and Statistical Learning	431
13.3.1	Variable Ranking with the Canonical Correlation Matrix $C$	432
13.3.2	Prediction with a Selected Number of Principal Components	434
13.3.3	Variable Ranking for Discriminant Analysis Based on $C$	438
13.3.4	Properties of the Ranking Vectors of the Naive $\hat{C}$ when $d$ Grows	442
13.4	Sparse Principal Component Analysis	449
13.4.1	The Lasso, SCoTLASS Directions and Sparse Principal Components	449
13.4.2	Elastic Nets and Sparse Principal Components	453
13.4.3	Rank One Approximations and Sparse Principal Components	458
13.5	(In)Consistency of Principal Components as the Dimension Grows	461
13.5.1	(In)Consistency for Single-Component Models	461
13.5.2	Behaviour of the Sample Eigenvalues, Eigenvectors and Principal Component Scores	465
13.5.3	Towards a General Asymptotic Framework for Principal Component Analysis	471
	<i>Problems for Part III</i>	476
	<i>Bibliography</i>	483
	<i>Author Index</i>	493
	<i>Subject Index</i>	498
	<i>Data Index</i>	503

---

## List of Algorithms

3.1	Partial Least Squares Solution	<i>page</i> 110
4.1	Discriminant Adaptive Nearest Neighbour Rule	153
4.2	Principal Component Discriminant Analysis	159
4.3	Discriminant Analysis with Variable Ranking	162
6.1	Hierarchical Agglomerative Clustering	187
6.2	Mode and Cluster Tracking with the SOPHE	201
6.3	The Gap Statistic	218
8.1	Principal Coordinate Configurations in $p$ Dimensions	253
10.1	Practical Almost Independent Component Solutions	326
11.1	Non-Gaussian Directions from Structure Removal and FastICA	367
11.2	An $M$ -Step Regression Projection Index	378
12.1	Kernel Independent Component Solutions	395
13.1	Independent Component Features in Supervised Learning	424
13.2	Sign Cluster Rule Based on the First Independent Component	427
13.3	An $IC_1$ -Based Test of Gaussianity	429
13.4	Prediction with a Selected Number of Principal Components	434
13.5	Naive Bayes Rule for Ranked Data	447
13.6	Sparse Principal Components Based on the Elastic Net	455
13.7	Sparse Principal Components from Rank One Approximations	459
13.8	Sparse Principal Components Based on Variable Selection	463

Cambridge University Press

978-0-521-88793-9 - Analysis of Multivariate and High-Dimensional Data

Inge Koch

Frontmatter

[More information](#)

---

## Notation

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

$\Gamma_k = [\boldsymbol{\eta}_1 \cdots \boldsymbol{\eta}_k]$	$d \times k$ matrix of (orthogonal) eigenvectors of $\Sigma$ , $k \leq d$
$\Lambda_k = \text{diag}(\lambda_1, \dots, \lambda_k)$	Diagonal $k \times k$ matrix with diagonal entries the eigenvalues of $\Sigma$ , $k \leq d$
$S = \widehat{\Gamma} \widehat{\Lambda} \widehat{\Gamma}^\top$	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
$\phi, \Phi$	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 $L(\theta \mathbb{X})$	Likelihood function of the parameter $\theta$ , given $\mathbb{X}$
$\mathbf{X} \sim (\boldsymbol{\mu}, \Sigma)$	Random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$
$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$	Random vector from the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$
$\mathbb{X} \sim \text{Sam}(\overline{\mathbf{X}}, S)$	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 $\Sigma_{\text{diag}}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ , $S_{\text{diag}}^{-1/2}(\mathbb{X} - \overline{\mathbf{X}})$
$\mathbf{X}^\circ, \mathbb{X}^\circ$	(Spatially) whitened random vector and data
$\mathbf{W}^{(k)} = [W_1 \cdots W_k]^\top$	Vector of first $k$ principal component scores
$\mathbb{W}^{(k)} = [\mathbf{W}_{\bullet 1} \cdots \mathbf{W}_{\bullet k}]^\top$	$k \times n$ 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
$\mathcal{O} = \{O_1, \dots, O_n\}$	Set of objects corresponding to data $\mathbb{X} = [\mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_n]$
$\{\mathcal{O}, \varrho\}$	Observed data, consisting of objects $O_i$ and dissimilarities $\varrho_{ik}$ between pairs of objects
$f, f(\mathbf{X}), f(\mathbb{X})$	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 $\ell = 1, 2$
$C = P \Upsilon Q^\top$	Singular value decomposition of $C$ with singular values $v_j$ and eigenvectors $\mathbf{p}_j, \mathbf{q}_j$
$\widehat{C} = \widehat{P} \widehat{\Upsilon} \widehat{Q}^\top$	Sample canonical correlation matrix and its singular value decomposition
$R^{[C,1]} = C C^\top, R^{[C,2]} = C^\top C$	Matrices of multivariate coefficients of determination, with $C$ the canonical correlation matrix



