Journal of Statistical Software

January 2016, Volume 69, Book Review 1. doi: 10.18637/jss.v069.b01

Reviewer: Abdolvahab Khademi
University of Massachusetts

Multivariate Statistics: Exercises and Solutions, 2nd Edition

Wolfgang Karl Härdle, Zdeněk Hlávka
Springer-Verlag, Berlin, 2015.
ISBN 9783642360046. xxiv+362 pp. USD 79.99 (P).
http://www.springer.com/9783642360046

High-dimensional data have frequently been used in many research areas and are even in more use today thanks to higher computing power and more optimized software and algorithms. All these applications and advances in statistical computing have made high-dimensional data analysis a valuable tool in fields as diverse as engineering, marketing, education, health sciences, biology, and computer science. This fact is reflected in multivariate statistics books in which in addition to the presentations of new methods, more traditional methods are revisited using new computing tools and application in more diverse areas.

There are several books on multivariate statistics that efficiently introduce the theory and application of multivariate statistics. This book solely focuses on problems, exercises and solutions in multivariate statistics. It is written in twenty chapters grouped into three parts: (I) descriptive techniques, (II) multivariate random variables, and (III) multivariate techniques. Because some exercises are solved through coding, the book companion websites make the code available both in R and MATLAB. Some solutions are coded in both languages while some exclusively in one. In addition, the data files used with the code can be found both at the book's project website and at the publisher's. A codebook is provided at the end of the book for the data sets used. A glossary of terms comes at the beginning of the book, which provides a glimpse into the fundamental concepts treated in the book. Part I of the book reminds us that despite the plethora of sophisticated statistical tools, simple diagrams are still the first step in any data analysis endeavor. In Part II, the authors lay the theoretical and mathematical foundations of multivariate techniques. And in Part III, the authors introduce the most common and general techniques used in high-dimensional data analysis, all through brief expositions yet extensive exercises.

Part I, *Descriptive Techniques*, comprises only Chapter 1 (*Comparison of Batches*). This chapter shows how to visualize multivariate data using boxplots, histograms, Chernoff-Flury faces, Andrew's curves and matrix plots. The purpose of this chapter is mainly to show the reader that examining the graphs of data before their numerical analysis can help us understand the patterns in the data, especially in the case of anomalous data points. Although the chapter is named descriptive, it primarily chooses a graphical approach to describe the data

rather than the traditional numerical methods. However, the exercises elicit both graphical and numerical responses from the reader.

Part II, *Multivariate Random Variables*, in six chapters lays the theoretical and mathematical foundation required to understand and use multivariate statistics. Readers need to have an adequate grasp of matrix algebra and multivariable calculus to be able to understand the concepts and solve the problems. In addition, knowledge of the programming languages R and MATLAB is assumed for completing some of the exercises. Chapter 2 (*A Short Excursion to Matrix Algebra*) is a concise refresher on matrix algebra needed for understanding the concepts and doing the exercises. Chapter 3 (*Moving to Higher Dimensions*) builds the premise that higher dimensional data are correlated data and hence understanding of correlation and covariance is essential for this purpose. The authors treat these concepts extensively through a linear model framework (ANOVA and regression) with least squares estimation.

In Chapter 4 (*Multivariate Distributions*), readers are introduced to joint distributions, joint cdf and pdf, dependency, multivariate moments, conditional moments, and transformations. The discussion of the multivariate normal distribution is dominant in this chapter. Chapter 5 (*Theory of the Multinormal*) solely treats the multivariate normal distribution. Properties of the multivariate distribution are expressed through theorems. This chapter closely extends the treatment in Chapter 4. In Chapter 6 (*Theory of Estimation*), the book transitions to inferential statistics. The maximum likelihood estimation (MLE) method, score function, Fisher information matrix and the Cramer-Rao inequality are presented. Once parameter estimation is introduced and practiced, Chapter 7 takes on *Hypothesis Testing* as a next step in inferential statistics. The likelihood ratio (LR) test and simultaneous confidence intervals are used along the MLE method to test different hypotheses. Every exercise in this chapter is worth the time to be done again and again.

Part III of the book, *Multivariate Techniques*, with thirteen chapters, is the most applied section of the book and perhaps appealing to a broader spectrum of audience due to its application in diverse fields. Chapter 8 (*Regression Analysis*) introduces linear regression and logistic regression as instances of generalized linear models (GLM). This chapter uses several data analysis examples with their respective outputs, interpretation of the results and graphs. The treatment of linear models is extended to Chapter 9 (*Variable Selection*) where the authors touch upon the multicollinearity problem and alternative methods such as regression on principal components, ridge regression, stepwise model selection, lasso, and elastic net.

