Embedding R in Windows applications, and executing R remotely

Thomas Baier and Erich Neuwirth

February 15, 2004

R is object oriented, and objects are the standard way of packing analysis results in R.

Many programming environments and applications programs in Windows can act as as (D)COM clients, which is the standard way of accessing objects exposed by (D)COM servers. Particularly, all Microsoft office programs are (D)COM clients and therefore can access any (D)COM server. Therefore, in encapsulating R as a (D)COM server is a natural choice to make R functionality accessible to other programs.

To embed R in other programs, therefore, one of the key questions is what kind of objects are exposed to these applications hosting R, and how these objects are exposed.

There are two different routes that can be taken here. We can either choose to expose R objects "as is", with all their richness, or we can choose a minimalist approach and only offer objects of types which can be handled easily by programs which normally do not employ rich object types for the data they usually handle.

The difference can be very well illustrated when R is embedded in Excel. A spreadsheet essentially has 6 data types, scalars, vectors, and matrices of either numbers or strings. If we want to make R functionality a part of the spreadsheet functionality, it is sufficient that the R (D)COM server exposed this type od data objects.

On the other hand, VBA (the programming language built into Excel) allows to work with any type of object. Therefore, the whole R object model, and even user-defined new object types, can be made accessible in VBA, and therefore be used in Excel.

The question is, how is R being used in connection with Excel. When the programmes "thinks R" and uses Excel just as a convenient data source and data editor, the full object model makes sense. Then, programming is done in R and VBA, and data and results are just transferred from time to time between worksheets and R. This way, Excel becomes a convenience item for R, but conceptually R is the center of the programming model.

If we want to use R as an extension of Excel worksheets, and only as subroutines accessible from VBA, the minimalist approach seems more adapted. In this case, calls to R will only return objects which can directly be embedded in worksheets. One of the key concepts or spreadsheet programs is automatic recalculation. Using data types which can immediately be embedded in the worksheet makes R calculations become part of Excel's automatic recalculation, thereby offering facilities not offered by R itself. Using only simple data types like arrays allows very fast implementation of the interface. Using the full R object model adds another level of complexity and therefore probably slows down calculation considerably. Calculation speed is very important for reasonable automatic recalculation, therefore this approach R leads to a less natural spreadsheet extension. Additionally, if the R server is executed on another machine than the client, transfer speed also plays an important role, and using only native Windows data types speeds up things considerably.

The relative merits of the 2 different approaches also heavily depend on the experience of the programmer using R as an embedded library. To be able to use the full R object hierarchy, one has to be rather knowledgeable about R's object model, and understand the relationships between different kinds of objects. Making R objects fully accessible in applications really puts just another kind of syntactic glue (in the case of Excel the glue is VBA) on top of R's objects.

Using the minimalist approach allows simpler access to R functions in other applications. The interface to R can be kept much simpler. Of course, the price to pay is that we do not have the full richness of R accessible in the application environment directly. It is, however, always possible to encapsulate everything in R code which only returns the simple data types.

If we separate R core functionality, especially the statistical methods needed in an application, from the data handling and interface code, it makes sense to write the core functionality as R functions returning only simple data types. Then, all the additional code (GUIs, spreadsheet functions) can be written without detailed knowledge of the R object hierarchy.

Another problem when we embed R into another application is who of the two partners is the authority for the state of data objects. When we transfer data to R, we assign the data to variables. What happens if the data in the hosting application is changed? Will these changes automatically propagate to R? As soon as we use variables in both applications, we have to be very careful about keeping variables synchronized.

If we apply a strict functional model, R only exposing (stateless) functions, and not (stateful) data objects, then we elegantly avoid this problem. To be able to apply that model, all the functions supplied by R have to return data types which can immediately be represented in the hosting application.

We will show some applications with both approaches, and we will demonstrate how the different approaches influence calculation speed.

Using R in a Distributed Computer Lab for Statistics Courses

Thomas Baier Erich Neuwirth

February 15, 2004

In many of today's computer labs it is not possible to provide a computer for every student. Fortunately many students already own a laptop computer which can be used for these labs. Universities tend to provide LAN or even WLAN access to the campus network for the student's computers. These facts already point out the solution to the problem.

We will discuss a software-architecture for using R in a distributed lab scenario and will describe the possible approaches and show the benefits and the problems arising from choosing one of them. The discussed scenarios will be

- lab-only installation and access of R
- lab installations and installations on student computers
- installation of R on lab computers and remote access via terminal services
- lab-provided installation of R and remote access via rich clients
- "repositories" of R on lab computers and static load balancing for access by rich clients on notebooks and lab computers

Our discussions will focus on using R as the computational engine while students are working in a spreadsheet-application on Windows platforms. Our main concerns are

- ease of installation both on lab computers and for the students' own computers
- transparency of computation
- maintenance of packages and installations
- administration of sample data and preparing exercises
- data storage and (semi-)automatic submission of results

Finally, we will show the implementation chosen for a course taking place in fall/winter 2004 at the University of Vienna, Austria and discuss future extensions using Web Service technology (SOAP/HTTP) as a portable client-interface.

Generic Functions for Spatial Data

Roger Bivand, Edzer J. Pebesma, Barry Rowlingson

A pre-DSC'03 workshop on spatial data handling in R addressed various forms of R/GIS integration, both loose and tight coupled. Topics discussed there included packages then already on CRAN (RArcInfo, GRASS, and the spatial statistics packages), and initiatives under development, especially Rmap, the R/GDAL package, functions to read and write shapefiles, TerraLib, and using StatConnector in ArcGIS, among others. This presentation describes work so far on a package of suitable classes and methods for spatial data. The classes document where the spatial location information resides, for 2D or 3D data. Utility functions are provided, e.g. for helping plotting data as maps, or spatial selection. The presentation will also refer to other work in progress on spatial data handling and spatial statistics in R.

hoa – A package bundle for higher order likelihood inference

Alessandra R. Brazzale[1] and **Ruggero Bellio**[2]

[1]Institute for Biomedical Engineering
Italian National Research Council
alessandra.brazzale@isib.cnr.it

[2]Department of Statistics
University of Udine, Italy
ruggero.bellio@dss.uniud.it

Since its introduction by Sir R. A. Fisher, the likelihood criterion has found extensive application in the analysis of data. The application of the central limit theorem to conclude that statistics such as the maximum likelihood estimator are approximately normally distributed, with mean and variance consistently estimable from the data, lead to the theory of first order asymptotic inference. Over the past twenty-five years or so very accurate approximations, generally called higher order approximations, to the distribution of the statistics involved have been developed. Although they are relatively easily derived using techniques adapted from the theory of asymptotic expansions, much application of likelihood inference still relies on first order asymptotics.

The purpose of this presentation is to illustrate how higher order likelihood theory can be applied in practice by using the software provided through the hoa package bundle. The applications considered are regression models, including logistic regression, non-normal linear regression and non-linear regression with normal errors and arbitrary variance function. These give rise to three of the four packages included in hoa, namely, in order, cond, marg and nlreg. A fourth packaged, called sampling, includes a Metropolis-Hastings sampler which can be used to simulate from the conditional distribution of the higher order statistics considered in marg.

On Multiple Comparisons in R

by Frank Bretz[1], Torsten Hothorn[2] and Peter Westfall[3]

[1] Universität Hannover, Lehrgebiet Bioinformatik, Hannover

[2] Friedrich-Alexander-Universität Erlangen-Nürnberg,
Institut für Medizininformatik,Biometrie und Epidemiologie

[3] Texas Tech University, Dept. of Information Systems and Quantitative Sciences, Lubbock, TX

The multcomp package for the R statistical environment allows for multiple comparisons of parameters whose estimates are generally correlated, including comparisons of k groups in general linear models. The package has many common multiple comparison procedures "hard-coded", including Dunnett, Tukey, sequential pairwise contrasts, comparisons with the average, changepoint analysis, Williams', Marcus', McDermott's, and tetrad contrasts. In addition, a free input interface for the contrast matrix allows for more general comparisons.

The comparisons itself are not restricted to balanced or simple designs. Instead, the programs are designed to suit general multiple comparisons, thus allowing for covariates, nested effects, correlated means, likelihood-based estimates, and missing values. For the homoscedastic normal linear models, the program accounts for the correlations between test statistics by using the exact multivariate t-distribution. The resulting procedures are therefore more powerful than the Bonferroni and Holm methods; adjusted p-values for these methods are reported for reference. For more general models, the program accounts for correlations using the asymptotic multivariate normal distribution; examples include multiple comparisons based on rank transformations, logistic regression, GEEs, and proportional hazards models. In the asymptotic case, the user must supply the estimates, the asymptotic covariance matrix, and the contrast matrix.

Basically, the package provides two functions. The first one computes confidence intervals for the common single-step procedures (simint). This approach is uniformly improved by the second function (simtest), which utilizes logical constraints and is closely related to closed testing. However, no confidence intervals are available for the simtest function.

In this talk we give an introduction to the multcomp package. We first provide a brief theoretical background on multiple comparisons and the multiple contrast representation. We then illustrate the use of the package by going through several examples.

Using R for Statistical Seismology

Ray Brownrigg

Statistical Seismology is a relatively new field which applies statistical methodology to earthquake data in an attempt to raise new questions about earthquake mechanisms, to characterise aftershock sequences and to make some progress towards earthquake prediction.

A suite of R packages, known as SSLib, is available for use with Statistical Seismology in general, but also with some applications outside this field. This presentation will introduce the packages, describe the way certain special features of R have been crafted to a framework for researchers in the field and demonstrate some of the exploratory data analysis functions available within the packages.

arrayMagic: two-colour DNA array quality control and preprocessing

Andreas Buneß, Wolfgang Huber, Klaus Steiner,
Holger Sültmann & Annemarie Poustka
Molecular Genome Analysis, German Cancer Research Center,
Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

arrayMagic is a software package which facilitates the analysis of two colour DNA microarray data. The package is written in R (http://www.r-project.org) and integrates into Bioconductor (http://www.bioconductor.org). The automated analysis pipeline comprises data loading, normalisation and quality diagnostics. The pipeline is flexbile and can be adjusted for specific needs.

The package takes advantage of the S4 class mechanism. The normalised data, as well as their annotation is stored in the class exprSetRG, an extension of the class exprSet of the library Biobase which accounts for the specific requirements of two colour microarray data and integrates well in the existing Bioconductor framework. Eventually, it will be merged into the new eSet class of Biobase.

Several quality diagnostic plots are generated on the fly and allow to assess the hybridisation quality and to discard low quality hybridisations from further analysis. Different normalisation methods are offered to remove systematic variations like hybridisation and dye effects. The pipeline supports to process microarray data at high throughput.

The `subselect` package - Selecting variable subsets in an exploratory data analysis.

Cadima, J.[*], Cerdeira, J.O.[†], Duarte Silva, A.P.[‡]& Minhoto, M.[§]

February 13, 2004

Identifying a subset of a large set of variables which can adequately replace the full data set is a problem that has been studied in different contexts and which is of widespread concern to analysts of large data sets.

The R package subselect provides functions which measure the quality of a given subset of k variables according to three criteria that are relevant for exploratory data analysis. These criteria measure: the similarity between subspaces spanned by a given subset of variables and a given subset of Principal Components of the full data set (via the mean of the squared canonical correlations of both sets of variables - Yanai's GCD); the similarity of the configurations of points obtained using all the original variables or only those in a given subset (via Escoufier's RV-coefficient); and the quality of a given variable subset as a predictor of all the original variables (McCabe's second criterion for Principal Variables). The subselect package also provides three different algorithms to search for optimal k-subsets, with respect to each of these criteria: a simulated annealing algorithm, a genetic algorithm and a modified local search algorithm.

New features of the package include an algorithm for a complete search (in the spirit of Furnival and Wilson's leaps-and-bounds algorithm for subset selection in multiple regression), which is viable for medium-sized data sets.

The issue of identifying a broad array of k-subsets that are maximal according to those three criteria taken simultaneously is also considered in the context of multi-criteria optimization.

[*] Departamento de Matemática, Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Lisbon, Portugal.

[†] Departamento de Matemática, Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Lisbon, Portugal.

[‡] Faculdade de Ciências Económicas e Empresariais, Universidade Católica Portuguesa, Oporto, Portugal.

[§] Departamento de Matemática, Universidade de Évora, Évora, Portugal.

Molecular signatures from gene expression data: algorithm development using R

David Casado and Ramón Díaz-Uriarte
Bioinformatics Unit, CNIO (Spanish National Cancer Centre)
Melchor Fernández Almagro 3
28029 Madrid
Spain.

April, 2004

Abstract

"Molecular signatures" or "gene-expression signatures" are used to model patients' clinically relevant information (e.g., prognosis, survival time) using expression data from coexpressed genes. Signatures are a key feature in cancer research because they can provide insight into biological mechanisms and have potential diagnostic use. However, available methods to search for signatures fail to address key requirements of signatures and signature components, especially the discovery of tightly coexpressed sets of genes.