## Notation

xvii

$\mathbf{U}^{(k)}, \mathbf{V}^{(k)}$	Pair of vectors of $k$ -dimensional canonical correlations
$\boldsymbol{\varphi}_k, \boldsymbol{\psi}_k$	$k$ th pair of canonical (correlation) transforms
$\mathbb{U}^{(k)}, \mathbb{V}^{(k)}$	$k \times n$ matrices of $k$ -dimensional canonical correlation data
$\mathcal{C}_\nu$	$\nu$ th class (or cluster)
$\tau$	Discriminant rule or classifier
$h, h_\beta$	Decision function for a discriminant rule $\tau$ ( $h_\beta$ depends on $\beta$ )
$\mathbf{b}, \mathbf{w}$	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 $B$
$\text{tr}(A)$	Trace of a matrix $A$
$\det(A)$	Determinant of a matrix $A$
$\text{dir}(\mathbf{X})$	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 $[\text{tr}(\boldsymbol{\Sigma})]^{1/2}$
$\ A\ _{\text{Frob}}$	Frobenius norm of a matrix $A$ given by $[\text{tr}(AA^\top)]^{1/2}$
$\Delta(\mathbf{X}, \mathbf{Y})$	Distance between vectors $\mathbf{X}$ and $\mathbf{Y}$
$\varrho(\mathbf{X}, \mathbf{Y})$	Dissimilarity of vectors $\mathbf{X}$ and $\mathbf{Y}$
$\alpha(\boldsymbol{\alpha}, \boldsymbol{\beta})$	Angle between directions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ : $\arccos(\boldsymbol{\alpha}^\top \boldsymbol{\beta})$
$\mathcal{H}, \mathcal{I}, \mathcal{J}, \mathcal{K}$	Entropy, mutual information, negentropy and Kullback-Leibler divergence
$\mathcal{Q}$	Projection index
$\mathbf{n} \succ \mathbf{d}$	Asymptotic domain, $d$ fixed and $n \rightarrow \infty$
$\mathbf{n} \succeq \mathbf{d}$	Asymptotic domain, $d, n \rightarrow \infty$ and $d = O(n)$
$\mathbf{n} \preceq \mathbf{d}$	Asymptotic domain, $d, n \rightarrow \infty$ and $n = o(d)$
$\mathbf{n} \prec \mathbf{d}$	Asymptotic domain, $n$ fixed and $d \rightarrow \infty$

Cambridge University Press

978-0-521-88793-9 - Analysis of Multivariate and High-Dimensional Data

Inge Koch

Frontmatter

[More information](#)

---

---

## 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](http://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

easily recognised form. If the assumptions of a theorem deviate substantially from properties of the data, the reader should recognise that care is required when applying a particular result or method to the data.

In the first part of the book I present proofs of many of the theorems and propositions. For the later parts, the emphasis changes, and in Part III in particular, I typically present theorems without proof and refer the reader to the relevant literature. This is because most of the theorems in Part III refer to recent results. Their proofs can be highly technical and complex without necessarily increasing the understanding of the theorem for the general readership of this book.

Many of the methods I describe do not require knowledge of the underlying distribution. For this reason, my treatment is mostly distribution-free. At times, stronger properties can be derived when we know the distribution of the data. (In practice, this means that the data come from the Gaussian distribution.) In such cases, and in particular in non-asymptotic situations I will explicitly point out what extra mileage knowledge of the Gaussian distribution can gain us. The development of asymptotic theory typically requires the data to come from the Gaussian distribution, and in these theorems I will state the necessary distributional assumptions. Asymptotic theory provides a sound theoretical framework for a method, and in my opinion, it is important to detail the asymptotic results, including the precise conditions under which these results hold.