The *Decomposition of Data Matrices by Factors* in Chapter 10 lays the foundation for the forthcoming chapters on dimension reduction techniques. The authors adopt a geometrical approach to project the data and to reduce it through matrix decomposition. The notion of inertia is used to indicate the amount of variance explained by the derived factors. Chapter 11 presents a mathematical treatment of *Principal Component Analysis* (PCA) followed by extensive theoretical and practical (data analysis) exercises. Rich graphical outputs and elaborate explanation and interpretation of the outputs walk the reader through the practice of PCA. Dimension reduction using correlation is treated in Chapter 12 (*Factor Analysis*), where the concept of common factor, estimation methods, and factor rotation are introduced.

Chapter 13 (*Cluster Analysis*) introduces cluster analysis methods mainly focusing on agglomerative procedures along with different ways of calculating the distance between clusters. The exercises focus on calculation of distance, derivation, proofs, and several applied ones using

data sets and practicing with different clustering algorithms and linkage methods. All applied exercises are accompanied by graphical outputs (dendrograms) and code available online. *Discriminant Analysis* is presented in Chapter 14 where ML and Fisher's linear discrimination function (and Bayes discriminant rule) are briefly discussed. Dimension reduction in categorical data is presented in Chapter 15 (*Correspondence Analysis*) with extensive exercises aimed at both theory and application.

The association between two linearly combined sets of variables is presented in Chapter 16 (*Canonical Correlation Analysis*) with several exercises on theory and application using code. *Multidimensional Scaling* is presented in Chapter 17 with the majority of exercises focusing on application of this method to data, with a greater emphasis on nonmetric solutions.

Conjoint analysis is a survey method used primarily in marketing (and other fields where respondents have stated preferences). This topic is briefly touched upon in Chapter 18 (*Conjoint Measurement Analysis*), with extensive hand calculation exercises and a few coded solutions. Next, the authors show the application of multivariate data analysis in quantitative finance, risk management, and portfolio optimization in Chapter 19 (*Applications in Finance*), focusing on efficient portfolios and capital asset price model. Finally, Chapter 20 (*Highly Interactive, Computationally Intensive Techniques*) discusses dimension reduction methods such as simplicial depth, exploratory projection pursuit, sliced inverse regression, classification and tree regression, and support vector machines. This chapter provides an excellent opportunity for those interested in computational statistics and statistical learning.

Multivariate Statistics: Exercises and Solutions, 2nd Edition is an exercise book intended to be used in conjunction with its parent textbook (Härdle and Simar 2015) which includes a more elaborate treatment of the corresponding chapters. The exercises in the book are designed both in the practice sense of the term and the mathematical rigor that challenges the more advanced reader or the aspiring expert in the field. Scholarly work in applied statistics ranges from studies that only use methods to those that develop new methods or adapt the existing ones for special problems. For the latter case, in-depth understanding and mastery of the mathematics of the methods cannot be overstated and this goal is well achieved in all theoretical exercises in the book.

One strength of the book is the diversity of topics that covers almost all the major applications of multivariate statistics. With twenty chapters, the topics will appeal to practitioners from fields as diverse as theoretical statistics, economics, finance, education, public health, and marketing. Practitioners will find the diverse exercises within each chapter intellectually rewarding and practically useful. All exercises come with fully worked-out solutions, which give an insight to the thought processes in solving both theoretical and practical problems in multivariate probability and statistics.

Another strength of the book is the integration of programming code in many of the exercises. The authors have included both R and MATLAB code (on the companion websites) to emphasize that learning and doing higher dimensional data analysis may not be possible without intense computation. Readers who may already have familiarity with multivariate statistics will have the opportunity to learn these popular programming languages through doing the applied exercises in the book.

As we know, data visualization is a great aid in representing and understanding data. This is one of the richest books in exhibiting many graphs to visualize data and solutions in multivariate analysis. In addition to visual appeal, the use of graphs is the most efficient way

to communicate the nature of data and the results of analyses to an audience with diverse levels of statistical literacy. This is well accomplished and implied in this book.

As was mentioned before, *Multivariate Statistics: Exercises and Solutions, 2nd Edition* is not written to teach multivariate statistics. Rather, it is intended to be accompanied by the authors' textbook (or any similarly structured book) for the purpose of reinforcing one's understanding of the theoretical and applied concepts in high dimensional data through solving numerous exercises, ranging from proofs to data analysis. Adopted together with the textbook, the exercises in this book will help graduate students in statistics and quantitative fields build a solid foundation to learn and do multivariate statistics. Multivariate statistics courses looking for a lab book should find this book well aligned.

References

Härdle WK, Simar L (2015). *Applied Multivariate Statistical Analysis*. Springer-Verlag, Berlin.

Reviewer:

Abdolvahab Khademi
University of Massachusetts
Department of Mathematics and Statistics
Amherst MA 01002, United States of America
E-mail: khademi@math.umass.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
January 2016, Volume 69, Book Review 1 *Published:* 2016-01-29
doi:10.18637/jss.v069.b01