We implement a method with good predictive performance that follows from the biologically relevant features of signatures. After identifying a seed gene with good predictive abilities, we search for a group of genes that is highly correlated with the seed gene, shows tight coexpression, and has good predictive abilities; this set of genes is reduced to a signature component using Principal Components Analysis. The process is repeated until no further component is found. Finally, to assess the stability of the obtained results, the bootstrap is used: biological interpretability is suspect if there is little overlap in the results from different bootstrap samples.

All the coding of the algorithm and its comparison with alternative approaches has been done using R and several packages available from CRAN (class —part of the VR bundle—, e1071) and Bioconductor (multtest), with a tiny bit of C++ code dynamically loaded into R. Right now, the predictive methods used include KNN and DLDA but, by using R, use of other methods is straightforward. This research and its code (released under the GNU GPL) is an example of the "turn ideas into software, quickly and faithfully" (Chambers, 1998, *"Programming with data"*) that is allowed by the S/R family of languages.

Calculating the autocovariances of fractional ARIMA model

Cheang Wai Kwong
Nanyang Technological University, Singapore

Abstract

Consider a stationary and invertible process $\{Y_t\}$ following a (long memory) fractional ARIMA(p,d,q) model,

$$\phi(B)(1-B)^d Y_t \;=\; \theta(B)\varepsilon_t,$$

where $\{\varepsilon_t\}$ is a white noise process with zero mean and finite variance σ_ε^2. The AR and MA operators are respectively $\phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$ and $\theta(B) = 1 - \sum_{j=1}^{q} \theta_j B^j$. Given a sample of T observations, for ML and REML estimation [e.g., Cheang and Reinsel (2003)] to be implemented efficiently for large T, we need efficient computation of the autocovariances $\gamma(l) = \text{Cov}(Y_t, Y_{t-l})$. Sowell (1992) derived from the spectral density a formula for computing $\gamma(l)$ when the AR polynomial $\phi(B)$ has distinct zeros. Recently, Doornik and Ooms (2003) implemented some refinements to this formula for numerically stable evaluation of the autocovariances. In this note we provide an alternate derivation of $\gamma(l)$, leading to recursive relations that allow $\gamma(l)$ to be evaluated accurately and efficiently. We will see how these autocovariances can be programmed in R for ARIMA(p,d,q) when $p = 0, 1$, and the problems encountered with evaluation of the hypergeometric function required when $p > 1$.

References

Cheang, W.-K., and Reinsel, G. C. (2003). Finite sample properties of ML and REML estimators in time series regression models with long memory noise, *Journal of Statistical Computation and Simulation*, **73**: 233–259.

Doornik, J. A. and Ooms, M. (2003). Computational aspects of maximum likelihood estimation of autoregressive fractionally integrated moving average models, *Computational Statistics and Data Analysis* **42**: 333–348.

Sowell, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models, *Journal of Econometrics* **53**: 165–188.

R and S-PLUS: Similar but different?

Jeff Coombs, CEO, Insightful Corporation

S-PLUS and R both include implementations of the S language, originally developed by Bell Labs (Lucent Technologies) in the 1980's. While there are many similarities between S-PLUS and R, there are also a number of differences. Jeff Coombs, CEO of Insightful Corporation (makers of S-PLUS) will discuss the differences between S-PLUS and R in light of the varying origins, development models and uses of the two packages. Jeff will also discuss ways that developments in R are leading to changes in S-PLUS, and the impact the S language is having on the use of statistics in commercial environments. Jeff will outline alternative business models that combine open source and commercial software and explain the approach that Insightful is taking with S-PLUS and R.

Using R for predicting air traffic

Clara Cordeiro
FCT/UALG, Departamento de Matemática
ccordei@ualg.pt

M. Manuela Neves
ISA/UTL, Departamento de Matemática

February 12, 2004

Abstract

Air traffic had an accentuated increase in the last decade, mainly in the last recent years. The air traffic controllers have to promote an organize flow airplanes; supply useful information and suggestions to promote the security of the flights; prevent collisions between airplanes also prevent collisions between airplanes and obstacles in the ground. With the aim of helping air traffic controllers we intend to get air traffic forecasts, in order to plan and to take decisions anticipated. Therefore, we use R and his packages to help us achieve air traffic forecasts. We obtain forecasts and confidence intervals monthly, for one year, using time series theory and then using the Bootstrap methodology.

Author Keywords: Time series; Forecasting methods; *Bootstrap*; Model-based resampling; Block resampling

Edward M. Debevec and Eric A. Rexstad
Institute of Arctic Biology
University of Alaska Fairbanks
Fairbanks, AK 99775
USA

Creating Online R Functions with deliveR

Web delivery of R functions has many advantages. Updates to the data and the functions themselves only need to happen on the server, users with no R experience can quickly implement the functions without having to learn R, and the functions are accessible worldwide. The disadvantage has been that the developer must be familiar with HTML and a CGI language such as PHP or Perl. At its most basic, an Online R Function (ORF) consists of two files: (1) an HTML web form to input arguments to the function and (2) a CGI script to create an R script that calls the function and handles the output, run the script in batch mode, and then deliver the function results to the user as images (png, jpg, or bmp) and/or files to be downloaded (comma delimited or HTML). We have created a tool called deliveR to generate these two ORF files without any coding required.

deliveR consists of three components: (1) an HTML web form to input details of the function to be delivered online, (2) a CGI file that creates the two ORF files to deliver the function, and (3) a library of R functions that are used by the ORF to create temporary directories, open and close graphics devices, and write output data frames to files. The function to be delivered online does not require any special modification; it can create any number of graphics and/or output a data frame or list of data frames. With the web form, the R developer defines input arguments to the function, whether they will appear in the ORF as text boxes or pull down select boxes, and their default text or select options. The developer also sets whether the output from the ORF will be graphics, data files, or both. When submitted, the CGI script creates the HTML and CGI files previously described that make up the ORF.

We demonstrate the use of deliveR using a suite of ecological analyses that produce multiple graphics and data output. A library of Online R Functions can be quickly created and made available to users worldwide. deliveR is available for Windows and Unix platforms and comes with both PHP and Perl CGI scripts.

Deployment of validated statistical algorithms within an end-user's application

Dr. Mark Demesmaeker, Spotfire Inc., www.spotfire.com

Modern life science research data demands sophisticated analysis. New technologies and new algorithmic developments require close collaboration between research- and statistical organizations. While once it was possible for statisticians and analytic experts to work one-on-one with researchers and their data sets, the pace of research now demands more rapid and broader deployment of analysis software to share analytic expertise across the organization. Increasingly statistical staffs are relying on their ability to influence the software used by researchers as their means of ensuring rigorous validation of experimental findings. Unfortunately, off-the-shelf software rarely contains the statistician's algorithm or statistical method of choice. Statisticians find themselves becoming application developers in order to provide the necessary analytics to their end-user colleagues. End-users struggle with the complexity of analysis applications, used only occasionally but at critical junctures, to validate their findings. This misalignment of resources can decrease the effectiveness of organizations better prepared to derive algorithms than developing and maintaining end-user software.

Spotfire DecisionSite is a visual, analytic application for dynamic, multi-dimensional data analysis. DecisionSite is highly configurable and supports guided analysis in the context of any customer process and data source using analytical tools and built-in data access capabilities.
Use DecisionSite and the R environment to interactively create and test R scripts. These scripts use standard R code and can be parameterized. Using DecisionSite's support for guided analysis, users deploy R scripts as specific end-user data analysis applications. Spotfire DecisionSite users run specific data analysis applications created and deployed to link with an R server. Users are prompted for and enter values for the specific parameters. Results are calculated on the R server and returned to DecisionSite.

The Spotfire Advantage Solution for R provides organizations the ability to deploy algorithms developed in R within a DecisionSite analytic application.

While many statistical tools are freely available, appropriate use of them may be quite difficult for individual users. By linking DecisionSite to an R server through use of the Spotfire Advantage Solution for R, the customer's analytic staff can define specific analysis routines in R and easily make them available to the rest of the organization—for example, various micro-array normalization processes available through Bioconductor.org or a novel high throughput screening
validation routine developed by internal statisticians.

R inside - let R drive your bioinformatic application

Janko Dietzsch, Matthias Zschunke and Kay Nieselt
Junior Research Group Proteomics Algorithms and Simulations
ZBIT (Center for Bioinformatics Tübingen)
University of Tübingen
Sand 14, 72076 Tübingen, Germany
Contact: dietzsch@informatik.uni-tuebingen.de

The emergence of Bioinformatics as a new independent field of science was and is driven by the use of new high throughput methods in life sciences. The huge amount of data produced by these methods causes the need for specialized methods to process, to organize and to visualize the generated data sets. These are the three pillars of the elaborate process of deriving knowledge from the gained information. Based on both its strengths in statistical computing and its rootage in the open source community R has a good chance to play an extraordinary role in Bioinformatics. Projects like Bioconductor already achieved a good acceptance in the research community. But R is a tool primarily designed for professionals and therefore lacks some features that are asked for by users in biological and clinical environments. In particular easy handling via fast reacting GUIs and an excellent visualization is requested by potential users. To deliver improved functionality in these fields we propose to combine R with implementations in an established, well proven language like Java. Java shares with R the independence of the used operating system and has a good support for GUIs, visualization and database connectivity.

We have developed Mayday, short for "MicroarrAY DAta analYsis", a plug-in-based platform in Java to analyse and visualize microarray data. It provides Swing-based GUIs and professional visualization features. On the basis of our plug-in architecture we have implemented a connection to R within the Java environment. With this specialized plug-in we offer advanced users the possibility to access R. The chosen way to connect R and our application does not depend on third party libraries and is independent of the underlying operating system. Since it is possible to generalize the implementation details, it can thus be regarded as a case study for integrating R in other software components and can serve as a template for other projects. Last but not least is it a good example of the application of R in a young, interesting and demanding research field - according to the conference motto: "useR!"

RMAGEML: integrating MAGE-ML format microarray data and Bioconductor

Steffen Durinck, Joke Allemeersch,
Yves Moreau and Bart De Moor

The microarray gene expression markup language (MAGE-ML) is a widely used XML standard for describing and exchanging information about microarray experiments. It can describe microarray designs, microarray experiments designs, gene expression data and data analysis results. Bioconductor is an open source project that provides a framework for the statistical analysis of genomic data in R. These R packages provide a wide range of microarray data analysis tools. Up till now it was not possible to import data stored in MAGE-ML format in Bioconductor. Because of the importance of both MAGE-ML and Bioconductor in the field of microarray data analysis, this acute gap had to be filled. We describe RMAGEML, a new Bioconductor package that provides a link between MAGE-ML format microarray data and Bioconductor. The current version enables MAGEML-import to the limma and marray Bioconductor packages. RMAGEML is available at http://www.bioconductor.org

Using R on Debian:
Past, Present, and Future

Douglas Bates
bates@stat.wisc.edu

Dirk Eddelbuettel
edd@debian.org

Albrecht Gebhardt
albrecht.gebhardt@uni-klu.ac.at

Submitted to useR! 2004

Abstract

In this paper, we discuss the R language and environment, and in particular its use on Debian Linux, as well as the Debian package management system.

1 Introduction

More and more research, both applied and theoretical, in physical, biological, and social sciences is being performed with computers. Researchers prefer to concentrate on their research, rather than on the provision of a suitable computing infrastructure. To this end, a 'nuts-to-bolts' distribution of software can help by providing high quality and up-to-date binary packages ready for immediate use. In this paper, we describe one such system: the R environment and language for 'programming with data', and in particular its availability for the Debian Linux distribution.

This paper is organized as follows. We first introduce the R environment and language, before examine its usage on a Debian system. The next section discusses creation of CRAN-based Debian packages in detail, before we conclude with a short review and an outlook.

2 About R

R (R Development Core Team, 2004) provides a powerful computing environment for doing statistical analysis, computation and graphics. It provides tools ranging from simple exploratory methods to complex modelling and visualisation functions. These tools can be used at the command-line, as well as via some graphical user interfaces. R is one implementation of the well established S language. One important feature of R is its extensibility. Users can write their own functions either in the S language, or (mostly for performance reasons, or to link with external libraries) code written in C, C++ or Fortran can be compiled and linked to R in a straightforward manner. Both R and compiled code can be made available in the form of so-called R packages. A large (over 320 as of February 2004) and fast-growing collection of contributed packages is available through the Comprehensive R Archive Network, or CRAN for short, which provides a series of mirrored file repositories around the globe.

R has been available for the Debian GNU/Linux distribution since the 0.61 release in 1997. Based on the original source code, several binary packages are provided which allows a Debian user to install all components of the R system, or just a subset. While that is similar to a Windows user opting to install only parts of the initial download, it is in fact based on separate binary packages.

Beyond the core R packages, additional utilties such as ESS (the 'Emacs Speaks Statistics' mode for a particularly powerful type of text editors) and the Xgobi (and now Ggobi) data visualization frontends have been made available, as have been other programs such as the postgresql-plr package which provides R as a procedural language embedded in the PostgreSQL relational database management system.

Given the large, and growing, extent of packages of CRAN, it was only a matter of time before individual packages would be integrated into Debian. This has happened over the course of the last year, see the Appendix for a current list of packages.