The formal framework and proofs of theorems should not deter readers who are more interested in an analysis of their data; the theory guides us regarding a method's appropriateness for specific data. However, the methods of this book can be used without dipping into a single proof, and most of the theorems are followed by remarks which explain aspects or implications of the theorems. These remarks can be understood without a deep knowledge of the assumptions and mathematical details of the theorems.

### Teaching from This Book

This book is suitable as a graduate-level textbook and can be taught as a one-semester course with an optional advanced second semester. Graduate students who use this book as their textbook should have a solid general knowledge of statistics, including statistical inference and multivariate regression methods, and a good background in real analysis and linear algebra and, in particular, matrix theory. In addition, the graduate student should be interested in analysing real data and have some experience with MATLAB or R.

Each part of this book ends with a set of problems for the preceding chapters. These problems represent a mixture of theoretical exercises and data analysis and may be suitable as exercises for students.

Some models for a one-semester graduate course are

- A 'classical' course which focuses on the following sections of Chapters 2 to 7
  - Chapter 2, Principal Component Analysis: up to Section 2.7.2
  - Chapter 3, Canonical Correlation Analysis: up to Section 3.7
  - Chapter 4, Discriminant Analysis: up to Section 4.8
  - Chapter 6, Cluster Analysis: up to Section 6.6.2
  - Chapter 7, Factor Analysis: up to Sections 7.6.2 and 7.6.5.

- A mixed ‘classical’ and ‘modern’ course which includes an introduction to Independent Component Analysis, based on Sections 10.1 to 10.3 and some subsections from Section 10.6. Under this model I would focus on a subset of the classical model, and leave out some or all of the following sections

Section 4.7.2 to end of Chapter 4

Section 6.6.

Section 7.6.

Depending on the choice of the first one-semester course, a more advanced one-semester graduate course could focus on the extension sections in Parts I and II and then, depending on the interest of teacher and students, pick and choose material from Chapters 8 and 10 to 13.

### Choice of Topics and How This Book Differs from Other Books on Multivariate Analysis

This book deviates from classical and newer books on multivariate analysis (MVA) in a number of ways.

1. To begin with, I have only a single chapter on background material, which includes summaries of pertinent material on the multivariate normal distribution and relevant results from matrix theory.
2. My emphasis is on the interplay of theory and data. I develop theoretical ideas and apply them to data. In this I differ from many books on MVA, which focus on either data analysis or treat theory.
3. I include analysis of high-dimension low sample size data and describe new and very recent developments in Principal Component Analysis which allow the dimension to grow faster than the sample size. These approaches apply to gene expression data and data from proteomics, for example.

Many of the methods I discuss fall into the category of *component analysis*, including the newer methods in Part III. As a consequence, I made choices to leave out some topics which are often treated in classical multivariate analysis, most notably ANOVA and MANOVA. The former is often taught in undergraduate statistics courses, and the latter is an interesting extension of ANOVA which does not fit so naturally into our framework but is easily accessible to the interested reader.

Another more recent topic, *Support Vector Machines*, is also absent. Again, it does not fit naturally into our component analysis framework. On the other hand, there is a growing number of books which focus exclusively on Support Vector Machines and which remedy my omission. I do, however, make the connection to Support Vector Machines and cite relevant papers and books, in particular in Sections 4.7 and 12.1.

#### And Finally ...

It gives me great pleasure to thank the people who helped while I was working on this monograph. First and foremost, my deepest thanks to my husband, Alun Pope, and sons, Graeme and Reiner Pope, for your love, help, encouragement and belief in me throughout the years it took to write this book. Thanks go to friends, colleagues and students who provided



*Preface*

xxv

valuable feedback. I specially want to mention and thank two friends: Steve Marron, who introduced me to Independent Component Analysis and who continues to inspire me with his ideas, and Kanta Naito, who loves to argue with me until we find *good* answers to the problems we are working on. Finally, I thank Diana Gillooly from Cambridge University Press for her encouragement and help throughout this journey.