Matrix-Based Introduction to Multivariate Data Analysis

Kohei Adachi

Matrix-Based Introduction to Multivariate Data Analysis

Kohei Adachi
Graduate School of Human Sciences
Osaka University
Osaka
Japan

ISBN 978-981-10-2340-8 ISBN 978-981-10-2341-5 (eBook)
DOI 10.1007/978-981-10-2341-5

Library of Congress Control Number: 2016948805

© Springer Nature Singapore Pte Ltd. 2016, corrected publication 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

A set of multivariate data can be expressed as a table, i.e., a matrix, of individuals (rows) by variables (columns), with the variables interrelated. Statistical procedures for analyzing such data sets are generally referred to as multivariate data analysis. The demand for this kind of analysis is increasing in a variety of fields. Each procedure in multivariate data analysis features a special purpose. For example, predicting future performance, classifying individuals, visualizing inter-individual relationships, finding a few factors underlying a number of variables, and examining causal relationships among variables are included in the purposes for the procedures.

The aim of this book is to enable readers who may not be familiar with matrix operations to understand major multivariate data analysis procedures in matrix forms. For that aim, this book begins with explaining fundamental matrix calculations and the matrix expressions of elementary statistics, followed by an introduction to popular multivariate procedures, with chapter-by-chapter advances in the levels of matrix algebra. The organization of this book allows readers without knowledge of matrices to deepen their understanding of multivariate data analysis.

Another feature of this book is its emphasis on the model that underlies each procedure and the objective function that is optimized for fitting the model to data. The author believes that the matrix-based learning of such models and objective functions is the shortest way to comprehend multivariate data analysis. This book is also arranged so that readers can intuitively capture for what purposes multivariate analysis procedures are utilized; plain explanations of the purposes with numerical examples precede mathematical descriptions in almost all chapters.

The preceding paragraph featured three key words: purpose, model, and objective function. The author considers that capturing those three points for each procedure suffices to understand it. This consideration implies that the mechanisms behind how objective functions are optimized must not necessarily be understood. Thus, the mechanisms are only described in appendices and some exercises.

This book is written with the following guidelines in mind:

- (1) Not using mathematics except matrix algebra
- (2) Emphasizing singular value decomposition (SVD)
- (3) Preferring a simultaneous solution to a successive one

Although the exceptions to (1) are found in Appendix A6, where differential calculus is used, and in some sections of Part III and Chap. 15, where probabilities are used, those exceptional parts only occupy a limited number of pages; the majority of the book is matrix-intensive. Matrix algebra is also exclusively used for formulating the optimization of objective functions in Appendix A4. For matrix-intensive formulations, ten Berge's (1983, 1993) theorem is considered to be the best starting fact, as found in Appendix A4.1.

Guideline (2) is due to the fact that SVD can be defined for any matrix, and a number of important properties of matrices are easily derived from SVD. In the former point, SVD is more general than eigenvalue decomposition (EVD), which is only defined for symmetric matrices. Thus, EVD is only mentioned in Sect. 6.2. Further, SVD takes on an important role in optimizing trace and least squares functions of matrices: The optimization problems are formulated with the combination of SVD and ten Berge's (1983, 1993) theorem, as found in Appendix A4.2 and Appendix A4.3.

Guideline (3) is particularly concerned with principal component analysis (PCA), which can be formulated as minimizing $\|\mathbf{X} - \mathbf{FA}'\|^2$ over PC score matrix $\mathbf{F}$ and loading matrix $\mathbf{A}$ for a data matrix $\mathbf{X}$. In some of the literature, PCA is described as obtaining the first component, the second, and the remaining components in turn (i.e., per column of $\mathbf{F}$ and $\mathbf{A}$). This can be called a successive solution. On the other hand, PCA can be described as obtaining $\mathbf{F}$ and $\mathbf{A}$ matrix-wise, which can be called a simultaneous solution. This is preferred in this book, as the above formulation is actually made matrix-wise and the simultaneous solution facilitates understanding PCA as a reduced rank approximation of $\mathbf{X}$.