The contribution of this paper is to outline the future direction of an integration of CRAN packages into Debian – either directly within Debian, or via an archive suitable for `apt-get` hostes on the CRAN mirrors. The setup describe here should also be suitable for an use with other code repositories (built on top of R) such as the BioConductor project.

3 So even if we use R, why with Debian?

Different users of Debian would probably give different answers to this question. However, non-users of Debian are often converging on a single answer: the (real or perceived) difficulty of the Debian installation process. The next section discusses the interplay between an easier initial installation versus an easier long-term maintenance and upgrade path. Ideally, a computing platform such as a Linux installation should excel in both aspects.

In the context of R and Debian, it may also be worthwhile to point out that a significant portion of the infrastructure of the R Project, including the main CRAN host, is running n Debian systems.

3.1 Installation versus longer-term administration

It has been said that a large number of Linux users get their first experiences by installing a distribution with a strong focus of ease-of-use during installation such as SuSE, Mandrake, RedHat or others. Some of these users may experience, over the course of a few years and after some upgrade/reinstallation cycles, that maintaining a system can be tedious, and even prone to failures requiring a full reinstallations. It is in this area that Debian is very clearly recognised for its ease of maintenance and upgradeability of a Debian GNU/Linux system.

Other advantages of Debian are the large collection of available packages – as of February 2004, about 8150 source packages with 13900 binary packages are reported by `http://www.debian.gr.jp/~kitame/maint.cgi` based on packages in the development branch of Debian. A further advantage is the robust and powerful mechanism for determining package inter-dependencies, an aspect that can become troubling on other types of systems (see also the next section) yet which has worked extremely well for Debian leading to a reputation for reliability. Debian is also a distribution supporting a wider variety of different hardware architectures: currently, ten different platforms ranging from x86 to S/390 are supported.

3.2 Why provide Debian packages of R packages?

One reason for providing a Debian package of an R package is to use Debian package dependencies to ensure that any system libraries or include files required to compile the code in the R package are available. For example, the Debian postgresql-dev package must be installed if the R package Rpgsql is to be installed successfully. By providing the meta-information of required packages in a control file, the build process can be largely automated.

The second reason is for ease of maintenance, as we first mentioned above. Someone who already uses Debian tools such as `apt-get` to update the packages on a Debian system may find installing or updating a Debian package to be more convenient than installing the r-base Debian package plus learning to update R packages from within R or externally using R CMD INSTALL. Because R is beginning to be used more widely in fields such as in biology (e.g. Bioconductor) and social sciences, we should not count on the typical user being an R guru. Having R packages controlled by `apt-get` seems worth the small amount of overhead in creating the Debian packages. This also applies to systems maintained by (presumably non-R using) system administrators who may already be more familiar with Debian's package mechanism. By using this system to distribute CRAN packages, another learning curve is avoided for those who may not actually use R but simply provide it for others.

The third reason is quality control. The CRAN team already goes to great length, including un-supervised nightly builds, to ensure the individual quality and coherence of an R package. Embedding a binary R package in the Debian package management system provides additional control over dependencies between required components or libraries, as well as access to a fully automated system of 'build daemons' which can recompile a source package for up to ten other architectures – which provides a good portability and quality control test.

The fourth reason is scalability. More and more users are using several machines, or may need to share work with co-workers. Being able to create, distribute and install identical binary packages makes it easier to keep machines synchronised in order to provide similar environments.

The fifth reason plays on Debian's strength as a common platform for other 'derived' systems. Examples for Debian derivatives include entire distributions such as Lindows or Libranet, as well as Knoppix and its own derivatives such as Quantian. Providing Debian packages of R packages allows others to use these in entirely new environments.

4 Debianising CRAN: How ?

As noted above, R itself has been a part of Debian since 1997. Binary Debian packages of contributed R libraries have been created since March 2003, starting with RODBC and tseries. Currrently, several Debian maintainers working more-or-less individually provide a variety of CRAN packages totalling about thirty-five packages (see the Appendix for a list). A proposed Debian R Policy (Bates and Eddelbuettel, 2003) is aiding in keeping these package in a coherent and homogeneous form.

Also, most Windows users rely on the R packages (currently maintained by Uwe Ligges) for their operating system distributed at CRAN. SuSE Linux users can grab their packages from CRAN. The increasing work load with packaging especially these SuSE Linux packages lead one of us (A. Gebhardt) some years ago to switching to a more automated process. This resulted in a Perl script[1] which does the whole job of building the RPM files. This script is still in use by Detlef Steuer who is currently in charge of being the package builder for contributed package for the SuSE Linux distribution.

More recently, this script has been modified[2] to perform a similar task for the Debian community. It performs the following steps:

1. Retrieve an up-to-date package list (`ftp://cran.r-project.org/pub/R/contrib/main/PACKAGES`).

2. Retrieve the library description files (`ftp://cran.r-project.org/pub/R/contrib/main/Descriptions/*.DESCRIPTION`).

3. Parse these description files using the `R::Dcf` Perl module.

4. Determine a package dependency graph based on the "Depends:" statements found in the description files.

5. Generate a Debian package build structure (according to the Debian R package policy).

6. Finally build and install the Debian packages in correct order (bottom to top in the dependency graph).

This process relies much on the format of the DESCRIPTION files of the individual libraries. It can cope automatically with inter-library dependencies. However, it can not yet deal with dependencies on non-R software, e.g. external programs like xgobi or GRASS. In these cases, some help is required from a hard-coded dependency list that can be kept in a small 'knowledge base', either as simple flat-file or a small RDBMS. The build process involves an extra run of "`R CMD check`" and insures that way that only correctly working packages will be distributed.

[1] Available at `http://cran.r-project.org/bin/linux/suse/build-R-contrib-rpms.pl`.

[2] The new version is available at `http://www.math.uni-klu.ac.at/~agebhard/build-R-contrib-debs.pl`.

It is our intention to extend the meta-information in such a way that we should be able to provide unattended, automatic creation of Debian packages for CRAN. This could then be provided as a side-effect of the exiting quality control builds at CRAN where all packages are already (re-)build every night in order to ensure and control code and packaging quality.

5 Experiences and Outlook

The 'Debianisation' of R may be perceived as being overly confusing to new users due to the schere number of components – just as the whole of Debian may be with its 13,000 packages. While this may be true in the narrow sense, we feel that the added flexibility of being able to customize and adapt an installation is worth the marginal difficulty it may add. We would also argue to few truly novice users ever install an entire system from scratch. Just as it is said that 'nobody is installing Windows' (given the prevalence of purchasing computers with an operating system pre-installed), few new users will attempt to go from a blank hard disk to a working system. On the other hand, we feel that more experienced users do in fact value the added flexibility.

For more than six years, the core R packages have been a part of Debian. The experiences from that process are very positive and encouraging. As R grows, and more and more user-contributed packages are added to the CRAN repositories, it becomes desirable to provide a similar level of quality and robustness for the contributed packages as there is for the core parts of R.

The advantage of prebuilt base and contrib R debian packages is clearly the ease of installation, maintenance and upgradeability. It will not only reduce adminstrative efforts in central networked installations, but also simplify the process of tailoring specific Knoppix based CD images which e.g. can be handed out to students. Quantian[3] is one example of a Knoppix-derived 'live cdrom' that makes use of both R and the existing Debian R and CRAN packages. Another one is the Debian based campus wide GNU/Linux installation at Klagenfurt University. A subset of this installation is also available as CD image[4] containing first versions of the above mentioned `r-cran-*` Debian packages.

Providing R for Debian has been a rewarding experience. Anecdotal evidence suggests that Debian is used for R work in variety of educational, industry and government institutions all of which should benefit from having the very rich set of CRAN packages only one command away.

References

Douglas Bates and Dirk Eddelbuettel. Debian R Policy: Draft proposal, 2003. URL http://lists.debian.org/debian-devel-0312/msg02332.html.

BioConductor. BioConductor: Open Source Software for BioInformatics, 2004. URL http://www.bioconductor.org.

CRAN. CRAN: The Comprehensive R Archive Network, 2004. URL http://CRAN.R-project.org.

Debian. Debian Project, 2004. URL http://www.debian.org.

R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2004. URL http://www.R-project.org. ISBN 3-900051-00-3.

[3] http://dirk.eddelbuettel.com/quantian.html

[4] ftp://ftp.uni-klu.ac.at/pub/unikluKNOPPIX.iso

A Existing packages

```
edd@homebud:˜> apt-cache search "ˆr-(.*cran|omegahat)-" | sort
r-cran-abind - GNU R abind multi-dimensional array combination function
r-cran-boot - GNU R package for bootstrapping functions from Davison and Hinkley
r-cran-car - GNU R Companion to Applied Regression by John Fox
r-cran-cluster - GNU R package for cluster analysis by Rousseeuw et al
r-cran-coda - Output analysis and diagnostics for MCMC simulations in R
r-cran-dbi - database interface for R
r-cran-design - GNU R regression modeling strategies tools by Frank Harrell
r-cran-effects - GNU R graphical and tabular effects display for glm models
r-cran-foreign - GNU R package to read / write data from other statistical systems
r-cran-gtkdevice - GNU R Gtk device driver package
r-cran-hmisc - GNU R miscellaneous functions by Frank Harrell
r-cran-its - GNU R package for handling irregular time series
r-cran-kernsmooth - GNU R package for kernel smoothing and density estimation
r-cran-lattice - GNU R package for 'Trellis' graphics
r-cran-lmtest - GNU R package for diagnostic checking in linear models
r-cran-mapdata - GNU R support for producing geographic maps (supplemental data)
r-cran-mapproj - GNU R support for cartographic projections of map data
r-cran-maps - GNU R support for producing geographic maps
r-cran-mcmcpack - routines for Markov Chain Monte Carlo model estimation in R
r-cran-mgcv - GNU R package for multiple parameter smoothing estimation
r-cran-nlme - GNU R package for (non-)linear mixed effects models
r-cran-qtl - [Biology] GNU R package for genetic marker linkage analysis
r-cran-rcmdr - GNU R platform-independent basic-statistics GUI
r-cran-rmysql - MySQL interface for R
r-cran-rodbc - GNU R package for ODBC database access
r-cran-rpart - GNU R package for recursive partitioning and regression trees
r-cran-rquantlib - GNU R package interfacing the QuantLib finance library
r-cran-statdataml - XML based data exchange format (R library)
r-cran-survival - GNU R package for survival analysis
r-cran-tkrplot - GNU R embedded Tk plotting device package
r-cran-tseries - GNU R package for time-series analysis and comp. finance
r-cran-vr - GNU R package accompanying the Venables and Ripley book on S
r-cran-xml - An XML package for the R language
r-noncran-lindsey - GNU R libraries contributed by Jim and Patrick Lindsey
r-omegahat-ggobi - GNU R package for the GGobi data visualization system
r-omegahat-rgtk - GNU R binding for Gtk
```

Programming with financial data: Connecting R to MIM and Bloomberg

Dirk Eddelbuettel[*]
edd@debian.org

Submitted to useR! 2004

Abstract

In this paper, we discuss uses of R for 'programming with data' in an investment banking trading floor environment. We outline connector packages that permit direct access to both the Market Information Server (MIM), and the Bloomberg data servers.

1 Introduction

Trading floor environments require access to historical market data. In our case, the principal market data warehouse for data on historical security prices, as well as various related time series, contains several hundred thousand series comprising several gigabytes of content.

R, an environment for 'programming with data', to use the title of one of the defining books on the S language, is an extremely approriate choice for research analysts charged with producing empirical work for support of sales and trading. R can really shine in such an environment as one of its core strengths is — by design, it should be stressed — on interactive, or exploratory, data work with a particular focus on compelling visual representation of relationships in the data. However, the traditional platform in a trading room environment, arguably more by default than choice, is based on spreadsheets of often enormous size (and hence some complexity). Users who are new to R are often frustrated with the difficulty of having to extract data from one system (e.g. a database) before transforming it in order to import it into another system (e.g. R). Direct access to the data can aid tremendously in this regard. The D(C)OM server by Thomas Baier (c.f. `http://cran.r-project.org/contrib/extra/dcom/`) is of some use in this area too. However, the focus of this paper is on slightly more platform-independent ways to access data directly from other data repositories.

The Market Information Server, or MIM for short, is a powerful database backend provided by Logical Information Machines (`http://www.lim.com`) or LIM for short. The MIM database is a hierarchical database system (as opposed to a relational database system queried by SQL) which offers very fast access to data in time series form. Another key aspect of the MIM system is the ability to query the data in a particular language which attempts to be both expressive and relatively easy for non-programmers. The potential for work in the S language with data stored in MIM has not been lost on LIM who is offering a plugin for S-PLUS as well, c.f. `http://www.lim.com/partners/splus_example.htm`.

Bloomberg (`http://www.bloomberg.com`) provides a very powerful 'terminal' with access to an incredible number of data series, as well as (financial) functions to price and analyse an immense number of different security types. Bloomberg is widely regarded as a benchmark by market particants in debt, equity, foreign exchange and commodity markets.

