Applied Multivariate Statistical Analysis

Wolfgang Karl Härdle • Léopold Simar

Applied Multivariate Statistical Analysis

Fourth Edition

Springer

Wolfgang Karl Härdle
C.A.S.E. Centre f. Appl. Stat. & Econ.
School of Business and Economics
Humboldt-Universität zu Berlin
Berlin, Germany

Léopold Simar
Center of Operations Research &
Econometrics (CORE)
Katholieke Univeristeit Leuven Inst.
Statistics
Leuven, Belgium

The majority of chapters have quantlet codes in Matlab or R. These quantlets may be downloaded from http://extras.springer.com or via a link on http://springer.com/978-3-662-45170-0 and from www.quantlet.de

ISBN 978-3-662-45170-0 ISBN 978-3-662-45171-7 (eBook)
DOI 10.1007/978-3-662-45171-7

Library of Congress Control Number: 2015933294

Mathematics Subject Classification (2000): 62H10, 62H12, 62H15, 62H17, 62H20, 62H25, 62H30, 62F25

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media (www.springer.com)

Preface to the Fourth Edition

The fourth edition of this book on *Applied Multivariate Statistical Analysis* offers a new sub-chapter on Variable Selection by using least absolute shrinkage and selection operator (LASSO) and its general form the so-called Elastic Net.

All pictures and numerical examples have been now calculated in the (almost) standard language R & MATLAB. The code for each picture is indicated with a small Q sign near the picture, e.g. Q MVAdenbank denotes the corresponding quantlet for reproduction of Fig. 1.9, where we display the densities of the diagonal of genuine and counterfeit bank notes. We believe that these publicly available quantlets (see also http://sfb649.wiwi.hu-berlin.de/quantnet/) create a valuable contribution to distribution of knowledge in the statistical science. The symbols and notations have also been standardised. In the preparation of the fourth edition, we received valuable input from Dedy Dwi Prastyo, Petra Burdejova, Sergey Nasekin and Awdesch Melzer. We would like to thank them.

Berlin, Germany	Wolfgang Karl Härdle
Louvain la Neuve, Belgium	Léopold Simar
January 2014	

Preface to the Third Edition

The third edition of this book on *Applied Multivariate Statistical Analysis* offers the following new features.

1. A new Chap. 8 on Regression Models has been added.
2. Almost all numerical examples have been reproduced in MATLAB or R.

The chapter on regression models focuses on a core business of multivariate statistical analysis. This contribution has not been subject of a prominent discussion in earlier editions of this book. We now take the opportunity to cover classical themes of ANOVA and ANCOVA analysis. Categorical responses are presented in Sect. 8.2. The spectrum of log linear models for contingency tables is presented in Sect. 8.2.2, and applications to count data, e.g. in the economic and medical science are presented there. Logit models are discussed in great detail, and the numerical implementation in terms of matrix manipulations is presented.

The majority of pictures and numerical examples has been now calculated in the (almost) standard language R & MATLAB. The code for each picture is indicated with a small Q sign near the picture, e.g. Q MVAdenbank denotes the corresponding quantlet for reproduction of Fig. 1.9, where we display the densities of the diagonal of genuine and counterfeit bank notes. We believe that these publicly available quantlets (see also www.quantlet.com) create a valuable contribution to distribution of knowledge in the statistical science. The symbols and notations have also been standardised. In the preparation of the third edition, we received valuable input from Song Song, Weining Wang and Mengmeng Guo. We would like to thank them.

Berlin, Germany
Louvain la Neuve, Belgium
June 2011

Wolfgang Karl Härdle
Léopold Simar

Contents

Part I Descriptive Techniques

1 Comparison of Batches 3
1.1 Boxplots 4
1.2 Histograms 11
1.3 Kernel Densities 15
1.4 Scatterplots 19
1.5 Chernoff-Flury Faces 22
1.6 Andrews' Curves 29
1.7 Parallel Coordinates Plots 32
1.8 Hexagon Plots 37
1.9 Boston Housing 40
1.10 Exercises 48

Part II Multivariate Random Variables

2 A Short Excursion into Matrix Algebra 53
2.1 Elementary Operations 53
2.2 Spectral Decompositions 60
2.3 Quadratic Forms 62
2.4 Derivatives 65
2.5 Partitioned Matrices 66
2.6 Geometrical Aspects 68
2.7 Exercises 76

3 Moving to Higher Dimensions 79
3.1 Covariance 80
3.2 Correlation 84
3.3 Summary Statistics 89
3.4 Linear Model for Two Variables 93
3.5 Simple Analysis of Variance 100

3.6 Multiple Linear Model ... 105
3.7 Boston Housing ... 110
3.8 Exercises ... 113

4 Multivariate Distributions ... 117
4.1 Distribution and Density Function ... 118
4.2 Moments and Characteristic Functions ... 123
4.3 Transformations ... 135
4.4 The Multinormal Distribution ... 137
4.5 Sampling Distributions and Limit Theorems ... 142
4.6 Heavy-Tailed Distributions ... 149
4.7 Copulae ... 166
4.8 Bootstrap ... 176
4.9 Exercises ... 179

