Claudiu VINȚE **Titus Felix FURTUNĂ**

Multivariate Data Analysis in Python

Colecția
Informatică

Editura ASE
București
2023

Academia de Studii Economice din București

Editura ASE
Piața Romană nr. 6, sector 1, București, România
cod 010374
www.ase.ro
www.editura.ase.ro
editura@ase.ro

Descrierea CIP a Bibliotecii Naţionale a României
VINŢE, CLAUDIU
Multivariate data analysis in Python / Claudiu Vinţe, Titus Felix Furtună. - București : Editura ASE, 2023
Conţine bibliografie
ISBN 978-606-34-0500-6

I. Furtună, Felix

004

Editura ASE
Redactor: Luiza Constantinescu
Tehnoredactor: Violeta Rogojan
Coperta: Violeta Rogojan

Autorii își asumă întreaga responsabilitate pentru: ideile exprimate, corectitudinea științifică, originalitatea materialului și sursele bibliografice menționate.

Foreword

Motto: *Data does not lie, but all the same will not provide information that we are not ready to grasp.*

This book is primarily addressed to students of the Department of Economic Informatics and Cybernetics, the Economic Informatics undergraduate programme in the English language, within the Faculty of Cybernetics, Statistics, and Economic Informatics of the Bucharest University of Economic Studies. This work covers the syllabus of the discipline titled *Software Development for Data Analysis*. It may also be useful for master's and doctoral students, researchers, and practitioners interested in methods for multivariate data analysis and their implementation in a programming language. In this sense, it can also be an introduction to data science through Python programming.

The goal is to complete the applicative component of the learning act in and to offer through this work hands-on support for the activity carried out in the laboratories, but also for the individual study of the multivariate data analysis discipline.

The work is structured into six chapters. Each chapter treats a multivariate data analysis method. In addition, carefully crafted to be readily accessible to an enthusiastic programmer, Python code implementations accompany the multivariate data analysis methods presented throughout the book.

The first chapter presents the principal component analysis (PCA). The process of determining the principal components is carried out in observation-driven and variable-driven approaches. The criteria for choosing the number of axes are discussed, along with the evaluation indicators. The second chapter deals with exploratory factor analysis (EFA). The third chapter examines canonical correlation analysis (CCA). The fourth chapter discusses factorial correspondence analysis (FCA) with an example regarding multiple correspondence analysis (MCA). Chapter five presents hierarchical cluster analysis (HCA) as an unsupervised classification method, and the sixth chapter is dedicated to discriminant analysis (DA) as one of the fundamental supervised classification methods in multivariate data analysis.

For each method addressed, a theoretical preamble is provided which introduces the fundamental results in the field of the discussed multivariate data analysis method and formalises the necessary concepts. In addition, within each analysis method, we present the methodological elements required for its implementation in Python. The presentation of each method is accompanied by a case study that represents a practical example of Python

implementation and processing, using the proposed IT solution, of one or more sets of input data. In each case study, the results obtained are discussed and interpreted. The Python code is written and tested with Python interpreter version 3.11 together with the corresponding compatible versions of the packages employed, taking December 1, 2023, as a reference.

We aim to provide those interested in multivariate data analysis methods and their implementation in one of today's most popular programming languages with an application-orientated material with immediate utility that opens the way to new applications or research activities from adjacent fields, such as pattern recognition, data mining, machine learning, etc.

In conclusion, the presented content underlines the idea of supporting and encouraging a learning way that stimulates the conscious and coherent construction of Python solutions for processing multivariate datasets. In this context, the bibliography we include at the end of the book is more likely to point readers to various useful sources to deepen the knowledge introduced through the practical examples presented in this work.

The authors are grateful to anonymous reviewers who, through their recommendations, helped us achieve better-organised content for the book.

We thank Editura ASE, the editorial team, and the printing team, who made this book possible through their professionalism and dedication.