This paper illustrates two in-house packages currently under development. These packages already permit highly efficient connections from R to both of these data services. We discuss the basic C language interface using the R `.call` function interface, as well as portability aspects between Windows and Solaris. A data quality control and validation example details one actual application before a summary concludes.

[*] The views expressed in this paper are those of its author, and do not necessarily reflect the positions or policies of his employer. The paper describes research by the author and is issued to elicit comments and to further discussion.

Quantian:
A single-system image scientific cluster computing environment

Dirk Eddelbuettel
edd@debian.org

Submitted to useR! 2004

Abstract

This paper introduces the openMosix extensions to the Quantian environment for quantitative and scientific computing. Quantian, originally based on Knoppix technology, allows one to boot virtually any recent commodity i386-class computer from a single cdrom containing a compressed iso image into a fully-configured graphical workstation equipped with over 2gb of software – including 500mb of applications with a focus on quantitative or scientific computing. With the additional foundation of ClusterKnoppix providing support for openMosix-based clustering, Quantian allows one to create ad-hoc single-system image computing clusters that are already loaded with a large number of software environments commonly used in quantitatively-oriented disciplines.

1 Introduction

Quantian (Eddelbuettel, 2003a) is a directly bootable and self-configuring Linux system based on a compressed cdrom image. Quantian is an extension of Knoppix (Knopper, 2001) and clusterKnoppix (Vandersmissen, 2003).

The Knoppix bootable Linux system on a single cdrom (also referred to as a so-called "live cdrom") provides a complete workstation with over two gigabytes of software along with fully automatic hardware detection and configuration.

ClusterKnoppix extends the basic Knoppix system with a kernel containing the openMosix patch, user-space utilities and several additional tools. If used on a local-area network with other computers running the same kernel version and openMosix patch (but possibly of different processor architecture), an auto-discovery mechanism permits these machines to instantaneously form a single-system image computing cluster. In such a cluster, any machine can be both a client or a server. In addition, any such node can act as network-boot server to additional nodes which could be booted via PXE from the initial machine.

To this powerful computing-cluster system, Quantian adds software with a quantitative, numerical or scientific focus: several computer-algebra systems are included, as are higher-level matrix languages, data visualization tools, and a variety of libraries and scientific applications. A particular focal point is the R (R Development Core Team, 2003) language and environment for which the base system as well as several add-on tools are installed.

This paper is organized as follows. In the next section, we briefly describe the single-system operation based on a Knoppix-style "live cdrom". In the following section, we detail how the ClusterKnoppix-based automatic setup of computing clusters can transform a set of personal computers into a powerful computing cluster. An example application illustrates Quantian before a summary concludes the paper.

2 Single-machine use of Quantian

Quantian (Eddelbuettel, 2003a) is a directly bootable cdrom using the technology introduced by the Knoppix system (Knopper, 2001). While the fundamental design and setup of Quantian, and its underlying Knoppix base, has been described by Eddelbuettel (2003b), we will review the key points before we outline the unique features added by Quantian.

2.1 The foundation provided by Knoppix

It is useful to reiterate some of the key points of the Knoppix (Knopper, 2001) 'live cdrom' system:

- usable on virtually any recent desktop or laptop;
- available for system recovery and forensics due to a large number of included utilities and tools for network and system administrators;
- usable in computer labs as machines can be rebooted quickly into identical states;
- can be used by students as no system administration skills are required for setup and installation;
- lowers barriers to entry for Linux and Unix technology as it provides a ready-to-use system;
- requires little retraining as the KDE desktop environment provides a familiar user experience to other common operation system interfaces;
- provides a terminalserver mode allowing other netboot-capable machines (potentially with hard-disk and/or cdrom) to be initialized and booted using the PXE protocol (or via etherboot) in a thin-client environment;
- provides a 'persistent home' mode where data can be written to USB storage devices (or disk partitions) to preserve states between sessions;
- permits one to try Linux risk-free as no information is written to the hard disk, or existing operating system;
- enables to try Linux on new hardware to reliably test its compatibility.

Knoppix is under active development and released several times a year.

2.2 Quantian contributions

The first two Quantian releases were based directly on Knoppix. In order to add quantitatively-focused software, existing programs have to be removed. The list of software that is removed from the base system is determined with the dual objectives of a) creating a maximum amount of capacity and b) removing applications with limited usefulness in a quantitative analysis setting. Consequently, we remove several large applications such as openoffice (which is also removed from the most-recent Knoppix versions), mozilla and gimp, internationalization packages, games, the bochs and wine emulators, several additional window managers, some networking tools as well as a few miscellaneous applications. In total, about 500mb of software are removed.

The base system stills contains around 1.5gb of software including a complete KDE environment with its window manager, browser, office suite, development environment and editors as well as a large number of other general-purpose tools, utilities and diagnostic applications.

Quantian then adds various sets of applications from different areas:

mathematical such as the giac, ginac, maxima, gap, pari and yacas computer-algebra systems, as well as the euler and xppaut applications;

statistical such as the R language and environment for 'programming with data' (along with several packages from CRAN, the Emacs Speaks Statistics (or ESS) mode for Emacs and XEmacs, and the Ggobi data visualization tool) as well as autoclass, gretl, mcl, multimix, x12a and xlispstat;

visualization such as the OpenDX and Mayavi visualizers, as well as gnuplot, grace, gri, xfig and plotutils for scientific plotting;

libraries such as the GNU Scientific Library (or GSL) and QuantLib, a quantitative finance library for risk management and trading;

matrix environments such as Octave, a matrix programming language and interactive environment included along with several add-on packages, scientific and numeric python, the perl data language (PDL) and yorick;

typesetting systems such as the lyx and kile frontends to LaTeX, as well as the auctex mode for XEmacs;

editors such as XEmacs, a general purpose programming editor and tool along with several add-on applications via their supplied elisp code, and the TeXmacs WYSIWYG mathematical editor and interface to several languages and systems;

python along with a variety of scientific, numerical or general-purpose Python libraries;

miscellaneous scientific programs such as aplus, aribas, ent, euler, evolver, felt, freefem, gambit, geg, geomview, ghemical, glp, gmt, gperiodic, ipe, lp-solve, lush, mpb, mpqc, and rasmol.

This unique combination of scientific applications, combined with the "live cdrom" technology from Knoppix enables any standard personal computer to be transformed into a scientific workstation.

3 Quantian and openMosix Single-System Image clustering

In the initial Quantian presentation, Eddelbuettel (2003b) conjectured that permitting several machines running Quantian to be combined in a single-system image openMosix-style cluster may be a natural extension. Shortly thereafter, the first releases of ClusterKnoppix became available. By basing Quantian on ClusterKnoppix instead of the standard Knoppix system, Quantian has added a new dimension: openMosix clustering support. The next section discusses ClusterKnoppix.

3.1 ClusterKnoppix contributions

ClusterKnoppix (Vandersmissen, 2003) extends Knoppix by combining it with an openMosix-enabled kernel. OpenMosix is a Linux kernel extension for single-system image (SSI) clustering. This SSI kernel extension turns a network of ordinary computers into a larger 'virtual' computer for Linux applications that presents itself to the user as a single, more powerful computer rather than a collection of machines.

Once openMosix is enabled and started, nodes in the cluster can start talking to one another while continuously attempting to optimize the resource allocation by migrating processes from 'busy' nodes to 'spare' nodes in order to split the total computing load evenly across the cluster. The resulting system is approximately linearly scalable in the number of nodes. Moreover, with the openMosix auto discovery mechanism, a new node can be added while the cluster is running. The cluster will automatically begin to use the new resource allowing for dynamic sizing of the cluster as well as some level of fault-tolerance.

Applications do not need to be programmed specifically for openMosix (as opposed to Beowulf systems which require explicit communication protocols such as MPI). Since all openMosix extensions are inside the kernel, every Linux application automatically and transparently benefits from the distributed computing concept of openMosix. The cluster behaves much like a Symmetric Multi-Processor (SMP) computer, but scales to well over a thousand nodes (which could themselves be SMPs for a larger number of total processors).

ClusterKnoppix provides these openMosix features in a framework with the following key benefits:

- includes openMosix terminal server using PXE, DHCP and tftp permitting linux clients to boot via the network (which permits to use nodes without hard disk, cdrom or floppy);
- operates openMosix in autodiscovery mode so that new nodes automatically join the cluster minimising the need for configuration or administration;
- contains cluster management tools such as openMosixview;
- setup such that every node has root access to every other node via ssh using RSA-encrypted keys;
- provides Mosix / Direct File System (MFS/DFSA) support which enables all nodes to see each others files;

- permits the choice for each node to be run as a graphical workstation (lab/demo setup) or as a text-only console conserving system memory.

The most recent version of clusterKnoppix also adds the tyd service from the CHAOS project Latter (2003a,b). CHAOS focuses on adding a virtual private network (VPN) layer on top of open network protocols in order for the openMosix-style cluster to be used securely on public networks. To this end, CHAOS includes the FreeSWAN kernel patches.

3.2 Quantian and openMosix

By providing the scientific applications listed in section 2.2 in the framework of an openMosix-enabled “live cdrom”, Quantian offers a synthesis of three distinct domains:

1. Knoppix auto-configuration and ease of installation on virtually any recent desktop or laptop;
2. a large number of analytically-focused applications pre-loaded and configured ready for deployment;
3. openMosix extensions to the Linux kernel permitting approximately linearly-scaled performance increases with additional nodes in a single-system image cluster.

4 Application example

To be done. Maybe use Monte Carlo simulation or Bootstrap example.

5 Summary

To be done.

References

CRAN. The Comprehensive R Archive network. URL `http://cran.r-project.org`.

Dirk Eddelbuettel. Quantian, 2003a. URL `http://dirk.eddelbuettel.com/quantian`.

Dirk Eddelbuettel. Quantian: A scientific computing environment. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 2003b. ISSN 1609-395X.

Klaus Knopper. Knoppix, 2001. URL `http://www.knopper.net/knoppix/index-en.html`.

Ian Latter. `http://itsecurity.mq.edu.au/chaos`, 2003a.

Ian Latter. Security and openmosix: Securely deploying ssi cluster technology over untrusted networking infrastructure. white paper, Macquarie University, 2003b.

openMosix. openmosix, 2002. URL `http://openmosix.sourceforge.net`.

R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2003. URL `http://www.R-project.org`. ISBN 3-900051-00-3.

Wim Vandersmissen. Clusterknoppix, 2003. URL `http://bofh.be/clusterknoppix`.

Structured Bradley-Terry models, and flat lizards fighting

David Firth
University of Warwick
http://www.warwick.ac.uk/go/dfirth

Abstract

The Bradley-Terry model (Bradley and Terry, 1952; Agresti, 2002) is a useful device for the scoring of 'tournaments' for which the data are the results of 'contests' between pairs of 'players'. Applications are many, ranging from bibliometrics (Stigler, 1994) in which the 'players' are academic journals, to genetics (for example, the allelic transmission/disequilibrium test of Sham and Curtis (1995) is based on a Bradley-Terry model in which the 'players' are alleles). The Bradley-Terry model in its simplest form, with no ties permitted, is a logistic regression with a specially structured design matrix: in any contest between players i and j,

$$\text{logit}[\text{pr}(i \text{ beats } j)] = \lambda_i - \lambda_j.$$

A simple elaboration allows an 'order' effect, for example to allow one player in each contest to enjoy an advantage on account of 'going first', or 'playing on home turf':

$$\text{logit}[\text{pr}(i \text{ beats } j \text{ in contest } t)] = \lambda_i - \lambda_j + \delta z_t,$$

where $z_t = 1$ if i has the supposed advantage and $z_t = -1$ if j has it. (If the 'advantage' is in fact a disadvantage, δ will be negative.) The scores λ_i then relate to ability in the absence of any such advantage.

Motivated by study of a large 'tournament' among male flat lizards (*Platysaurus broadleyi*) in the wild, we consider structured forms of the Bradley-Terry model in which the ability parameters $\lambda_1, \ldots, \lambda_K$ are determined by measured attributes of the K players involved, *viz.*

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir} \quad (i = 1, \ldots, K).$$

Special attention is needed in the case of contests involving any lizard i whose explanatory values $x_{i1}, \ldots, x_{ip}$ are incomplete.

This talk will describe the *BradleyTerry* package for *R* (Firth, 2004), which provides facilities for the specification and fitting of such models, for model selection by standard methods, and for model criticism *via* special-purpose residuals. Use of the package will be illustrated using the lizard data ($K = 77$), analysis of which reveals biologically-important evidence on the role played by bright colours on a male lizard's body.

(Arising from joint work with M J Whiting, D M Stuart-Fox, D O'Connor, N Bennett and S Blomberg.)

References

Agresti, A. (2002). *Categorical Data Analysis* (Second edition). Wiley.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika 39*, 324–45.

Firth, D. (2004). *Bradley-Terry models in R.* Submitted for publication.

Sham, P. C. and D. Curtis (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of Human Genetics 59*(3), 323–336.

Stigler, S. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science 9*, 94–108.