This book is appropriate for undergraduate students who have already learned introductory statistics, as the author has used preliminary versions of the book in a course for such students. It is also useful for graduate students and researchers who are not familiar with the matrix-intensive formulations of multivariate data analysis.

I owe this book to the people who can be called the "matricians" in statistics, more exactly, the ones taking matrix-intensive approaches for formulating and developing data analysis procedures. Particularly, I have been influenced by the Dutch psychometricians, as found above, in that I emphasize the theorem by Jos M.F. ten Berge (Professor Emeritus, University of Groningen). Yutaka Hirachi of Springer has been encouraging me since I first considered writing this book. I am most grateful to him. I am also thankful to the reviewers who read through drafts of this book. Finally, I must show my gratitude to Yoshitaka Shishikura of the publisher Nakanishiya Shuppan, as he readily agreed to the use of the numerical examples in this book, which I had originally used in that publisher's book.

*The original version of the book was revised:
Belated corrections have been incorporated.
The erratum to the book is available at
https://doi.org/10.1007/978-981-10-2341-5_17*

Contents

Part I Elementary Statistics with Matrices

1	Elementary Matrix Operations	3
1.1	Matrices	3
1.2	Vectors	5
1.3	Sum of Matrices and Their Multiplication by Scalars	6
1.4	Inner Product and Norms of Vectors	7
1.5	Product of Matrices	8
1.6	Two Properties of Matrix Products	11
1.7	Trace Operator and Matrix Norm	11
1.8	Vectors and Matrices Filled with Ones or Zeros	13
1.9	Special Square Matrices	14
1.10	Bibliographical Notes	15
2	Intravariab le Statistics	17
2.1	Data Matrices	17
2.2	Distributions	19
2.3	Averages	20
2.4	Centered Scores	20
2.5	Variance and Standard Deviation	23
2.6	Standard Scores	25
2.7	What Centering and Standardization Do for Distributions	25
2.8	Matrix Representation	26
2.9	Bibliographical Notes	28
3	Inter-variable Statistics	29
3.1	Scatter Plots and Correlations	29
3.2	Covariance	30
3.3	Correlation Coefficient	32
3.4	Variable Vectors and Correlations	34
3.5	Covariances and Correlations for Standard Scores	35
3.6	Matrix Expressions of Covariances and Correlations	36

3.7	Unbiased Covariances	37
3.8	Centered Matrix	38
3.9	Ranks of Matrices: Intuitive Introduction	39
3.10	Ranks of Matrices: Mathematical Definition	40
3.11	Bibliographical Notes	42

Part II Least Squares Procedures

4	Regression Analysis	47
4.1	Prediction of a Dependent Variable by Explanatory Variables	47
4.2	Least Squares Method	50
4.3	Predicted and Error Values	52
4.4	Proportion of Explained Variance and Multiple Correlation	54
4.5	Interpretation of Regression Coefficients	56
4.6	Standardization	57
4.7	Geometric Derivation of Regression Coefficients	59
4.8	Bibliographical Notes	60
5	Principal Component Analysis (Part 1)	63
5.1	Reduction of Variables into Components	63
5.2	Singular Value Decomposition	65
5.3	Formulation with a Weight Matrix	67
5.4	Constraints for Components	68
5.5	Interpretation of Loadings	70
5.6	Interpretation of Weights	71
5.7	Percentage of Explained Variance	72
5.8	High-Dimensional Data Analysis	73
5.9	Bibliographical Notes	75
6	Principal Component Analysis (Part 2)	79
6.1	Reformulation with Different Constraints	79
6.2	Maximizing the Sum of Variances	80
6.3	Weighted Composite Scores with Maximum Variance	82
6.4	Projecting Three-Dimensional Vectors Onto Two-Dimensional Ones	83
6.5	Visualization of Invisible Distributions	86
6.6	Goodness of Projection	89
6.7	Bibliographical Notes	90
7	Cluster Analysis	93
7.1	Membership Matrices	93
7.2	Example of Clustering Results	94
7.3	Formulation	96
7.4	Iterative Algorithm	98

7.5	Obtaining Cluster Features	100
7.6	Obtaining Memberships	101
7.7	Brief Description of Algorithm	102
7.8	Bibliographical Notes	103

