Use R!

Series Editors:
Robert Gentleman Kurt Hornik Giovanni Parmigiani

For other titles published in this series, go to
http://www.springer.com/series/6991

Ron Wehrens

Chemometrics with R

Multivariate Data Analysis in the Natural Sciences and Life Sciences

Springer

Ron Wehrens
Fondazione Edmund Mach
Research and Innovation Centre
Via E. Mach 1
38010 San Michele all'Adige
Italy
ron.wehrens@iasma.it

Series Editors:
Robert Gentleman
Program in Computational Biology
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue, N. M2-B876
Seattle, Washington 98109
USA

Kurt Hornik
Department of Statistik and Mathematik
Wirtschaftsuniversität Wien
Augasse 2-6
A-1090 Wien
Austria

Giovanni Parmigiani
The Sidney Kimmel Comprehensive
Cancer Center at Johns Hopkins University
550 North Broadway
Baltimore, MD 21205-2011
USA

ISBN 978-3-642-17840-5 e-ISBN 978-3-642-17841-2
DOI 10.1007/978-3-642-17841-2
Springer Heidelberg Dordrecht London New York

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

For Odilia, Chris and Luc

Preface

The natural sciences, and the life sciences in particular, have seen a huge increase in the amount and complexity of data being generated with every experiment. It is only some decades ago that scientists were typically measuring single numbers – weights, extinctions, absorbances – usually directly related to compound concentrations. Data analysis came down to estimating univariate regression lines, uncertainties and reproducibilities. Later, more sophisticated equipment generated complete spectra, where the response of the system is wavelength-dependent. Scientists were confronted with the question how to turn these spectra into useable results such as concentrations. Things became more complex after that: chromatographic techniques for separating mixtures were coupled to high-resolution (mass) spectrometers, yielding a data matrix for every sample, often with large numbers of variables in both chromatographic and spectroscopic directions. A set of such samples corresponds to a data cube rather than a matrix. In parallel, rapid developments in biology saw a massive increase in the ratio of variables to objects in that area as well.

As a result, scientists today are faced with the increasingly difficult task to make sense of it all. Although most will have had a basic course in statistics, such a course is unlikely to have covered much multivariate material. In addition, many of the classical concepts have a rough time when applied to the types of data encountered nowadays – the multiple-testing problem is a vivid illustration. Nevertheless, even though data analysis has become a field in itself (or rather: a large number of specialized fields), scientists generating experimental data should know at least some of the ways to interpret their data, if only to be able to ascertain the quality of what they have generated. Cookbook approaches, involving blindly pushing a sequence of buttons in a software package, should be avoided. Sometimes the things that deviate from expected behaviour are the most interesting in a data set, rather than unfortunate measurement errors. These deviations can show up at any time point during data analysis, during data preprocessing, modelling, interpretation... Every phase in this pipeline should be carefully executed and results, also at an intermediate stage, should be checked using common sense and prior knowledge.

This also puts restrictions on the software that is being used. It should be sufficiently powerful and flexible to fit complicated models and handle large and complex data sets, and on the other hand should allow the user to exactly follow what is being calculated – black-box software should be avoided if possible. Moreover, the software should allow for reproducible results, something that is hard to achieve with many point-and-click programs: even with a reasonably detailed prescription, different users can often obtain quite different results. R [1], with its rapidly expanding user community, nicely fits the bill. It is quickly becoming the most important tool in statistical bioinformatics and related fields. The base system already provides a large array of useful procedures; in particular, the high-quality graphics system should be mentioned. The most important feature, however, is the package system, allowing users to contribute software for their own fields, containing manual pages and examples that are directly executable. The result is that many packages have been contributed by users for specific applications; the examples and the manual pages make it easy to see what is happening.

Purpose of this book.