Computational cost and time threads

Xavier Font Aragonés
Escola Universitària Politècnica de Mataró
Av. Puig I Cadaflach 101-111
08303 Mataró (Barcelona) - Spain
font@eupmt.es

This paper investigates some of the well known general problems a Ph.D. student usually finds when is conducting a research. First, we concentrate on efficient algorithm construction and second the simulation and generation of a survey tables. Both situations under examination take an example from the field of time series analysis and filtering.

All the paper turns around using R as a statistical environment to help us focus on the real statement of the research.

The starting point of the first part shows some interesting improvement expressed in a mathematical fashion that has to be evaluated within a computer programming language (R). Here three different approaches will be presented. The first approach comes from the direct translation of the mathematical formulation into our user software platform. The second approach takes the first one and makes a vectorization (where it is possible) and the third approach tries to avoid repeated or unnecessary evaluations.

The second part stands on extracting meaningful conclusions about the three algorithms and with the problem of generating tables of the results. It shows how to select the best algorithm and how to automate the generation of tables in order to add then in your latex document.

The conclusion drawn about the first part of the paper takes an easy expression: "*do not make computer work as much as you would*". The second conclusion, close to the first one visualizes the need of: "*do not loose your time with simulation and with the effort of translating the best simulation ratios into your latex tables if someone can do it for you*".

Relaxing Error Assumptions on Dynamic models

Xavier Font Aragonés
Escola Universitària Politècnica de Mataró
Av. Puig I Cadaflach 101-111
08303 Mataró (Barcelona) - Spain
font@eupmt.es

The state space representation and its application on time series analysis provides us with new ways for analyzing time series. From pioneering first research by Akaike (1974) to most recent work Kitagawa (1998), Durbin (1998) or Arulampalam (2002), researchers have worked based on state space representation and have tried to model situations, where time series goes from simple stationary situations to the most complex non-stationary, non-linear and non-gaussian situations. Recently there has been a great interest in particle filtering methods to perform filtering and prediction in general state space models. Our goal shifts the interest through the usual assumptions on error distributions (related with the system and observation equation).

A way to obtain filtering densities without imposing assumptions on error distributions by means of kernel density estimation (kde) is presented. The advantage of this kind of solution is clear: self adaptation of the estimated densities and a wides range of possible ways to perform this estimation.

The paper presents three ways for dealing with kde. The firs one comes from the fact that once all the data has been processed, estimated errors will define the data used to estimate the new density. Then we repeat the filtering process with the new estimated densities and again with all the data. This process goes on until we get to the best filtering density. To overcome the problem of too much iteration with all data a second approach which tries to reduce the kde step just for the data available is presented. This means that once all the data has been processed we obtain the estimated densities. The last approach takes previously knowledge or the need to impose gaussian assumptions.

Getting Started With the R Commander:
A Basic-Statistics Graphical User Interface to R

John Fox*
McMaster University
Hamilton, Ontario, Canada

2 January 2004

Abstract

Unlike S-PLUS, R does not include a statistical graphical user interface (GUI), but it does include tools for building GUIs. Based on the `tcltk` package (which furnishes an interface to the Tcl/Tk GUI builder) the `Rcmdr` package provides a basic-statistics graphical user interface to R called the "R Commander."

The design objectives of the R Commander were as follows: to support, through an easy-to-use, extensible, cross-platform GUI, the statistical functionality required for a basic-statistics course (though its current functionality has grown to include support for linear and generalized-linear models); to make it relatively difficult to do unreasonable things; to render visible the relationship between choices made in the GUI and the R commands that they generate.

The R Commander uses a simple and familiar menu/dialog-box interface. Top-level menus include *File*, *Edit*, *Data*, *Statistics*, *Graphs*, *Models*, *Distributions*, and *Help*, with the complete menu tree given in the paper. Each dialog box includes a *Help* button, which leads to a relevant help page.

Menu and dialog-box selections generate R commands, which are recorded in a simple log/script window and are echoed, along with output, to the R session window. The log/script window also provides the ability to edit and re-execute commands.

Data sets in the R Commander are simply R data frames, and may be read from attached packages or imported from files. Although several data frames may reside in memory, only one is "active" at any given time.

The purpose of this paper is to introduce and describe the basic use of the R Commander GUI and the manner in which it can be extended.

*Please address correspondence to `jfox@mcmaster.ca`. The work described in this paper was supported by a grant for the Social Science Research Board of McMaster University.

UBS Investment Bank

Portfolio modelling of operational losses

John Gavin[1]

February 2004

Abstract

Basel II is an emerging regulatory regime for financial institutions. It stipulates that banks must quantify the capital they need to offset operational losses, such as fraud, computer viruses or transaction processing errors. One statistical approach is to assume that risk can be described in terms of a distribution of possible outcomes, over some time horizon, such as one year. The capital charge for operational losses is then based on the difference between the 99.9th percentile and the mean of this distribution.

The distribution of aggregated portfolio losses is a compound distribution, combining loss event frequencies and severities. A key assumption is that historic operational events contain some information about future potential operational risk exposure. So frequency and severity distributions are independently parameterised using this historic data. Then Monte Carlo simulation is used to convolute the distributions. All calculations are implemented in R.

In this paper, the frequency distribution is a negative binomial and the severity distribution is semi-parametric, combining the empirical distribution, for losses below some high threshold, with a generalized Pareto distribution, for excesses above that threshold.

[1] Quantitative Risk Models and Statistics, UBS Investment Bank, 100 Liverpool St., London, EC2M 2RH. U.K.

Bayesian Methods in Geostatistics: Using prior knowledge about trend parameters for Kriging and Design of Experiments

Albrecht Gebhardt[*]
agebhard@uni-klu.ac.at

Claudia Gebhardt[†]
cgebhard@uni-klu.ac.at

Submitted to useR! 2004

Abstract

This paper covers an approach to incorporate prior knowledge about trend parameters into Kriging estimation. The implementation in R covers prediction as well as monitoring network design.

1 Introduction

Kriging, the central method for spatial prediction of a regionalised variable $Z(\underline{x}), \underline{x} \in \mathbb{R}^d$, is based on some prerequisites: translation invariant covariance structure (second order stationarity)

$$\mathrm{Cov}\,(Z(\underline{x}), Z(\underline{x}+\underline{h})) = C(\underline{h}), \quad \underline{h} \in \mathbb{R}^d \tag{1}$$

and knowledge about an underlying trend. Depending on the existence or absence of a trend function one has to choose between Ordinary or Universal Kriging. Additional prior knowledge about the trend parameter can be incorporated into the prediction using a Bayesian approach. This approach has been implemented in R for prediction and monitoring network design purposes.

2 Bayesian Kriging

Given a data set $\underline{Z} = (Z(\underline{x}_i))_{i=1,\ldots,n}$ and using a linear trend model

$$Z(\underline{x}) \quad = \quad \underline{f}(\underline{x})^\top \underline{\theta} \; + \varepsilon(\underline{x}), \qquad \underline{\theta} \in \Theta \subseteq \mathbb{R}^r \tag{2}$$

Universal Kriging takes a weighted average $\hat{Z}(\underline{x}_0) = \underline{w}^\top \underline{Z}$ of the data as estimator at location $\underline{x}_0$. It assumes unbiasedness $\mathrm{E}\,\hat{Z}(\underline{x}_0) = \mathrm{E}\,Z(\underline{x}_0)$ which leads to the so called universality constraint $\mathbf{F}^\top \underline{w} = \underline{f}_0$. Choosing the weights $\underline{w}$ which minimise the variance of the prediction yields finally the Universal Kriging estimator

$$\hat{Z}(\underline{x}_0) \quad = \quad \underline{c}_0^\top \mathbf{C}^{-1}(\underline{Z} - \mathbf{F}\hat{\underline{\mu}}) + \underline{f}_0^\top \hat{\underline{\theta}} \tag{3}$$

using the notations $\underline{f}_0 = \underline{f}(\underline{x}_0)$, $\mathbf{F} = (\underline{f}(\underline{x}_1), \underline{f}(\underline{x}_2), \ldots, \underline{f}(\underline{x}_n))^\top$, $(\mathbf{C})_{ij} = C(\underline{x}_i - \underline{x}_j)$ and $(\underline{c}_0)_i = C(\underline{x}_i - \underline{x}_0)$ $i, j = 1, \ldots, n$ where $\hat{\underline{\theta}} = (\mathbf{F}^\top \mathbf{C}^{-1}\mathbf{F})^{-1}\mathbf{F}^\top \mathbf{C}^{-1}\underline{Z}$ corresponds to the generalised least squares estimator of $\underline{\theta}$.

[*] University of Klagenfurt, Institute of Mathematics, Austria

[†] University of Klagenfurt, University Library, Austria

Bayesian Kriging requires knowledge about the prior distribution of the trend parameter θ. This approach (see Omre (1987)) makes only use of the first two moments $\mathrm{E}\,\underline{\theta} = \underline{\mu}$ and $\mathrm{Cov}\,\underline{\theta} = \mathbf{\Phi}$. In contrast to Universal Kriging the condition of unbiasedness gets now thrown away in favour of a modified Bayesian Kriging estimator involving a bias component w_0

$$\hat{Z}(\underline{x}_0) \;=\; \underline{w}^\top \, \underline{Z} + w_0 \; . \tag{4}$$

Again the weights are chosen to minimise the prediction variance. The solution of the corresponding Lagrange system is

$$\hat{Z}(\underline{x}_0) \;=\; \tilde{\underline{c}}_0^\top \, \tilde{\mathbf{C}}^{-1} \, (\underline{Z} - \mathbf{F}\underline{\mu}) + \underline{f}_0^\top \underline{\mu} \tag{5}$$

which involves the following modified covariance terms

$$\tilde{C}_0 = C_0 + \underline{f}_0^\top \, \mathbf{\Phi} \, \underline{f}_0 \quad \tilde{\underline{c}}_0 = \underline{c}_0 + \mathbf{F} \, \mathbf{\Phi} \, \underline{f}_0 \quad \tilde{\mathbf{C}} = \mathbf{C} + \mathbf{F} \, \mathbf{\Phi} \, \mathbf{F}^\top \; . \tag{6}$$

3 Monitoring network design

Monitoring networks evolve over time. After starting with an initial set of measurement sites it is a common task to expand (or shrink) this network in an optimal way. Kriging does not only return estimated values but also the associated prediction variance. As this prediction variance only depends on the (planned) measurement locations it is possible to use G- and I-optimality criteria minimising maximum or mean prediction variance to choose an optimal subset fo a given set of candidate measurement points.

4 Implementation Details and Example

The implementation (see Gebhardt (2003)) is primarily based on the `sgeostat` library. In a first step a library `rgeostat` implementing universal kriging estimation in Fortran code based on LAPACK subroutines using the same data structures as in `sgeostat` has been written. Another library `baykrig` implements the above shown Bayesian approach based on empirical prior data (see Pilz (1991)). Finally library `kdesign` puts all things together and provides functions for optimal network design based on complete enumeration among a finite set of candidate measurement points. The usage of these libraries will be demonstrated in a short example.

Preliminary versions of these libraries are available at `ftp://ftp.uni-klu.ac.at/pub/R`.

References

C. Gebhardt. *Bayessche Methoden in der geostatistischen Versuchsplanung.* PhD thesis, University of Klagenfurt, 2003.

H. Omre. Bayesian Kriging - merging observations and qualified guesses in kriging. *Mathematical Geology*, 19:25–39, 1987.

J. Pilz. Ausnutzung von a-priori-Kenntnissen in geostatistischen Modellen. In G. Peschel, editor, *Beiträge zur Mathematischen Geologie und Geoinformatik*, volume 3, pages 74–79, Köln, 1991. Verlag Sven von Loga.

The Extreme Value Toolkit (extRemes):
Weather and Climate Applications of Extreme Value Statistics

Eric Gilleland, Rick Katz and Greg Young

The Extremes toolkit, **extRemes**, is part of the recent Weather and Climate Impact Assessment Science Initiative undertaken at the National Center for Atmospheric Research (NCAR). The toolkit reflects an effort to develop software for the fitting of meteorological extremes in a form accessible to the broader atmospheric science community. The software used by the toolkit for fitting meteorological extremes was provided by Stuart Coles and ported into R by Alec Stephenson as the R package **ismev**. The primary exciting new applications in R are graphical user interface (GUI) dialogs via **tcltk** for the ismev functions (plus a few new functions) with an accompanying tutorial on how to use the toolkit, with examples of modeling meteorological extremes. The package has been implemented and tested on unix, linux and Windows operating systems. Although the emphasis of the toolkit, and particularly the accompanying tutorial, is on extreme weather and climate events, the toolkit itself can be used with any sort of data where extreme value methods are appropriate.

An important part of the initiative focuses on extreme weather and climate events, and the toolkit is integral to meeting the goals for this component of the initiative. Specifically, it is hoped that the toolkit will prompt more use of (i) extreme value methodology when analzying weather and climate extremes, (ii) information about upper tail of distribution (via point process approach instead of block maxima), (iii) climate change detection by way of extreme value methodology with trends introduced through covariates and (iv) development of more physically realistic statistical models for extremes (via covariates for annual and diurnal cycles as well as for physical variables such as El Niño events).

