

Preface

Not so long ago, multivariate analysis consisted solely of linear methods illustrated on small to medium-sized data sets. Moreover, statistical computing meant primarily batch processing (often using boxes of punched cards) carried out on a mainframe computer at a remote computer facility. During the 1970s, interactive computing was just beginning to raise its head, and exploratory data analysis was a new idea. In the decades since then, we have witnessed a number of remarkable developments in local computing power and data storage. Huge quantities of data are being collected, stored, and efficiently managed, and interactive statistical software packages enable sophisticated data analyses to be carried out effortlessly. These advances enabled new disciplines called data mining and machine learning to be created and developed by researchers in computer science and statistics.

As enormous data sets become the norm rather than the exception, statistics as a scientific discipline is changing to keep up with this development. Instead of the traditional heavy reliance on hypothesis testing, attention is now being focused on information or knowledge discovery. Accordingly, some of the recent advances in multivariate analysis include techniques from computer science, artificial intelligence, and machine learning theory. Many of these new techniques are still in their infancy, waiting for statistical theory to catch up.

The origins of some of these techniques are purely algorithmic, whereas the more traditional techniques were derived through modeling, optimiza-

tion, or probabilistic reasoning. As such algorithmic techniques mature, it becomes necessary to build a solid statistical framework within which to embed them. In some instances, it may not be at all obvious why a particular technique (such as a complex algorithm) works as well as it does:

When new ideas are being developed, the most fruitful approach is often to let rigor rest for a while, and let intuition reign — at least in the beginning. New methods may require new concepts and new approaches, in extreme cases even a new language, and it may then be impossible to describe such ideas precisely in the old language.

— Inge S. Helland, 2000

It is hoped that this book will be enjoyed by those who wish to understand the current state of multivariate statistical analysis in an age of high-speed computation and large data sets. This book mixes new algorithmic techniques for analyzing large multivariate data sets with some of the more classical multivariate techniques. Yet, even the classical methods are not given only standard treatments here; many of them are also derived as special cases of a common theoretical framework (multivariate reduced-rank regression) rather than separately through different approaches. Another major feature of this book is the novel data sets that are used as examples to illustrate the techniques.

I have included as much statistical theory as I believed is necessary to understand the development of ideas, plus details of certain computational algorithms; historical notes on the various topics have also been added wherever possible (usually in the **Bibliographical Notes** at the end of each chapter) to help the reader gain some perspective on the subject matter. **References** at the end of the book should be considered as extensive without being exhaustive.

Some common abbreviations used in this book should be noted: “iid” means *independently and identically distributed*; “wrt” means *with respect to*; and “lhs” and “rhs” mean *left-* and *right-hand side*, respectively.

Audience

This book is directed toward advanced undergraduate students, graduate students, and researchers in statistics, computer science, artificial intelligence, psychology, neural and cognitive sciences, business, medicine, bioinformatics, and engineering. As prerequisites, readers are expected to have had previous knowledge of probability, statistical theory and methods, multivariable calculus, and linear/matrix algebra. Because vectors and matrices play such a major role in multivariate analysis, Chapter 3 gives the matrix notation used in the book and many important advanced concepts in matrix theory. Along with a background in classical statistical theory

and methods, it would also be helpful if the reader had some exposure to Bayesian ideas in statistics.

There are various types of courses for which this book can be used, including data mining, machine learning, computational statistics, and for a traditional course in multivariate analysis. Sections of this book have been used at Temple University as the basis of lectures in a one-semester course in applied multivariate analysis to statistics and graduate business students (where technical derivations are skipped and emphasis is placed on the examples and computational algorithms) and a two-semester course in advanced topics in statistics given to graduate students from statistics, computer science, and engineering. I am grateful for their feedback (including spotting typos and inconsistencies).

Although there is enough material in this book for a two-semester course, a one-semester course in traditional multivariate analysis can be drawn from the material in Sections 1.1–1.3, 2.1–2.3, 2.5, 2.6, 3.1–3.5, 5.1–5.7, 6.1–6.3, 7.1–7.3, 8.1–8.7, 12.1–12.4, 13.1–13.9, 15.4, and 17.1–17.4; additional parts of the book can be used as appropriate.

Software

Software for computing the techniques described in this book is publicly available either through routines in major computer packages or through download from Internet websites. I have used primarily the R, S-PLUS, and MATLAB packages in writing this book. In the **Software Packages** section at the ends of certain chapters, I have listed the relevant R/S-PLUS routines for the respective chapter as well as the appropriate toolboxes in MATLAB. I have also tried to indicate other major packages wherever relevant.

Data Sets

The many data sets that illustrate the multivariate techniques presented in this book were obtained from a wide variety of sources and disciplines and will be made available through the book's website. Disciplines from which the data were obtained include astronomy, bioinformatics, botany, chemometrics, criminology, food science, forensic science, genetics, geoscience, medicine, philately, physical anthropology, psychology, soil science, sports, and steganography. Part of the learning process for the reader is to become familiar with the classic data sets that are associated with each technique. In particular, data sets from popular data repositories are used to compare and contrast methodologies. Examples in the book involve small data sets (if a particular point or computation needs clarifying) and large data sets (to see the power of the techniques in question).

Exercises

At the end of every chapter (except Chapter 1), there is a number of exercises designed to make the reader (a) relate the problem to the text and fill in the technical details omitted in the development of certain techniques,

(b) illustrate the techniques described in the chapter with real data sets that can be downloaded from Internet websites, and (c) write software to carry out an algorithm described in the chapter. These exercises are an integral part of the learning experience. The exercises are not uniform in level of difficulty; some are much easier than others, and some are taken from research publications.

Book Website

The book's website is located at:

`http://astro.ocis.temple.edu/~alan/MMST`

where additional materials and the latest information will be available, including data sets and R and S-PLUS code for many of the examples in the book.

Acknowledgments

I would like to thank David R. Brillinger, who instilled in me a deep appreciation of the interplay between theory, data analysis, computation, and graphical techniques long before attention to their connections became fashionable.

There are a number of people who have helped in the various draft stages of this book, either through editorial suggestions, technical discussions, or computational help. They include Bruce Conrad, Adele Cutler, Gene Fiorini, Burt S. Holland, Anath Iyer, Vishwanath Iyer, Joseph Jupin, Chuck Miller, Donald Richards, Cynthia Rudin, Yan Shen, John Ulicny, Allison Watts, and Myra Wise. Special thanks go to Richard M. Heiberger for his invaluable advice and willingness to share his expertise in all matters computational. Thanks also go to Abraham “Adi” Wyner, whose conversations at Border’s Bookstore kept me fueled literally and figuratively. Thanks also go to the reviewers and to all the students who read through various drafts of this book. Individuals who were kind enough to allow me to use their data or with whom I had e-mail discussions to clarify the nature of the data are acknowledged in footnotes at the place the data are first used. I would also like to thank the *Springer* editor John Kimmel, who provided help and support during the writing of this book, and the *Springer* L^AT_EX expert Frank Ganz for his help.

Finally, I thank my wife Betty-Ann and daughter Kayla whose patience and love these many years enabled this book to see the light of day.

Alan Julian Izenman
Philadelphia, Pennsylvania
April 2008



<http://www.springer.com/978-0-387-78188-4>

Modern Multivariate Statistical Techniques
Regression, Classification, and Manifold Learning

Izenman, A.J.

2008, XXV, 733 p., Hardcover

ISBN: 978-0-387-78188-4