5 Theory of the Multinormal ... 183
5.1 Elementary Properties of the Multinormal ... 183
5.2 The Wishart Distribution ... 191
5.3 Hotelling's T^2-Distribution ... 193
5.4 Spherical and Elliptical Distributions ... 195
5.5 Exercises ... 197

6 Theory of Estimation ... 201
6.1 The Likelihood Function ... 202
6.2 The Cramer–Rao Lower Bound ... 206
6.3 Exercises ... 210

7 Hypothesis Testing ... 213
7.1 Likelihood Ratio Test ... 214
7.2 Linear Hypothesis ... 224
7.3 Boston Housing ... 242
7.4 Exercises ... 246

Part III Multivariate Techniques

8 Regression Models ... 253
8.1 General ANOVA and ANCOVA Models ... 255
8.1.1 ANOVA Models ... 255
8.1.2 ANCOVA Models ... 260
8.1.3 Boston Housing ... 262
8.2 Categorical Responses ... 263
8.2.1 Multinomial Sampling and Contingency Tables ... 263
8.2.2 Log-Linear Models for Contingency Tables ... 264
8.2.3 Testing Issues with Count Data ... 268
8.2.4 Logit Models ... 271
8.3 Exercises ... 279

9 Variable Selection 281
9.1 Lasso.......... 282
9.1.1 Lasso in the Linear Regression Model 282
9.1.2 Lasso in High Dimensions 292
9.1.3 Lasso in Logit Model.......... 293
9.2 Elastic Net.......... 297
9.2.1 Elastic Net in Linear Regression Model.......... 298
9.2.2 Elastic Net in Logit Model 299
9.3 Group Lasso.......... 300
9.4 Exercises 304

10 Decomposition of Data Matrices by Factors 305
10.1 The Geometric Point of View 306
10.2 Fitting the p-Dimensional Point Cloud.......... 307
10.3 Fitting the n-Dimensional Point Cloud 310
10.4 Relations Between Subspaces.......... 312
10.5 Practical Computation 314
10.6 Exercises 317

11 Principal Components Analysis 319
11.1 Standardised Linear Combination.......... 320
11.2 Principal Components in Practice.......... 324
11.3 Interpretation of the PCs.......... 327
11.4 Asymptotic Properties of the PCs 331
11.5 Normalised Principal Components Analysis 335
11.6 Principal Components as a Factorial Method.......... 336
11.7 Common Principal Components 342
11.8 Boston Housing 346
11.9 More Examples 348
11.10 Exercises 357

12 Factor Analysis 359
12.1 The Orthogonal Factor Model.......... 360
12.2 Estimation of the Factor Model 367
12.3 Factor Scores and Strategies.......... 376
12.4 Boston Housing 378
12.5 Exercises 382

13 Cluster Analysis 385
13.1 The Problem.......... 386
13.2 The Proximity Between Objects 387
13.3 Cluster Algorithms.......... 392
13.4 Boston Housing 400
13.5 Exercises 404

14 Discriminant Analysis 407
14.1 Allocation Rules for Known Distributions 407
14.2 Discrimination Rules in Practice 415
14.3 Boston Housing 421
14.4 Exercises 423

15 Correspondence Analysis 425
15.1 Motivation 426
15.2 Chi-Square Decomposition 428
15.3 Correspondence Analysis in Practice 432
15.4 Exercises 441

16 Canonical Correlation Analysis 443
16.1 Most Interesting Linear Combination 443
16.2 Canonical Correlation in Practice 448
16.3 Exercises 454

17 Multidimensional Scaling 455
17.1 The Problem 455
17.2 Metric MDS 460
17.3 Nonmetric MDS 465
17.4 Exercises 472

18 Conjoint Measurement Analysis 473
18.1 Introduction 473
18.2 Design of Data Generation 475
18.3 Estimation of Preference Orderings 478
18.4 Exercises 485

19 Applications in Finance 487
19.1 Portfolio Choice 487
19.2 Efficient Portfolio 488
19.3 Efficient Portfolios in Practice 496
19.4 The Capital Asset Pricing Model 497
19.5 Exercises 499

20 Computationally Intensive Techniques 501
20.1 Simplicial Depth 502
20.2 Projection Pursuit 505
20.3 Sliced Inverse Regression 511
20.4 Support Vector Machines 519
20.5 Classification and Regression Trees 534
20.6 Boston Housing 552
20.7 Exercises 554

Part IV Appendix

21 Symbols and Notations 557

22 Data 561
22.1 Boston Housing Data 561
22.2 Swiss Bank Notes 562
22.3 Car Data 562
22.4 Classic Blue Pullovers Data 563
22.5 US Companies Data 563
22.6 French Food Data 563
22.7 Car Marks 564
22.8 French Baccalauréat Frequencies 564
22.9 Journaux Data 564
22.10 US Crime Data 565
22.11 Plasma Data 566
22.12 WAIS Data 566
22.13 ANOVA Data 567
22.14 Timebudget Data 567
22.15 Geopol Data 568
22.16 US Health Data 569
22.17 Vocabulary Data 570
22.18 Athletic Records Data 570
22.19 Unemployment Data 570
22.20 Annual Population Data 570
22.21 Bankruptcy Data I 571
22.22 Bankruptcy Data II 571

References 573

Index 577