To accomplish these goals, the toolkit uses GUI dialogs to make it as easy as possible for scientists not familiar with R to get started and have access to extreme value software with a small learning curve; provides an in-depth accompanying tutorial to introduce scientists who may be unfamiliar with extreme value methods to some of the basic principles; and through both the tutorial and the design of the software facilitates easy adaptation of the code for specific problems.

So far the toolkit only handles univariate data, except that covariates may be incorporated into model parameters. One possible future addition to the toolkit would be to add software that can directly support multivariate models; because weather and climate studies generally occur spatially and it is desired to incorporate spatial dependence into the models.

Creating graphs of meta-analyses with R

Christian Gold

Meta-analysis is a type of research aiming at statistically summarising information from several previous studies. It is becoming increasingly important in the medical field as the amount of information available increases [1, 2]. Commercially available computer programmes for meta-analysis usually offer limited flexibility, and the quality of graphs is sometimes poor. R offers almost unlimited flexibility in creating high-quality graphs.
I show how I used R to create a meta-analysis graph [3] that, while presenting information in a fairly standard way, had some idiosyncratic features that would have been impossible to deal with when using standard meta-analysis software (such as Comprehensive Meta-Analysis [4], but also Thomas Lumley's new R package rmeta). A major drawback of our procedure is that as high as its flexibility is, as low is its user-friendliness – it would have to be re-programmed for every new study or even update of this study.
I conclude that R is an excellent tool for creating the type of high-quality individualised graphs that are needed when submitting completed research to a scientific journal. However, there is room for improvement concerning ease of use, and there is a gap to fill between readymade, easy-to-use but relatively inflexible software and pure programming language. The package rmeta is a very welcome step in this process, but there is also a need for more generalised packages or functions that combine flexibility with ease of use.

References:
1. Cooper, H., & Hedges, L. (Eds.). (1994). *The handbook of research synthesis*. New York: Russel Sage.
2. The Cochrane Collaboration (2002). *Cochrane reviewers' handbook* (4.1.5.). Available: http://www.cochrane.org/software/Documentation/Handbook/handbook.pdf.
3. Gold, C., Voracek, M., & Wigram, T. (in press). Effects of music therapy for children and adolescents with psychopathology: A meta-analysis. *Journal of Child Psychology and Psychiatry and Allied Disciplines*.
4. Borenstein, M., & Rothstein, H. (1999). *Comprehensive meta-analysis: A computer program for research synthesis* (Version 1.0.23). Englewood, NJ: Biostat.

Author's address:

Christian Gold, PhD
Faculty of Health Studies
Sogn og Fjordane University College
6823 Sandane
Norway
Phone: +47-57866834
Fax: +47-57866801
Email: christian.gold@hisf.no

`BayesMix`: An R package for Bayesian Mixture Modelling

Bettina Grün[*], Friedrich Leisch
Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien
Wiedner Hauptstraße 8-10, 1040 Wien, Österreich
{Bettina.Gruen, Friedrich.Leisch}@ci.tuwien.ac.at

Finite mixture models are a popular method for modeling unobserved heterogeneity as well as for parametric approximations of multimodal distributions. Areas of application are, e.g., biology, medicine, economics and engineering among many others. For ML estimation the EM algorithm is most frequently used which can be done in R, e.g., with the package `mclust`.

Bayesian estimation has become feasible with the advent of Markov Chain Monte Carlo (MCMC) simulation and the R package `BayesMix` provides facilities for estimating univariate Gaussian finite mixtures with MCMC methods. It has been developed as accompanying material to the forthcoming book Frühwirth-Schnatter (2005).

The model class which can be estimated with `BayesMix` is a special case of a graphical model where the nodes and their distributions are fixed and the user only needs to specify the values of the constant nodes, the data and the initial values. Small variations of the model are allowed with respect to the segment specific priors. The MCMC sampling is done by JAGS (Just Another Gibbs Sampler; Plummer, 2003) and its output can be analyzed in R using functionality from the package `coda`. In addition to the visualization of the MCMC chains there are diagnostic plots implemented which can be used for determining the appropriate number of segments or a suitable variable for ordering the segments as in Bayesian mixture modelling it makes in general a difference which constraint is imposed for ordering the segments due to label switching.

`BayesMix` can be seen as a prototype for a special purpose interface to the software JAGS. Its advantage is that a user who "only" wants to estimate finite mixtures of Gaussian distributions can use JAGS as sampling engine, but does not need to know the BUGS syntax which is used by JAGS for specifying general Bayesian hierarchical models. `BayesMix` offers the opportunity to be a starting point for learning the BUGS syntax as the model specifications are written into separate files and can be inspected or modified by the user.

Keywords: Finite mixture models, Bayesian modelling

References

S. Frühwirth-Schnatter. *Bayesian Mixture Modelling.* Springer, 2005. Forthcoming.

M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, and A. Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Technische Universität Wien, Vienna, Austria, 2003. URL `http://www.ci.tuwien.ac.at/Conferences/DSC.html`. ISSN 1609-395X.

[*] supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modeling in Economics and Management Science').

THE STATISTICAL LAB
UNDERSTANDING STATISTICS

Free University Berlin – Center for Digital Systems (CeDiS) –
Competence Center e-Learning / Multimedia
http://www.statistiklabor.de

2004-04-14

The Statistical Lab - Teaching and Learning Statistics

(Albert Geukes, Christian Grune, Negar Razi)

Overview & Functionality

The Statistical Lab as a working and learning environment supports teachers and students in elementary statistics' studies. It facilitates working with abstract statistical questions by offering a graphical user interface. The Lab supports students and teachers in:

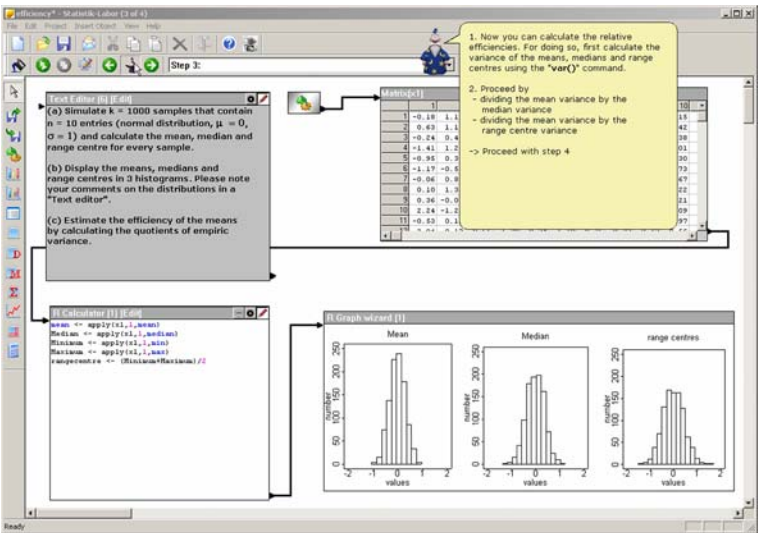
- creating and solving statistical exercises
- carrying out statistical calculations and analysis
- exploring different computational approaches for statistical problems
- easily visualizing data
- composing statistical reports

The Statistical Lab has been designed for usage throughout the entire teaching process. It can be used for data presentation, for individual learning at home, and as working environment in courses and workshops. The Lab has also been applied as examination tool.

The Statistical Lab can be extended and flexibly adjusted to meet specific needs by using the statistical programming language R: The Lab provides the "R Calculator" as co-operative interface to R. This interface is based on release level 1.4.1 of R and leaves R unchanged.

Idea & Concept

The Statistical Lab has been created to sustainable improve the statistical education at universities. One of its most important aims is to promote and support creative thinking by utilizing constructive elements in a step-by-step solution process.



(a) Simulate k = 1000 samples that contain n = 10 entries (normal distribution, µ = 0, σ = 1) and calculate the mean, median and range centre for every sample.

(b) Display the means, medians and range centres in 3 histograms. Please note your comments on the distributions in a "Text editor".

(c) Estimate the efficiency of the means by calculating the quotients of empiric variance.

1. Now you can calculate the relative efficiencies. For doing so, first calculate the variance of the means, medians and range centres using the "**var()**" command.

2. Proceed by
- dividing the mean variance by the median variance
- dividing the mean variance by the range centre variance

-> Proceed with step 4

```
mean <- apply(x1,1,mean)
Median <- apply(x1,1,median)
Minimum <- apply(x1,1,min)
Maximum <- apply(x1,1,max)
rangecentre <- (Minimum+Maximum)/2
```

The Statistical Lab consequently focuses on manipulation and analysis of real data instead of one-dimensional methods, symbols, formulas and abstract equations. Authentic statistical questions can easily be processed. The open structure of the working space supports individual approaches to these problems. Understanding theoretical problems is made easier by interactive statistical experiments and simulations, which can be adapted to the individual needs of teachers and students.

Most steps of data analysis are made visible by user interface objects. These objects can be datasets, frequency tables or graphical diagrams. In addition to these standard objects, the Statistical Lab offers numerous possibilities for individual experimentation. Almost every R function and user-made extensions (so called userlibs) can be accessed via an interface to R.

History & Future development

The Statistical Lab's core functionality has been continuously developed since 1997 at Free University of Berlin. A network of German universities (Free University, University Hamburg and University Bielefeld) has used the Lab since 1998 in the context of the project "DIALEKT: Statistik interaktiv!" in regular curricular activities. This project was supported by the German Research Network (DFN).

In 2000 a succeeding project started as a consortium of 10 German universities, again headed by the Center for Digital Systems (FU) and supported by the Federal Ministry of Education and Research: "Neue Statistik" (New Statistics). An early version of the Statistical Lab used a self-developed S-like "Lab Definition Language" (LDL) to enable statistical programming. In the context of "Neue Statistik" the Lab's statistical engine has been changed. Recent versions of the Lab, now implemented in Visual C, have an interface to an unmodified version of R 1.4.1. This interface enables statistical computing and user programming.

The change to R has been made for technical and "strategic" reasons: Since the early implementations of the Lab were based on Visual Basic, performance and stability were poor. Moreover, the growing demand among our project partners for an open statistical programming interface in order to expand the functional capabilities of the Statistical Lab was another strong motivation.
The project "Neue Statistik" ended in 2003. Together with old and new, national and international partners we are now going to establish the Statistical Lab in academic statistical education.

Future steps may consist of cooperation between the communities of the Statistical Lab and R in several fields. There can be found many points of common interests for professional statistical computing as well as for problem-based teaching statistics with real data. So it may improve both the usability and accessibility of R for beginners and make the Statistical Lab more powerful for experts' needs.

References

Statistical Lab's Homepage
http://www.statistiklabor.de

Homepage of the project "Neue Statistik"
http://www.neuestatistik.de

Center of Digital Systems at FU Berlin
http://www.cedis.fu-berlin.de

Federal Ministry of Education and Research
http://www.bmbf.de

German Research Network (DFN)
http://www.dfn.de

Why R Project Succeeded?

A Comparison of Open Source Software Projects Using R.

Yutaka Hamaoka

hamaoka@fbc.keio.ac.jp

Faculty of Business and Commerce

Keio University

Though thousands of open software projects (OSSPs) are initiated, most of them fail. On the contrary, R project succeeded to develop mature software, to build user community, and to keep active user-to-user interaction. This research compared 2,000 OSSPs, including R project, to explore key factor of success of OSSPs. Publicly available data, e.g., messages posted to mailing lists, CVS log files that track development activities, were compiled. Based on social network theory, coordination among developers, communication pattern between users, and growth of community were examined.

PRSG_Nirvana — A parallel tool to assess differential gene expression

David Henderson

PRSG_Nirvana is a statistical package to assess statistical significance in differential gene expression data sets. The basis of the method is a two stage mixed linear model approach popularized in a fairly recent publication by R. Wolfinger. The package takes advantage of the SNOW and Rmpi packages to create a parallel computing environment on common commodity computing clusters. Without the added computing power of cluster computing, bootstrapping of residuals to utilize the empirical distribution of a test statistic would not be practical. One would then have to rely on asymptotic thresholds for less than large data sets as inference is based upon gene specific observations. Additional features of the package are a custom version of the now ubiquitous shrinkage estimators for individual gene variances and a dependent test false discovery rate adjustment of bootstrap p-values based upon work by Benjamini and Yekutieli. The performance of the package for computational speed is compared to other single processor compiled packages. The resulting lists of genes from each method are also compared using a small example data set where results are assumed known.

spatialreg — A package to fit spatial regression models with isotropic and anisotropic covariance functions

Nadine Henkenjohann, Department of Statistics
University of Dortmund, Germany

Abstract

In regression analysis the assumption of uncorrelated residuals is sometimes untenable. The feature of correlated data should be investigated and integrated in the modelling and analysis of the data. Spatial regression models (SRMs) are an appropriate choice to depict the structure of such data.

The classical optimization approach in Design of Experiments facilitates the use of second-order models. If rather complex relationships are to be expected, the benefit of second-order models is limited. In such cases, a further appealing property of SRMs is revealed. These models provide smooth, data-faithful approximations of complex response functions based on a relatively small number of design points.

