

Preface

Why We Wrote This Book

This book is about using graphs to explore and model continuous multivariate data. Such data are often modelled using the multivariate normal distribution and, indeed, there is a literature of weighty statistical tomes presenting the mathematical theory of this activity. Our book is very different. Although we use the methods described in these books, we focus on ways of exploring whether the data do indeed have a normal distribution. We emphasize outlier detection, transformations to normality and the detection of clusters and unsuspected influential subsets. We then quantify the effect of these departures from normality on procedures such as discrimination and cluster analysis.

The normal distribution is central to our book because, subject to our exploration of departures, it provides useful models for many sets of data. However, the standard estimates of the parameters, especially the covariance matrix of the observations, are highly sensitive to the presence of outliers. This is both a blessing and a curse. It is a blessing because, if we estimate the parameters with the outliers excluded, their effect is appreciable and apparent if we then include them for estimation. It is however a curse because it can be hard to detect which observations are outliers. We use the forward search for this purpose.

The search starts from a small, robustly chosen, subset of the data that excludes outliers. We then move forward through the data, adding observations to the subset used for parameter estimation. As we move forward we

monitor statistical quantities such as parameter estimates, Mahalanobis distances and test statistics. In this way we can immediately detect the presence of outliers and clusters of observations and determine their effect on inferences drawn from the data. We can then improve our models.

This book is a companion to “*Robust Diagnostic Regression Analysis*” by Atkinson and Riani published by Springer in 2000. In the preface to that book we wrote “This bald statement . . . masks the excitement we feel about the methods we have developed based on the forward search. We are continuously amazed, each time we analyze a new set of data, by the amount of information the plots generate and the insights they provide”. Although more years have passed than we intended before the completion of our new book, in which process we have become three authors rather than two, this statement of our enthusiasm still holds.

For Whom We Wrote It

We have written our book to be of use and interest both to professional statisticians and other scientists concerned with data analysis as well as to postgraduate students. Because data analysis requires software we have a web site <http://stat.econ.unipr.it/riani/arc> which includes programs and the data. The programming was done in GAUSS, with most graphs for publication prepared in S-Plus.

The programs on our web site are in S-Plus. In addition Luca Scrucca of the University of Perugia (Italy) has translated the forward search routines into the R language (<http://www.r-project.org>). His routines are at <http://www.stat.unipg.it/luca/fwd>. Also, Stanislav Kolenikov of the University of North Carolina has created in STATA (<http://www.stata.com>) a module for forward search in regression available on <http://ideas.repec.org/c/bocode/s414902.html>. Links to forward search routines in other languages will be put on the web site of the book as they become known to us.

Our book is intended to serve as the text for a postgraduate course on modern multivariate statistics. The theoretical material is complemented by exercises with detailed solutions. In this way we avoid interrupting the flow of our data analytical arguments. We give references to the statistical literature, but believe that our book is reasonably self-contained. It should serve as a textbook even for courses in which the emphasis is not on the forward search. We trust such courses will decrease in number.

What Is In Our Book

The first chapter of this book introduces the forward search and contains four examples of its use for multivariate data analysis. We show how outliers and groups in the data can be identified and introduce some important plots. The second chapter, on theory, is in two parts. The first gives the distributional theory for a single sample from a multivariate normal distribution, with particular emphasis on the distributions of various Mahalanobis distances. The second part of the chapter contains a detailed description of the forward search and its properties. An understanding of all details of this chapter is not essential for an appreciation of the uses of the forward search in the later chapters. If you feel you know enough statistical theory for your present purposes, continue to Chapter 3.

The next three chapters describe methods for a sample believed to be from a single multivariate normal distribution. Chapter Three continues, extends and amplifies the analyses of the four examples from Chapter 1. In Chapter 4 we apply the forward search to multivariate transformations to normality. Analyses of three of the examples from earlier chapters are supplemented by the analysis of three new examples. Chapter 5 contains our first use of the forward search in a procedure depending on multivariate normality, that of principal components analysis. We are particularly interested in how the components are affected by outliers and other unsuspected structure in the data.

The two following chapters describe the forward search for data in several groups rather than one. In Chapter 6 the subject is discriminant analysis and in Chapter 7 cluster analysis, where the number of groups, as well as their composition, is unknown. Here the forward search enables us to see how individual observations are distorting the boundaries between our putative clusters. Finally, in Chapter 8 we consider the analysis of spatial data, which has something in common with the regression analysis of our earlier book.

Our Thanks

As with our first book, the writing of this book and the research on which it is based, have been both complicated and enriched by the fact that the authors are separated by half of Europe. Our travel has been supported by grants from the Italian Ministry for Scientific Research, by the Department of Economics of the University of Parma and by the Staff Research Fund of the London School of Economics. We are grateful to our numerous colleagues for their help in many ways. In England we especially thank Dr Martin Knott at the London School of Economics, who has been a steadfast source of help with both statistics and computing. Kjell Konis, currently at

Oxford University, helped greatly with the S-Plus programming. In Italy we thank Professor Sergio Zani of the University of Parma for his continuing support and his colleagues Aldo Corbellini and Fabrizio Laurini for help with computing including L^AT_EX. We also thank our families who have endured our absences and provided hospitality. Their support has been vital.

Our book was read by three anonymous reviewers for Springer-Verlag. We are very grateful both for their enthusiasm for our project and for their detailed comments, many of which we have incorporated to improve readability, flow and focus. Unfortunately one of their suggested objectives escaped us - to produce a shorter volume. We trust that the 390 figures in our book will make it seem not at all like a tome. In reviewing our first, and shorter, book for the Journal of the Royal Statistical Society, Gabrielle Kelly wrote “I read this (hardback) book, compulsive reading such as it was, in three sittings”. Even if it takes them more than three sittings, we hope many readers will find this new book similarly enjoyable.

Anthony Atkinson

`a.c.atkinson@lse.ac.uk`

`http://stats.lse.ac.uk/atkinson/`

London, England

Marco Riani

`mriani@unipr.it`

`http://www.riani.it`

`http://economia.unipr.it/docenti/riani`

`http://stat.econ.unipr.it/riani`

Andrea Cerioli

`andrea.cerioli@unipr.it`

`http://economia.unipr.it/docenti/cerioli`

`http://stat.econ.unipr.it/cerioli`

Parma, Italy

June 2003

Exploring Multivariate Data with the Forward Search

Atkinson, A.; Riani, M.; cccc, A.

2004, XXIV, 624 p., Hardcover

ISBN: 978-0-387-40852-1