

Wolfgang Härdle  
Zdeněk Hlávka

# Multivariate Statistics:

Exercises and Solutions



Springer

---

## Preface

There can be no question, my dear Watson, of the value of exercise before breakfast.

Sherlock Holmes in “The Adventure of Black Peter”

The statistical analysis of multivariate data requires a variety of techniques that are entirely different from the analysis of one-dimensional data. The study of the joint distribution of many variables in high dimensions involves matrix techniques that are not part of standard curricula. The same is true for transformations and computer-intensive techniques, such as projection pursuit.


The purpose of this book is to provide a set of exercises and solutions to help the student become familiar with the techniques necessary to analyze high-dimensional data. It is our belief that learning to apply multivariate statistics is like studying the elements of a criminological case. To become proficient, students must not simply follow a standardized procedure, they must compose with creativity the parts of the puzzle in order to see the big picture. We therefore refer to Sherlock Holmes and Dr. Watson citations as typical descriptors of the analysis.

Puerile as such an exercise may seem, it sharpens the faculties of observation, and teaches one where to look and what to look for.

Sherlock Holmes in “Study in Scarlet”

Analytic creativity in applied statistics is interwoven with the ability to see and change the involved software algorithms. These are provided for the student via the links in the text. We recommend doing a small number of problems from this book a few times a week. And, it does not hurt to redo an exercise, even one that was mastered long ago. We have implemented in these links software quantlets from XploRe and R. With these quantlets the student can reproduce the analysis on the spot.

This exercise book is designed for the advanced undergraduate and first-year graduate student as well as for the data analyst who would like to learn the various statistical tools in a multivariate data analysis workshop.

The chapters of exercises follow the ones in Härdle & Simar (2003). The book is divided into three main parts. The first part is devoted to graphical techniques describing the distributions of the variables involved. The second part deals with multivariate random variables and presents from a theoretical point of view distributions, estimators, and tests for various practical situations. The last part is on multivariate techniques and introduces the reader to the wide selection of tools available for multivariate data analysis. All data sets are downloadable at the authors' Web pages. The source code for generating all graphics and examples are available on the same Web site. Graphics in the printed version of the book were produced using XploRe. Both XploRe and R code of all exercises are also available on the authors' Web pages. The names of the respective programs are denoted by the symbol .

In Chapter 1 we discuss boxplots, graphics, outliers, Flury-Chernoff faces, Andrews' curves, parallel coordinate plots and density estimates. In Chapter 2 we dive into a level of abstraction to relearn the matrix algebra. Chapter 3 is concerned with covariance, dependence, and linear regression. This is followed by the presentation of the ANOVA technique and its application to the multiple linear model. In Chapter 4 multivariate distributions are introduced and thereafter are specialized to the multinormal. The theory of estimation and testing ends the discussion on multivariate random variables.

The third and last part of this book starts with a geometric decomposition of data matrices. It is influenced by the French school of data analysis. This geometric point of view is linked to principal component analysis in Chapter 9. An important discussion on factor analysis follows with a variety of examples from psychology and economics. The section on cluster analysis deals with the various cluster techniques and leads naturally to the problem of discrimination analysis. The next chapter deals with the detection of correspondence between factors. The joint structure of data sets is presented in the chapter on canonical correlation analysis, and a practical study on prices and safety features of automobiles is given. Next the important topic of multidimensional scaling is introduced, followed by the tool of conjoint measurement analysis. Conjoint measurement analysis is often used in psychology and marketing to measure preference orderings for certain goods. The applications in finance (Chapter 17) are numerous. We present here the CAPM model and discuss efficient portfolio allocations. The book closes with a presentation on highly interactive, computationally intensive, and advanced nonparametric techniques.