The SRM is defined as a linear mixed model that omits random effects but incorporates correlated errors. Restricting attention to isotropic covariance functions, a convenient way to fit such a model in R is to apply the function *glm* in the package *nlme* (Pinhero and Bates, 2000). If the assumption of isotropy is too restrictive, anisotropic covariance functions improve the fit of the model. SRMs including these functions are estimable with the R-package *spatialreg*. The two geometrically anisotropic and separable covariance functions implemented in this package were introduced by Sacks, Welch, et al. (1989) and Zimmermann and Harville (1991). A capability of the *spatialreg* package is the fitting of SRMs independently of the dimension of the data. For the deterministic mean model one can choose between polynomials of various degrees. In order to estimate the parameters of the covariance function, the classical ML method and the residual maximum likelihood method (REML) are available. An additional feature of the package is the visualization of the predicted response surface for two non-fixed variables. Examples are provided which show the flexibility of the SRM to approximate complex relationships.

References

O'Connell, M. A. and Wolfinger, R. D. (1997). Spatial Regression Models, Response Surfaces, and Process Optimization. *Journal of Computational and Graphical Statistics*, 6, 224-241.

Pinheiro, J. C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York, Springer-Verlag.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and Analysis of Computer Experiments. *Statistical Science*, 4, 409-435.

Zimmermann, D. L. and Harville, D. A. (1991). A Random Field Approach to the Analysis of Field-Plot Experiments and Other Spatial Experiments. *Biometrics*, 47, 223-239.

Microeconomic Analysis with R

Arne Henningsen
Department of Agricultural Economics
University of Kiel, Germany

Since its first public release in 1993, the “R language and environment for statistical computing” (R Development Core Team, 2003) has been used more and more for statistical analyses. However, in the first years it was not much used by economists and econometricians, but this situation has been changing in recent years. One cornerstone was the article “R: Yet another econometric programming environment” by Cribari-Neto and Zarkos (1999). Three years later Racine and Hyndman (2002) published the article “Using R to teach econometrics”. And Arai (2002) has a “A brief guide to R for beginners in econometrics” in the web that has been updated and improved several times within the past one and a half years.

Over the last years the number of R packages that are useful for economists also increased. One of these packages is called “systemfit” (Hamann and Henningsen, 2004) and provides functions to estimate systems of linear or non-linear equations. Many economic models consist of several equations, which should be estimated simultaneously, because the disturbance terms are likely contemporaneously correlated or the economic theory requires cross-equation restrictions. This is especially the case in microeconomic modeling.

We extended the “systemfit” package to make it suitable for microeconomic modeling (e.g. incorporating cross-equation restrictions). Subsequently, we used it for several microeconomic demand and production analyses. The demand analyses were done with the “Almost Ideal Demand System” (AIDS) (Deaton and Muellbauer, 1980) and the production analyses with the “symmetric normalized quadratic” profit function. On the useR! conference we want to demonstrate this on a poster and on a laptop computer. Furthermore a first release of a new R package will be presented that contains functions for microeconomic modeling (e.g. a function that carries out a full demand analysis with the AIDS with only a single command). Applied economists interested in microeconomic modeling will be invited to contribute to this package by providing functions for other functional forms.

References

Arai, M. (2002). A brief guide to R for beginners in econometrics. http://people.su.se/~ma/R_intro/R_intro.pdf.

Cribari-Neto, F. and Zarkos, S. G. (1999). R: Yet another econometric programming environment. *Journal of Applied Econometrics*, 14:319–329.

Deaton, A. and Muellbauer, J. (1980). An Almost Ideal Demand System. *The American Economic Review*, 70:312–326.

Hamann, J. D. and Henningsen, A. (2004). systemfit - simultaneous equation estimation package for R. http://cran.r-project.org/src/contrib/PACKAGES.html#systemfit.

R Development Core Team (2003). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

Racine, J. and Hyndman, R. (2002). Using R to teach econometrics. *Journal of Applied Econometrics*, 17:175–189.

Irregular Time Series in a Financial Risk Model

Giles Heywood

In financial market data, measurements that are equally spaced in time are the exception rather than the rule. The irregularities result from a number of sources, including order flow, market opening and closing times, timezone differences, weekends and holidays, and calendar irregularities, to name but a few. Depending on the nature of the analysis, these irregularities can represent anything form a minor irritation to a vital part of the analysis.

The **its** package consists of a single S4 class that extends the matrix class, plus a number of methods and functions for handling irregular time series. In this paper we give an overview of the package and a case study of its application to a financial problem concerned with missing data, in part arising from the irregularity of the raw data available. The case study is intended to illustrate the facilities for handling irregular time series, rather than the merits of different statistical methods for handling missing data.

The case study considers the use of the **its** package to solve a realistic practical problem in finance, namely asynchronous holidays. Starting with a panel of daily time-series data for a set of fifty European stocks, we compute the historical covariance matrix, and from that estimate a multi-factor risk model by maximum likelihood. Holidays that are pan-European are excluded entirely from the analysis, but for national holidays, the 'missing' price is estimated using the remaining data plus least squares.

Multi-state Markov modelling with R

Christopher Jackson
Department of Epidemiology and Public Health
Imperial College, London

April 15, 2004

A *multi-state model* expresses the movement of an individual between a finite set of states. It is most commonly used in the form of a *Markov model*, where the movement follows a Markov process. It can be seen as an extension of survival or time-to-event models, where there are several events.

A major application of multi-state models is investigating the progression of chronic diseases. It is of interest to estimate average transition rates between disease states, and also to investigate explanatory variables for the rates of transition. Usually the data are only observed as a series of snapshots of the process in continuous time, such as irregular doctor or hospital visits. Then we do not see when each transition occured, only a set of intervals in which certain transitions must have occurred.

The *hidden Markov model* is a powerful extension to the basic multi-state model. This means that the underlying progression of the individual through the states is not observed, while the observed data are generated conditionally on the underlying states. This can be useful in disease screening applications, where the state of disease is only observed through some error-prone biological marker, and the screening test is subject to error.

The R package `msm` [1] can be used to fit continuous-time multi-state Markov models to irregularly observed longitudinal data, using likelihood maximisation. Any form of transition intensity matrix can be estimated, as long as there are data available to identify the model. Transition rates can be modelled in terms of explanatory variables through a proportional-intensities model. `msm` also handles a special case of the hidden Markov model, where the observed states are misclassifications of the true states. Interested users are encouraged to download `msm` from CRAN, try it out and offer suggestions for improvement.

References

[1] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society, Series D: The Statistician*, 52(2):1–17, 2003.

Opening the treasury of R language to a wider scientific community: *GCDkit*, a package for interpretation of geochemical data

V. JANOUŠEK[1], C. M. FARROW[2], V. ERBAN[3]

[1] *Institut für Mineralogie, Universität Salzburg, Hellbrunnerstraße 34, A-5020 Salzburg, Austria,*
vojtech.janousek@sbg.ac.at

[2] *Computing Service, University of Glasgow, Glasgow G12 8QQ, Scotland,*
c.farrow@compserv.gla.ac.uk

[3] *Czech Geological Survey, Klárov 3, 118 21 Prague 1, Czech Republic,*
erban@cgu.cz

Similarly to many branches of science, the flood of numerical data on the one hand, and the dearth of potent and flexible software for their recalculation and plotting on the other, is often a limiting factor to a creative work in igneous geochemistry. The Geochemical Data Toolkit for Windows *(GCDkit)* is our answer to the challenge, taking advantage of the tools available in R and exploiting its functions facilitating Windows-like interaction. The current version does:

1. Offer an alternative, more user-friendly interface to powerful statistical and graphical functions built in R. The R is provided with a command line interface, which on the one hand allows excellent control for experienced users but on the other tends to discourage many scientists, accustomed to software with graphical user interface. All functions of *GCDkit* are accessible via pull-down menus, as well as in an interactive regime.
2. Provide core routines for effortless data management, i.e. loading and saving of free form text files, copying data from/to clipboard, data editing, searching and generation of subsets using regular expressions and Boolean logic.
3. Allow the analyses to be organised into unique groups, which are subsequently utilised by the statistical and plotting functions. This can be done on the basis of various attributes (locality, rock type,…), ranges of a numerical variable, by cluster analysis or using a selected classification diagram.
4. Contain flexible high-level graphical functions. Most of the diagrams are defined as templates for *Figaro*, a set of graphical utilities for R. *Figaro* provides a means to create figure objects, which contain both the data and methods to make subsequent changes to the plot (zooming and scaling, adding comments or legend, identifying data points, altering the size or colour of the plotting symbols…). The templates can be used also as a basis for classification; the general algorithm looks for the name of the polygon within the diagram, into which the analysis falls according to its x–y coordinates.
5. Offer brand new calculation, modelling and plotting tools designed specifically for geochemists.
6. Permit expansions by the end users. Text files containing R code defining new calculation or plotting options when placed in the sub-directory *Plugin* are automatically loaded at the system start-up. These can be made accessible via newly appended menu items.
7. Avoid any licensing problems. The *GCDkit* is distributed as freeware via the WWW; the current version can be downloaded from *http://www.gla.ac.uk/gcdkit.*

The whole *GCDkit* system, which is modular and straightforward to modify, provides potentially a platform for DIY additions written by R literate geochemists for their less fortunate colleagues. Our mission is to amass eventually a platform-independent version of *GCDkit*, using Tcl/Tk-based interface on Unix/Linux and Macintosh (System X). Moreover we intend to extract a core of the system (without geochemical bias), creating a simple GUI front-end to R, intended specifically for use in natural sciences.

This project was supported by the grant FWF 15133–GEO.

`metrics`: Towards a package for doing econometrics in R

Hiroyuki Kawakatsu
School of Management and Economics
25 University Square
Queen's University, Belfast
Belfast BT7 1NN
Northern Ireland
`hkawakat@qub.ac.uk`

January 20, 2004

Abstract

This paper proposes a package `metrics` for doing econometrics in R. The discussion is in two parts. First, I discuss the current state of `metrics`, which is just a small collection of some useful functions for doing econometrics in R. Second, I discuss how the `metrics` package should evolve in the future.

The `metrics` package currently contains four main functions: robust (heteroskedasticity and/or autocorrelation consistent) standard errors, general (nonlinear) hypothesis testing, linear instrumental variables (IV) estimation, and maximum likelihood estimation of binary dependent variable models. These are the minimum necessary functions I needed to use R for teaching an undergraduate level econometrics class. I discuss current implementation and example usage of these functions in `metrics`. The key features of these functions are as follows. The heteroskedasticity and autocorrelation consistent (HAC) covariance supports data-based automatic bandwidth selection rules of Andrews (1991) and Newey and West (1994). The hypothesis test function `wald.test` provides an interface where users specify restrictions as R `expressions`. The Wald χ^2 statistic is computed via the delta method using the symbolic (or "algorithmic") derivative routine `deriv` in the base package. The IV estimator is implemented as a linear GMM estimator, providing robust standard errors and an over-identification test statistic. The binary dependent variable models are estimated by maximum likelihood using hard-coded analytic derivatives. A variety of options for computing the asymptotic covariance matrix is available and can be feeded into `wald.test` for general hypothesis testing.

As for the future of `metrics`, I identify several aspects of econometric analyses for which an interface needs to be developed for R to be of general use to econometricians. These include handling of panel or longitudinal data sets and a general interface for GMM and ML estimation with support for a variety of inference procedures.

JEL classification: C61, C63.

Keywords: R, econometrics.

Proposed Abstract for the UseR Symposium Vienna May

FLR, A Framework For Fisheries Management In R Using S4 Classes

Philippe Grosjean, Laurence Kell, David Die, Robert Scott and Jose De Oliveira.

The various models for assessment of fisheries dynamics and evaluation of management strategies are currently implemented in separate software programs and their respective input and output formats are often incompatible although many are performing similar tasks. Most of these packages provide basic analysis tools (model estimation, graphing, result reporting) that are already available in various software platforms. Comparing the results of such models is difficult and requires exporting them to an environment that has more efficient analytical tools. Moreover, integration of such different models into a single simulation environment that allows evaluation of the whole fishery system has been impossible.

The EC project "FEMS" (Framework for Evaluation of Management Strategies), currently in its second year, has decided to use R, a common, feature-rich environment, both to run fishery models and to analyse their output. The latest object-oriented features of R (named S4 objects, or classes) allow for the definition of complex and flexible objects with a structure and arithmetic that is appropriate to fishery models. R also, allows access to objects (fishery models) already written in C/C++ or FORTRAN and recompilation of these objects into the R environment using a wrapper.

