

I.T. Jolliffe

Principal Component Analysis

Second Edition

With 28 Illustrations



Springer

Contents

Preface to the Second Edition	v
Preface to the First Edition	ix
Acknowledgments	xv
List of Figures	xxiii
List of Tables	xxvii
1 Introduction	1
1.1 Definition and Derivation of Principal Components	1
1.2 A Brief History of Principal Component Analysis	6
2 Properties of Population Principal Components	10
2.1 Optimal Algebraic Properties of Population Principal Components	11
2.2 Geometric Properties of Population Principal Components	18
2.3 Principal Components Using a Correlation Matrix	21
2.4 Principal Components with Equal and/or Zero Variances	27
3 Properties of Sample Principal Components	29
3.1 Optimal Algebraic Properties of Sample Principal Components	30
3.2 Geometric Properties of Sample Principal Components .	33
3.3 Covariance and Correlation Matrices: An Example . . .	39
3.4 Principal Components with Equal and/or Zero Variances	43

3.4.1	Example	43
3.5	The Singular Value Decomposition	44
3.6	Probability Distributions for Sample Principal Components	47
3.7	Inference Based on Sample Principal Components	49
3.7.1	Point Estimation	50
3.7.2	Interval Estimation	51
3.7.3	Hypothesis Testing	53
3.8	Patterned Covariance and Correlation Matrices	56
3.8.1	Example	57
3.9	Models for Principal Component Analysis	59
4	Interpreting Principal Components: Examples	63
4.1	Anatomical Measurements	64
4.2	The Elderly at Home	68
4.3	Spatial and Temporal Variation in Atmospheric Science	71
4.4	Properties of Chemical Compounds	74
4.5	Stock Market Prices	76
5	Graphical Representation of Data Using Principal Components	78
5.1	Plotting Two or Three Principal Components	80
5.1.1	Examples	80
5.2	Principal Coordinate Analysis	85
5.3	Biplots	90
5.3.1	Examples	96
5.3.2	Variations on the Biplot	101
5.4	Correspondence Analysis	103
5.4.1	Example	105
5.5	Comparisons Between Principal Components and other Methods	106
5.6	Displaying Intrinsically High-Dimensional Data	107
5.6.1	Example	108
6	Choosing a Subset of Principal Components or Variables	111
6.1	How Many Principal Components?	112
6.1.1	Cumulative Percentage of Total Variation	112
6.1.2	Size of Variances of Principal Components	114
6.1.3	The Scree Graph and the Log-Eigenvalue Diagram	115
6.1.4	The Number of Components with Unequal Eigenvalues and Other Hypothesis Testing Procedures	118
6.1.5	Choice of m Using Cross-Validatory or Computationally Intensive Methods	120
6.1.6	Partial Correlation	127
6.1.7	Rules for an Atmospheric Science Context	127
6.1.8	Discussion	130

6.2	Choosing m , the Number of Components: Examples . . .	133
6.2.1	Clinical Trials Blood Chemistry	133
6.2.2	Gas Chromatography Data	134
6.3	Selecting a Subset of Variables	137
6.4	Examples Illustrating Variable Selection	145
6.4.1	<i>Alate adelges</i> (Winged Aphids)	145
6.4.2	Crime Rates	147
7	Principal Component Analysis and Factor Analysis	150
7.1	Models for Factor Analysis	151
7.2	Estimation of the Factor Model	152
7.3	Comparisons Between Factor and Principal Component Analysis	158
7.4	An Example of Factor Analysis	161
7.5	Concluding Remarks	165
8	Principal Components in Regression Analysis	167
8.1	Principal Component Regression	168
8.2	Selecting Components in Principal Component Regression	173
8.3	Connections Between PC Regression and Other Methods	177
8.4	Variations on Principal Component Regression	179
8.5	Variable Selection in Regression Using Principal Compo- nents	185
8.6	Functional and Structural Relationships	188
8.7	Examples of Principal Components in Regression	190
8.7.1	Pitprop Data	190
8.7.2	Household Formation Data	195
9	Principal Components Used with Other Multivariate Techniques	199
9.1	Discriminant Analysis	200
9.2	Cluster Analysis	210
9.2.1	Examples	214
9.2.2	Projection Pursuit	219
9.2.3	Mixture Models	221
9.3	Canonical Correlation Analysis and Related Techniques .	222
9.3.1	Canonical Correlation Analysis	222
9.3.2	Example of CCA	224
9.3.3	Maximum Covariance Analysis (SVD Analysis), Redundancy Analysis and Principal Predictors . .	225
9.3.4	Other Techniques for Relating Two Sets of Variables	228