Something of this philosophy also can be found in the way this book is set up. The aim is to present a broad field of science in an accessible way, mainly using illustrative examples that can be reproduced immediately by the reader. It is written with several goals in mind:

- **An introduction to multivariate analysis.** On an abstract level, this book presents the route from raw data to information. All steps, starting from the data preprocessing and exploratory analysis to the (statistical) validation of the results, are considered. For students or scientists with little experience in handling real data, this provides a general overview that is sometimes hard to get from classical textbooks. The theory is presented as far as necessary to understand the principles of the methods and the focus is on immediate application on data sets, either from real scientific examples, or specifically suited to illustrate characteristics of the analyses.
- **An introduction to R.** For those scientists already working in the fields of bioinformatics, biostatistics and chemometrics but using other software, the book provides an accessible overview on how to perform the most common analyses in R [1]. Many packages are available on the standard repositories, CRAN[1] and BIOCONDUCTOR[2], but for people unfamiliar with the basics of R the learning curve can be pretty steep – for software, power and complexity are usually correlated. This book is an attempt to provide a more gentle way up.

[1] `http://cran.r-project.org`

[2] `http://www.bioconductor.org`

- **Combinding multivariate data analysis and R.** The combination of the previous two goals is especially geared towards university students, at the beginning of their specialization: it is of prime importance to obtain hands-on experience on real data sets. It does take some help to start reading R code – once a certain level has been reached, it becomes more easy. The focus therefore is not just on the use of the many packages that are available, but also on showing how the methods are implemented. In many cases, simplified versions of the algorithms are given explicitly in the text, so that the reader is able to follow step-by-step what is happening. It is this insight in (at least the basics of) the techniques that is essential for fruitful application.

The book has been explicitly set up for self-study. The user is encouraged to try out the examples, and to substitute his or her own data as well. If used in a university course, it is possible to keep the classical "teaching" of theory to a minimum; during the lessons, teachers can concentrate on the analysis of real data. There is no substitute for practice.

Prior knowledge.

Some material is assumed to be familiar. Basic statistics, for example, including hypothesis tests, the construction of confidence intervals, analysis of variance and least-squares regression are referred to, but not explained. The same goes for basic matrix algebra. The reader should have some experience in programming in general (variables, variable types, functions, program control, etcetera). It is assumed the reader has installed R, and has a basic working knowledge of R, roughly corresponding to having worked through the excellent "Introduction to R" [2], which can be found on the CRAN website. In some cases, less mundane functions will receive a bit more attention in the text; examples are the `apply` and `sweep` functions. We will only focus on the comman-line interface: Windows users may find it easier to perform actions using point-and-click.

The R package **ChemometricsWithR***.*

With the book comes a package, too: **ChemometricsWithR** contains all data sets and functions used in this book. Installing the package will cause all other packages used in the book to be available as well – an overview of these packages can be found in Appendix A. In the examples it is always assumed that the **ChemometricsWithR** package is loaded; where functions or data sets from other packages are used for the first time, this is explicitly mentioned in the text.

More information about the data sets used in the book can be found in the references – no details will be given about the background or interpretation of the measurement techniques.

Acknowledgements.

This book has its origins in a reader for the Chemometrics course at the Radboud University Nijmegen covering exploratory analysis (PCA), clustering (hierarchical methods and k-means), discriminant analysis (LDA, QDA) and multivariate regression (PCR, PLS). Also material from a later course in Pattern Recognition has been included. I am grateful for all the feedback from the students, and especially for the remarks, suggestions and criticisms from my colleagues at the Department of Analytical Chemistry of the Radboud University Nijmegen. I am indebted to Patrick Krooshof and Tom Bloemberg, who have contributed in a major way in developing the material for the courses. Finally, I would like to thank all who have read (parts of) the manuscript and with their suggestions have helped improving it, in particular Tom Bloemberg, Karl Molt, Lionel Blanchet, Pietro Franceschi, and Jan Gerretzen.

Trento, *Ron Wehrens*
September 2010

Contents

Part III Modelling

Part IV Model Inspection

Part V Applications

Part VI Appendices