Authors

Table of Contents

1. Principal Component Analysis (PCA) 11

1.1 Data to be Processed and Stages of the Analysis 11

1.1.1 Observation-driven Approach 12

1.1.2 Variable-driven Approach 16

1.2 Rationale – Criteria for Choosing the Number of Axes 18

1.2.1 Coverage Percentage Criterion 18

1.2.2 Kaiser's Criterion 19

1.2.3 Cattell's Criterion 19

1.3 Results Evaluation 20

1.3.1 Scores 20

1.3.2 Quality of Observations Representation 20

1.3.3 Contribution of Individuals to the Variance of Axes 21

1.3.4 Communality 21

1.3.5 Factor Loadings 21

1.4 Graphical Representations 22

1.4.1 Representation of Individuals (Observations) 22

1.4.2 Representation of Variables 22

1.5 Additional Variables and Observations 23

1.6 Reconstitution of Initial Data 23

1.6.1 Reconstitution of the Initial Data Table 23

1.6.2 Reconstitution of the Correlation Matrix 24

1.7 Unstandardised Principal Component Analysis 24

1.8 Weighted Principal Component Analysis 25

1.9 PCA – Case Study 25

1.9.1 Data Set 26

1.9.2 Python Implementation 27

1.9.3 Results and Discussion 37

2. Exploratory Factor Analysis (EFA) 47

2.1 Data to be Processed 47

2.2 Model Assumptions 48

2.3 Estimation of Factors Existence 49

2.4 Estimation of Model Parameters. Factor Extraction 50

2.5 Method of Maximum Likelihood 51

2.6 Method of Least-squares 51

2.7 Method of Principal Components 52

2.8 Estimate the Number of Factors. Bartlett's Test 53

2.9 Factor Rotation 54

2.10 EFA – Case Study 55

2.10.1 Data Set 55

2.10.2 Python Implementation 57

2.10.3 Results and Discussion 64

3. Canonical Correlation Analysis (CCA) 69
3.1 Data to be Processed 69
3.2 Stages of Analysis 70
3.3 Canonical Factors 72
3.4 Relations Between Factors 74
3.5 Explained Variance and Informational Redundancy 75
3.6 Standardisation of Canonical Factors 76
3.7 Relevance of Canonical Roots. Bartlett's Chi-square Test 76
3.8 Generalised Canonical Correlation Analysis (GCCA) 77
3.9 CCA – Case Study 79
3.9.1 Data Sets 79
3.9.2 Python Implementation 82
3.9.3 Results and Discussion 87
4. Factorial Correspondence Analysis (FCA) 91
4.1 Canonical Analysis of Complete Disjunctive Tables 91
4.2 Principal Component Analysis of Profile Tables 94
4.3 Principal Component Analysis of Inertia Matrix 96
4.4 Multiple Correspondences Analysis (MCA) 98
4.5 Principal Component Analysis of Standard Deviation Matrix 100
4.6 MCA – Case Study 101
4.6.1 Data Set 101
4.6.2 Python Implementation 106
4.6.3 Results and Discussion 111
5. Hierarchical Cluster Analysis (HCA) 115
5.1 Hierarchical Algorithms 116
5.2 Hierarchical Grouping Methods 117
5.3 Distances for Proximity Calculation 120
5.4 Choosing the Optimal Number of Clusters 121
5.5 HCA – Case Study 124
5.5.1 Data Set 124
5.5.2 Python Implementation 126
5.5.3 Results and Discussion 132
6. Discriminant Analysis (DA) 138
6.1 Data Organisation and Notation 138
6.2 Indicators of Variability and Dispersion 139
6.3 Model Meaning. Statistical Tests 141
6.4 Linear Discriminant Analysis (LDA) – Fisher Functions 143
6.5 Discriminant Analysis – Special Case of Canonical Correlation Analysis 146
6.6 Classification Functions 147
6.7 Bayesian Discrimination 149
6.7.1 Total Probability and Bayesian Probability 149
6.7.2 Bayesian Classifier 149
6.7.3 Nonparametric Estimation of Conditional Probabilities. Histogram Method 150
6.7.4 Parametric Methods 151
6.8 Discriminat Analysis – Case Study 152
6.8.1 Data Set 152
6.8.2 Python Implementation 153
6.8.3 Results and Discussion 161
Bibliography 167
Appendix 1 169
Appendix 2 176