10	Outlier Detection, Influential Observations and Robust Estimation	232
10.1	Detection of Outliers Using Principal Components	233
10.1.1	Examples	242
10.2	Influential Observations in a Principal Component Analysis	248
10.2.1	Examples	254
10.3	Sensitivity and Stability	259
10.4	Robust Estimation of Principal Components	263
10.5	Concluding Remarks	268
11	Rotation and Interpretation of Principal Components	269
11.1	Rotation of Principal Components	270
11.1.1	Examples	274
11.1.2	One-step Procedures Using Simplicity Criteria . .	277
11.2	Alternatives to Rotation	279
11.2.1	Components with Discrete-Valued Coefficients . .	284
11.2.2	Components Based on the LASSO	286
11.2.3	Empirical Orthogonal Teleconnections	289
11.2.4	Some Comparisons	290
11.3	Simplified Approximations to Principal Components . .	292
11.3.1	Principal Components with Homogeneous, Contrast and Sparsity Constraints	295
11.4	Physical Interpretation of Principal Components	296
12	PCA for Time Series and Other Non-Independent Data	299
12.1	Introduction	299
12.2	PCA and Atmospheric Time Series	302
12.2.1	Singular Spectrum Analysis (SSA)	303
12.2.2	Principal Oscillation Pattern (POP) Analysis . .	308
12.2.3	Hilbert (Complex) EOFs	309
12.2.4	Multitaper Frequency Domain-Singular Value Decomposition (MTM SVD)	311
12.2.5	Cyclo-Stationary and Periodically Extended EOFs (and POPs)	314
12.2.6	Examples and Comparisons	316
12.3	Functional PCA	316
12.3.1	The Basics of Functional PCA (FPCA)	317
12.3.2	Calculating Functional PCs (FPCs)	318
12.3.3	Example - 100 km Running Data	320
12.3.4	Further Topics in FPCA	323
12.4	PCA and Non-Independent Data—Some Additional Topics	328
12.4.1	PCA in the Frequency Domain	328
12.4.2	Growth Curves and Longitudinal Data	330
12.4.3	Climate Change—Fingerprint Techniques	332
12.4.4	Spatial Data	333
12.4.5	Other Aspects of Non-Independent Data and PCA	335

13 Principal Component Analysis for Special Types of Data	338
13.1 Principal Component Analysis for Discrete Data	339
13.2 Analysis of Size and Shape	343
13.3 Principal Component Analysis for Compositional Data	346
13.3.1 Example: 100 km Running Data	349
13.4 Principal Component Analysis in Designed Experiments	351
13.5 Common Principal Components	354
13.6 Principal Component Analysis in the Presence of Missing Data	363
13.7 PCA in Statistical Process Control	366
13.8 Some Other Types of Data	369
14 Generalizations and Adaptations of Principal Component Analysis	373
14.1 Non-Linear Extensions of Principal Component Analysis	374
14.1.1 Non-Linear Multivariate Data Analysis—Gifi and Related Approaches	374
14.1.2 Additive Principal Components and Principal Curves	377
14.1.3 Non-Linearity Using Neural Networks	379
14.1.4 Other Aspects of Non-Linearity	381
14.2 Weights, Metrics, Transformations and Centerings	382
14.2.1 Weights	382
14.2.2 Metrics	386
14.2.3 Transformations and Centering	388
14.3 PCs in the Presence of Secondary or Instrumental Variables	392
14.4 PCA for Non-Normal Distributions	394
14.4.1 Independent Component Analysis	395
14.5 Three-Mode, Multiway and Multiple Group PCA	397
14.6 Miscellanea	400
14.6.1 Principal Components and Neural Networks	400
14.6.2 Principal Components for Goodness-of-Fit Statistics	401
14.6.3 Regression Components, Sweep-out Components and Extended Components	403
14.6.4 Subjective Principal Components	404
14.7 Concluding Remarks	405
A Computation of Principal Components	407
A.1 Numerical Calculation of Principal Components	408
Index	458
Author Index	478