Journal of Statistical Software

January 2007, Volume 18, Issue 1. *http://www.jstatsoft.org/*

An Introduction to the Special Volume "Spectroscopy and Chemometrics in R"

Katharine M. Mullen
Vrije Universiteit Amsterdam

Ivo H. M. van Stokkum
Vrije Universiteit Amsterdam

Abstract

This special volume collates ten issues under the rubric "Spectroscopy and Chemometrics in R". In so doing, it provides an overview of the breadth, depth and state of the art of R-based software projects for spectroscopy and chemometrics applications. Just as the authors have contributed to R their documentation and source code, so has R contributed to the quality, standardization and dissemination of their software, as this volume attests. We hope that the volume is inspiring to both computational statisticians interested in applications of their methodologies and to spectroscopists or chemometricians in need of solutions to their data analysis problems.

Keywords: spectroscopy, chemometrics, R.

1. Introduction

Statistical computing has assumed a central role in spectroscopy and chemometrics research. The R language and environment for statistical computing (R Development Core Team 2006) is at the forefront of the development of software for statistical computing, and has become a powerful tool for research in spectroscopy and chemometrics. The present special volume gives a flavour of the state of the art of the use of R in these areas.

The summer of 2006 was a watershed moment in the history of the interaction of R and the spectroscopy and chemometrics fields. Prior to then, researchers in these fields had found ample application for the extremely wide range of general-purpose statistical techniques available in R to their research, and a variety of R packages (i.e., modular software components) specifically targeted to the analysis of spectroscopy and chemometrics data had been released. However, no overview of the use of R in these application areas yet existed. Jan de Leeuw took a step toward changing this at the useR! conference in Vienna in June by suggesting the compilation of this special volume. After the conference we invited contributions from researchers in the community and the outlines of the present volume began to take shape.

Then, in August, a special issue of *R News* concentrating on applications in chemistry was released, guest edited by Ron Wehrens. This was the first such compilation in the literature, and began with the observation by Wehrens that, "R has become the standard for statistical analysis in biology and bioinformatics, but it is also gaining popularity in other fields of natural sciences" (Wehrens 2006). Three of the articles in the R News special issue have been expanded into contributions here.

For the uninitiated, some explanation may be required as to why an overview of spectroscopy and chemometrics tools and techniques in R *in particular* is necessary. After all, it may be argued that anything possible in one (Turing-complete) programming language can surely be done in another, and that one's choice of language is therefore not of much interest. However, the features of R in comparison to other currently available systems for statistical computing uniquely advantage spectroscopy and chemometrics researchers working in R. The hundreds of available packages and the fact that R is the *lingua franca* of computational statisticians means that an implementation of even very new statistical techniques is likely to be found in R. Furthermore, R is available under the terms of the GNU General Public License in source code form, and compiles and runs on all major operating systems. This means that R need not be treated as a black box: if necessary, researchers can examine or modify the techniques it implements at a very low level. It also means that without cost and without regard to choice of operating system, researchers using R can reproduce each other's results. Moreover, R (and its package system in particular) facilitates the standardization and distribution of software for spectroscopy and chemometrics, which up until now has too often remained proprietary to the particular department or laboratory in which it was developed, impeding the verification of research results and slowing progress in the dissemination of analysis methodology.

The useR! conference during the summer of 2006 not only gave birth to this special volume. It also brought home the point that *Journal of Statistical Software* (JSS) is an extremely suitable forum for publications describing statistical software based in R. The conference concluded with a panel discussion by representatives of well-established journals in computational statistics, discussing how "Getting recognition for excellence in computational statistics" might be accomplished. In the course of this discussion Jan de Leeuw made a compelling case for the advantages of publishing R-based research in JSS: like R, JSS is free, so that all interested may access it (assuming an internet connection), it supports reproducible research and the validation of software tools by both reviewing and publishing of the software associated with a manuscript, and the electronic format does away with archaic limitations of print journals like page and color restrictions. We are especially pleased that this compendium describing spectroscopy and chemometrics in R appears in JSS.

2. An outline of the contributions

The volume begins with a contribution from Mevik and Wehrens (2007) on the **pls** package for principal component and partial least squares regression. These multivariate regression methods are widely applied in spectroscopy and many other fields. The user interface of **pls** is modeled after the traditional formula interface used in R functions, as exemplified by the `lm` function. The authors provide an introduction to the algorithms as well as several case studies. Next, Mullen and van Stokkum (2007) present the **TIMP** package for modeling multi-way spectroscopic measurements. The package employs a partitioned variable projection algorithm for fitting the separable nonlinear models that are common in spectroscopy