Part III Maximum Likelihood Procedures

8	Maximum Likelihood and Multivariate Normal Distribution	109
8.1	Model, Parameter, Objective Function, and Optimization	109
8.2	Maximum Likelihood Method	110
8.3	Probability Density Function.	112
8.4	Multivariate Normal Distribution	114
8.5	Maximum Likelihood Method for Normal Variables	117
8.6	Maximum Likelihood Estimates of Means and Covariances	118
8.7	Model Selection	120
8.8	Assessment of Between-Group Heterogeneity.	121
8.9	Bibliographical Notes	123
9	Path Analysis	127
9.1	From Multiple Regression Analysis to Path Analysis	127
9.2	Matrix Expression.	132
9.3	Distributional Assumptions	133
9.4	Likelihood for Covariance Structure Analysis.	134
9.5	Maximum Likelihood Estimation	135
9.6	Estimated Covariance Structure.	136
9.7	Unstandardized and Standardized Solutions	137
9.8	Other and Extreme Models	139
9.9	Model Selection	141
9.10	Bibliographical Notes	141
10	Confirmatory Factor Analysis	145
10.1	Example of Confirmatory Factor Analysis Model.	145
10.2	Matrix Expression.	146
10.3	Distributional Assumptions for Factors.	152
10.4	Distributional Assumptions for Errors.	153
10.5	Maximum Likelihood Method	153
10.6	Solutions.	154
10.7	Other and Extreme Models	156
10.8	Model Selection	157
10.9	Bibliographical Notes	158
11	Structural Equation Modeling	161
11.1	Causality Among Factors	161
11.2	Observed Variables as Indicator of Factors.	162
11.3	SEM Model	165

11.4	Matrix Expression	166
11.5	Distributional Assumptions	167
11.6	Maximum Likelihood Method	168
11.7	Solutions	169
11.8	Model Selection	171
11.9	Bibliographical Notes	172
12	Exploratory Factor Analysis	175
12.1	Example of Exploratory Factor Analysis Model	175
12.2	Matrix Expression	176
12.3	Distributional Assumptions	177
12.4	Maximum Likelihood Method	178
12.5	Indeterminacy of EFA Solutions	179
12.6	Two-Stage Procedure	180
12.7	Interpretation of Loadings	180
12.8	Interpretation of Unique Variances	181
12.9	Selecting the Number of Factors	182
12.10	Difference to Principal Component Analysis	184
12.11	Bibliographical Notes	186
 Part IV Miscellaneous Procedures		
13	Rotation Techniques	193
13.1	Geometric Illustration of Factor Rotation	193
13.2	Oblique and Orthogonal Rotation	196
13.3	Rotation to Simple Structure	197
13.4	Varimax Rotation	199
13.5	Geomin Rotation	200
13.6	Orthogonal Procrustes Rotation	202
13.7	Bibliographical Notes	203
14	Canonical Correlation and Multiple Correspondence Analyses	207
14.1	Block Matrices	207
14.2	Canonical Correlation Analysis	210
14.3	Generalized Canonical Correlation Analysis	212
14.4	Multivariate Categorical Data	216
14.5	Multiple Correspondence Analysis	217
14.6	Homogeneity Assumption	219
14.7	Bibliographical Notes	221
15	Discriminant Analysis	225
15.1	Modification of Multiple Correspondence Analysis	225
15.2	Canonical Discriminant Analysis	227
15.3	Minimum Distance Classification	229
15.4	Maximum Probability Classification	230
15.5	Normal Discrimination for Two Groups	231

15.6	Interpreting Solutions	233
15.7	Generalized Normal Discrimination	236
15.8	Bibliographical Notes	239
16	Multidimensional Scaling	243
16.1	Linking Coordinates to Quasi-Distances	243
16.2	Illustration of an MDS Solution	245
16.3	Iterative Algorithm	246
16.4	Matrix Expression for Squared Distances	247
16.5	Inequality for Distances	249
16.6	Majorization Algorithm.	251
16.7	Bibliographical Notes	252
Erratum to: Matrix-Based Introduction to Multivariate		
Data Analysis		E1
Appendix		255
References		291
Index		297