A book of this kind would not have been possible without the help of many friends, colleagues, and students. For many suggestions on how to formulate the exercises we would like to thank Michal Benko, Szymon Borak, Ying

Chen, Sigbert Klinke, and Marlene Müller. The following students have made outstanding proposals and provided excellent solution tricks: Jan Adamčák, David Albrecht, Lütfiye Arslan, Lipi Banerjee, Philipp Batz, Peder Egemen Baykan, Susanne Böhme, Jan Budek, Thomas Diete, Daniel Drescher, Zeno Enders, Jenny Frenzel, Thomas Giebe, LeMinh Ho, Lena Janys, Jasmin John, Fabian Kittman, Lenka Komárková, Karel Komorád, Guido Krbetschek, Yulia Maletskaya, Marco Marzetti, Dominik Michálek, Alena Myšičková, Dana Novotny, Björn Ohl, Hana Pavlovičová, Stefanie Radder, Melanie Reichelt, Lars Rohrschneider, Martin Rolle, Elina Sakovskaja, Juliane Scheffel, Denis Schneider, Burcin Sezgen, Petr Stehlík, Marius Steininger, Rong Sun, Andreas Uthemann, Aleksandrs Vatajins, Manh Cuong Vu, Anja Weiß, Claudia Wolff, Kang Xiaowei, Peng Yu, Uwe Ziegenhagen, and Volker Ziemann. The following students of the computational statistics classes at Charles University in Prague contributed to the R programming: Alena Babiaková, Blanka Hamplová, Tomáš Hovorka, Dana Chromíková, Kristýna Ivanková, Monika Jakubcová, Lucia Jarešová, Barbora Lebdušková, Tomáš Marada, Michaela Maršálková, Jaroslav Pazdera, Jakub Pečánka, Jakub Petrásek, Radka Picková, Kristýna Sionová, Ondřej Šedivý, Tereza Těšitelová, and Ivana Žohová.

We acknowledge support of MSM 0021620839 and the teacher exchange program in the framework of Erasmus/Sokrates.

We express our thanks to David Harville for providing us with the LaTeX sources of the starting section on matrix terminology (Harville 2001). We thank John Kimmel from Springer Verlag for continuous support and valuable suggestions on the style of writing and the content covered.

Berlin and Prague,  
April 2007

*Wolfgang K. Härdle*  
*Zdeněk Hlávka*

---

# Contents

Symbols and Notation .....	1
Some Terminology .....	5

---

## Part I Descriptive Techniques

---

1 Comparison of Batches .....	15
-------------------------------	----

---

## Part II Multivariate Random Variables

---

2 A Short Excursion into Matrix Algebra.....	33
3 Moving to Higher Dimensions .....	39
4 Multivariate Distributions .....	55
5 Theory of the Multinormal .....	81
6 Theory of Estimation .....	99
7 Hypothesis Testing .....	111

---

**Part III Multivariate Techniques**

---

<b>8</b>	<b>Decomposition of Data Matrices by Factors</b> .....	147
<b>9</b>	<b>Principal Component Analysis</b> .....	163
<b>10</b>	<b>Factor Analysis</b> .....	185
<b>11</b>	<b>Cluster Analysis</b> .....	205
<b>12</b>	<b>Discriminant Analysis</b> .....	227
<b>13</b>	<b>Correspondence Analysis</b> .....	241
<b>14</b>	<b>Canonical Correlation Analysis</b> .....	263
<b>15</b>	<b>Multidimensional Scaling</b> .....	271
<b>16</b>	<b>Conjoint Measurement Analysis</b> .....	283
<b>17</b>	<b>Applications in Finance</b> .....	291
<b>18</b>	<b>Highly Interactive, Computationally Intensive Techniques</b> .....	301
<b>A</b>	<b>Data Sets</b> .....	325
	A.1 Athletic Records Data.....	325
	A.2 Bank Notes Data.....	327
	A.3 Bankruptcy Data.....	331
	A.4 Car Data.....	333
	A.5 Car Marks.....	335
	A.6 Classic Blue Pullover Data.....	336
	A.7 Fertilizer Data.....	337
	A.8 French Baccalauréat Frequencies.....	338
	A.9 French Food Data.....	339

A.10	Geopol Data . . . . .	340
A.11	German Annual Population Data . . . . .	342
A.12	Journals Data . . . . .	343
A.13	NYSE Returns Data . . . . .	344
A.14	Plasma Data . . . . .	347
A.15	Time Budget Data . . . . .	348
A.16	Unemployment Data . . . . .	350
A.17	U.S. Companies Data . . . . .	351
A.18	U.S. Crime Data . . . . .	353
A.19	U.S. Health Data . . . . .	355
A.20	Vocabulary Data . . . . .	357
A.21	WAIS Data . . . . .	359
	<b>References . . . . .</b>	<b>361</b>
	<b>Index . . . . .</b>	<b>363</b>