data analysis. Models are constructed in **TIMP** with a rich variety of options, e.g., for compartmental models, spectral constraints, and multiexperiment parameterizations. Kirchner, Saussen, Steen, Steen, and Hamprecht (2007) follow with a description of the **amsrpm** package for preprocessing mass spectra measured in-line with liquid chromatography systems (LC/MS). Using a robust point matching algorithm the package compensates for instrumental distortions, and provides retention time alignment of the LC/MS data. Guha (2007) then introduces the **rcdk** package that provides R users easy access to the Chemistry Development Kit **CDK**, a Java framework for chemical informatics, as well as the **rpubchem** package that allows access to the data in PubChem, a public repository of molecular structures and associated assay data. The next contribution from Binsl, Mullen, van Beek, and Heringa (2007) describes the **FluxSimulator** package that is able to simulate isotopomer distributions given a specification of a metabolic network of arbitrary complexity. It is intended for use with isomer labeled substrates measured by mass or NMR spectroscopy to quantify metabolic networks. The sixth contribution by Laptenok, Mullen, Borst, van Stokkum, Apanasovich, and Visser (2007) shows the utility of the **TIMP** package in fluorescence lifetime imaging microscopy (FLIM), an important technique for the *in situ* and *in vivo* study of physico-chemical processes in biological objects. Case studies on real and simulated data evidence the applicability of the package for parameter estimation and also the limitations imposed by practical signal to noise ratios. Next, Fraley and Raftery (2007) review clustering techniques for chemometrics data analysis applications, illustrating their **mclust** package. They present case studies of density estimation and discriminant analysis. Ritter and Gilliard (2007) then describe **ImpuR**, a collection of diagnostic tools to explore time-intensity matrices with respect to their bilinear structure and departures from it. It is applied to chromatography combined with (UV) absorption spectroscopy. This is followed by a description by Krastev (2007) of **spectrino**, an R-based Windows application to organize and pre-process spectra. **spectrino** also offers a set of features to create data structures and visually manipulate/compare spectra, in particular from mass spectroscopy. The volume concludes with a contribution by Babu and Mahabal (2007) describing **VOStat**, an R-based online software suite written mainly for the Virtual Observatory, in which astronomers share large scale spectroscopic data.

3. Conclusions and outlook

The present volume showcases some of the breadth and depth of software tools for spectroscopy and chemometrics applications for the R language and environment. We hope the collection is inspiring to both computational statisticians interested in applications of their methodologies and to spectroscopists or chemometricians in need of solutions to their data analysis problems. Feedback to the authors may provide an incentive to further develop their methods and software for a broader applicability.

References

Babu GJ, Mahabal A (2007). "Using R-based **VOStat** as a Low-Resolution Spectrum Analysis Tool." *Journal of Statistical Software*, **18**(11). URL http://www.jstatsoft.org/v18/i11/.

Binsl TW, Mullen KM, van Beek JHGM, Heringa J (2007). "pkgFluxSimulator: An R Package to Simulate Isotopomer Distributions in Metabolic Networks." *Journal of Statistical Software*, **18**(7). URL http://www.jstatsoft.org/v18/i07/.

Fraley C, Raftery A (2007). "Model-based Methods of Classification: Using the **mclust** Software in Chemometrics." *Journal of Statistical Software*, **18**(6). URL http://www.jstatsoft.org/v18/i06/.

Guha R (2007). "Chemical Informatics Functionality in R." *Journal of Statistical Software*, **18**(5). URL http://www.jstatsoft.org/v18/i05/.

Kirchner M, Saussen B, Steen H, Steen JA, Hamprecht FA (2007). "**amsrpm**: Robust Point Matching for Retention Time Alignment of LC/MS Data with R." *Journal of Statistical Software*, **18**(4). URL http://www.jstatsoft.org/v18/i04/.

Krastev T (2007). "**Spectrino** Software: Spectra Visualization and Preparation for R." *Journal of Statistical Software*, **18**(10). URL http://www.jstatsoft.org/v18/i10/.

Laptenok S, Mullen KM, Borst JW, van Stokkum IHM, Apanasovich VV, Visser AJWG (2007). "Fluorescence Lifetime Imaging Microscopy (FLIM) Data Analysis with **TIMP**." *Journal of Statistical Software*, **18**(8). URL http://www.jstatsoft.org/v18/i08/.

Mevik BH, Wehrens R (2007). "The **pls** Package: Principal Component and Partial Least Squares Regression in R." *Journal of Statistical Software*, **18**(2). URL http://www.jstatsoft.org/v18/i02/.

Mullen KM, van Stokkum IHM (2007). "**TIMP**: An R Package for Modeling Multi-way Spectroscopic Measurements." *Journal of Statistical Software*, **18**(3). URL http://www.jstatsoft.org/v18/i03/.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ritter C, Gilliard J (2007). "**ImpuR**: A Collection of Diagnostic Tools Developed in R in the Context of Peak Impurity Detection in HPLC-DAD but Potentially Useful with Other Types of Time–Intensity Matrices." *Journal of Statistical Software*, **18**(10). URL http://www.jstatsoft.org/v18/i10/.

Wehrens R (2006). "Editorial." *R News*, **6**(3), 1–2.

Affiliation:

Katharine M. Mullen
Department of Physics and Astronomy
Faculty of Sciences
Vrije Universiteit Amsterdam
De Boelelaan 1081
1081 HV Amsterdam, The Netherlands
E-mail: kate@nat.vu.nl
URL: http://www.nat.vu.nl/~kate/

Ivo H. M. van Stokkum
Department of Physics and Astronomy
Faculty of Sciences
Vrije Universiteit Amsterdam
De Boelelaan 1081
1081 HV Amsterdam, The Netherlands
E-mail: ivo@nat.vu.nl
URL: http://www.nat.vu.nl/~ivo/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 18, Issue 1 *Submitted:* 2007-01-16
January 2007 *Accepted:* 2007-01-16