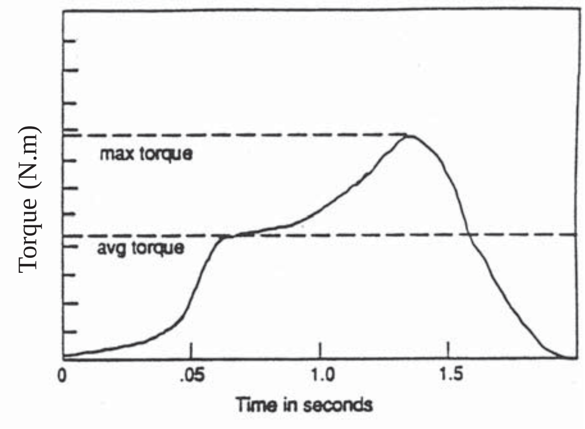
Currently FEMS has implemented selected key components of the framework including, FLQUANT, a flexible data object with key fishery dimensions (time, age, space and stock) and for example FLSTOCK, a collection of FLQUANTs for selected biological properties for a population (weight, catch, survival).

The current implementation of the FLR library has proved to be convenient, flexible and capable of using fisheries models in R. The FLR framework is currently been evaluated by international fisheries agencies, including the International Council for the Exploration of the Sea (ICES) and the International Commission for the Conservation of Atlantic Tunas (ICCAT). If this evaluation is successful, FLR may become an ideal environment for the evaluation of different fisheries models and management structures within a simulation framework including all the relevant components of the fishery system.

Performance of Spinal Supporting Muscles under Different Angles of Spinal Flexion for 4 Combination of Flexion/Extension and Sitting / Standing Positions.

Keyhani, M. Reza* (MSc. Biomefry); Dr. Farahini, Hossain** (Orthopedist);
Ebrahimi, Ismaeel* (PhD. Physiotherapy); Taheri, Navid *** (MSc. Physiotherapy)

Low back pain is a common problem throughout the world. To provide, or evaluate, a preventive, or therapeutic, exercise program, a physiotherapist should be familiar with the functional biomechanics of spine and muscles supporting it. This study was designed, as an explanatory research, to provide necessary information about the performance of spinal supporting muscles, when the lumbar spine flexes at angles of 0, 20, 40 and 60 degrees in four combination of flexion/ extension and sitting/standing positions. To measure the performance of the muscles, a modified ISO-Station 200 – B instrument was used . This instrument measures the performance via maximun and average torques.



max. torque & avg.torque are calculated by ISO-Station 200-B

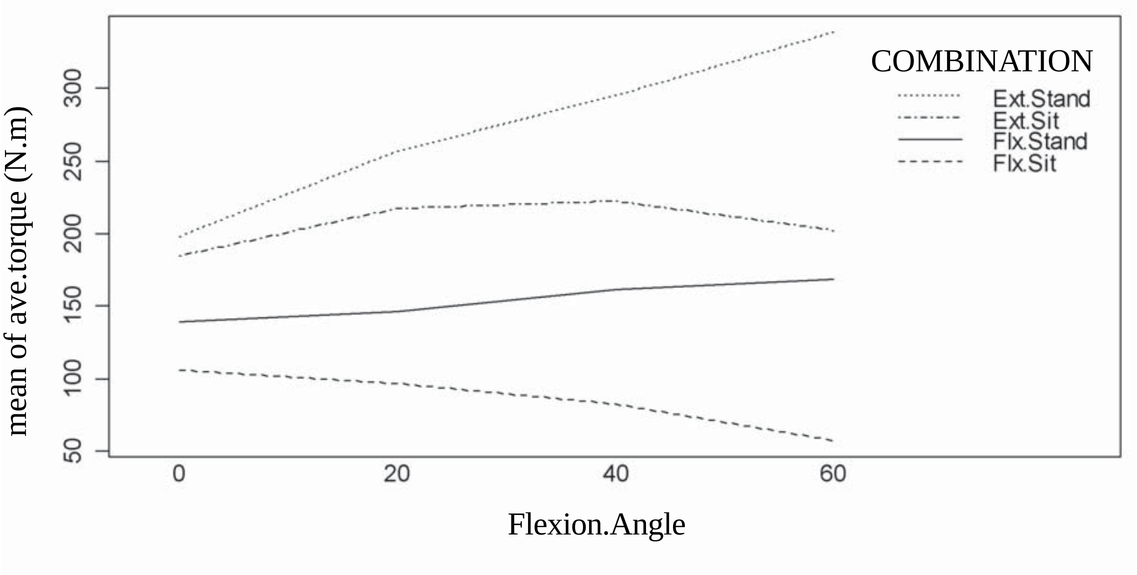
Behavior of max.torques was the same as ave.torques; hence the latter was selected as the response variable of the study. Explanatory variables of the study are:

(1) Flexion.Angle of the lumbar spine having values of 0, 20, 40 & 60 degree.
(2) Standing position with values: 1 = Standing , 0 = Sitting.
(3) Extension position with values: 1=Extension, 0 = Flexion.
(4) COMBINATION with values: Ext.Sit, Flex.Sit, Ext.Stand, Flex.Stand.

This is a repeated measure problem where the response variable is measured at 16 different conditions for each participant. Participants were 30 healthy male students with age range of 20-30 years.

Besides various summary statistics & graphs, basically we can explain the whole picture of the study, by using three R- functions:

(1) interaction.plot (Flexion.Angle, COMBINATION, ave.torque)
(2) summary (aov (ave.torque~ (Flexion.Angle + Stand + Ext)^2 + Error(participant)))



* School of Rehabilitation Sciences – Iran Medical Sciences University – Tehran – Iran
** School of Medicine – Iran Medical Sciences University – Tehran – Iran
*** School of Rehabilitation Sciences – Isfahan Medical Sciences University – Isfahan – Iran

Anova for Repeated Measures and Skewed Response Variable: Two Examples from Medical Fields

Keyhani, M.Reza*(MSc Biometry);Roohi-Azizi,Mahtab*(MSc Human Physiology);Ebrahimi,Ismaeel*
(PhD,Dean of SRS);Milani,Mina*(MSc Audiology);Zowghi,Elaheh*&Talebi,Hossain*(BSc Audiology)

I- Effects of Velocity and Direction of Exerting Pressure on Ear Canal Admittance of a Tympanogram

A. Research Problem: A tympanogram is the plotted acoustic admittance measures(mho) versus exerted ear canal pressure (Pa).It is plotted for each person and has many parameters. One of them is ear canal admittance (**ECA**).

The aim is to study the behavior of ECA (**response variable**) under the following conditions (**explanatory variables**):

(1) **gender :** 2-levels, **M**=male, **F**=female . (2) **ear :**2-levels, **R**=right ear, **L**=left ear.
(3) **direction** (of exerting pressure) **:** 2-levels,**U**=Up(min to max),**D**=Down(max to min).
(4) **prssure.velocity :** 4-levels,**50,100,200,400** (daPa/s).

This is a repeated measure problem with **16** observation per case (4x2x2) i.e. , pressure.velocity * direction * ear.

B. Result Expectations: As ECA is also equivalent ear canal volume; It is expected that its behavior is constant under different conditions, except for gender.

C. R-Solution

a: Computational Solution: aov() with “error term”=Error(cases)

As a repeated measures problem we have to create the column of **cases(**1 to 60),each value being repeated 16 times. With no interaction found, the final format of **aov( )** is:

> aov(ECA~(pressure.velocity+gender+direction+ear+Error(cases))

b: Graphical Solution: interaction.plot( )

Three categorical variables of gender, ear and direction were combined into a single factor, named **GED** with eight levels. Therefore, we have the following format:

> interaction.plot(pressure.velocity,GED,ECA)

*****

II. Residual Analysis of a Skewed Response Variable

A. Research Problem: There is a lymphocyte ratio of CD4 (helper) counts to CD8 (suppressor or killer) counts. It is used to evaluate the immune status of patients.

The aim is to show the behavior of CD4/CD8 ratio (**CD4.CD8** is our **response variable**) for three types of pulmonary diseases, i.e. we have a 3-level factor variable (**disease**).

B. Result Expectations: It is expected that average value of CD4/CD8 ratio is different for different pulmonary diseases; and hence this ratio can be used as a differential test among these three types of pulmonary diseases.

C. R-Solution

a. Computational: aov( ) , glm(…,family=Gamma,…)

Since the response variable is a ratio, its distribution is highly skewed. The assumption of normality is violated;further,the residuals from a simple anova are far from being normal.

With using glm(…,family=Gamma,…), the distribution of residuals becomes normal.

> summary(result.glm<-glm(CD4.CD8 ~ disease, family=Gamma))

> summary(result.aov<-aov(CD4.CD8 ~ disease))

b. Graphical: hist( ),boxplot( )

> **hist(CD4.CD8);hist(result.glm$resid);hist(result.aov$resid)**

> boxplot(CD4.CD8 ~ disease)

* School of Rehabilitation Sciences (SRS); Iran Medical Sciences University,Tehran-Iran
contact <more_keyhani@hotmail.com>

Distributions of finite and infinite quadratic forms in R

Christian Kleiber
Universität Dortmund, Fachbereich Statistik
Vogelpothsweg 78
D-44227 Dortmund, Germany
`kleiber@statistik.uni-dortmund.de`

Distributions of quadratic forms in normal random variables, or equivalently, distributions of weighted sums of χ^2 random variables, arise in a number of situations.

Distributions of finite quadratic forms occur in connection with power and robustness studies in the linear model, for example when one is interested in the size of the F test when disturbances are nonspherical. They are also encountered in connection with the determination of exact critical values for certain tests, a prominent example is the Durbin-Watson test.

Distributions of infinite quadratic forms arise in diagnostic checking of linear models, for instance in testing for structural change against random coefficient alternatives, and also in testing for stationarity against integrated alternatives in time series econometrics. Further applications include goodness of fit tests, for example the Cramér-von Mises and Anderson-Darling statistics.

We provide R functions for the evaluation of the CDF of these distributions. Users may choose between two routines: (i) the numerical method of Imhof (*Biometrika*, 1961) and (ii) the saddlepoint approximation of Kuonen (*Biometrika*, 1999). Procedures for the computation of power functions of some diagnostic tests in the linear model will be made available in a forthcoming version of the R package `lmtest`.

Statistical Process Control and R: A cautious beginning

Sven Knoth

Department of Statistics, European University Viadrina, Frankfurt (Oder), Germany

A lot of monographs were written about the area of statistics called Statistical Process Control. Moreover, in nearly all commercial statistical packages some SPC routines are included. Up to now, it is not possible to call a R function for computing, e. g., the Average Run Length (ARL) of an EWMA control chart. Additionally, R provides a fine framework for timely incorporating new results of computing SPC relevant quantities. Thus, it is time for designing a SPC package for R. The talk will give an impression of a very early state of a SPC package, which allows to compute the (zero-state) ARL and steady-state ARL of EWMA and CUSUM control charts. Naturally, the ideas and solutions have to be discussed more intensively than for mature R packages and each SPC and/or R user/developer is invited to debate the whole thing.

RUnit – A Software Development And Test Package

Thomas König, Klaus Jünemann, and Matthias Burger
Epigenomics AG, Kleine Präsidentenstraße 1, D-10718 Berlin

Abstract

Software development for production systems presents a challenge to the development team as the quality of the coded package(s) has to be constantly monitored and verified. We present a generic approach to software testing for the R language modelled after successful examples such as JUnit, CppUnit, and PerlUnit. The aim of our approach is to facilitate development of reliable software packages and provide a set of tools to analyse and report the software quality status. The presented framework is completely implemented with R and does not rely on external tools or other language systems. The basic principle is that every function or method is accompanied with a test case that queries many calling situations including incorrect invocations. A test case can be executed instantly without reinstalling the whole package – a feature that is necessary for parallel development of functionality and test cases. On a second level one or more packages can be tested in a single test run, the result of which is reported in an easy to understand test protocol.

To verify the coverage of the test framework a code inspector is provided that monitors the code coverage of executed test cases. The result of individual test invocations as well as package wide evaluations can be compiled into a summary report exported to HTML. This report details the executed tests, their failure or success, as well as the code coverage. Taking it one step further and combining the build system with a development and release procedure with defined code status description this approach opens the way for a principled software quality monitoring and risk assessment of the developed application. For our code development we have utilized the described system with great benefit w.r.t. code reliability and maintenance efforts in a medium sized development team.

ADDITIVE MODELS FOR NON-PARAMETRIC REGRESSION: ADVENTURES IN SPARSE LINEAR ALGEBRA

ROGER KOENKER

ABSTRACT. The development of efficient algorithms for sparse linear algebra has significantly expanded the frontiers of statistical computation. This is particularly true of non-parametric regression where penalty methods for additive models require efficient solution of large, sparse least-squares problems. Sparse methods for estimating additive non-parametric models subject to total variation roughness penalties will be described. The methods are embodied in the R packages `SparseM` and `nprq`.

Models are structured similarly to the `gss` package of Gu and the `mgcv` package of Wood. Formulae like

$$y \sim qss(z_1) + qss(z_2) + X$$

are interpreted as a partially linear model in the covariates of X, with nonparametric components defined as functions of z_1 and z_2. When z_1 is univariate fitting is based on the total variation penalty methods described in Koenker, Ng and Portnoy (1994). When z_2 is bivariate fitting is based on the total variation penalty (triogram) methods described in Koenker and Mizera (2003). There are options to constrain the qss components to be monotone and/or convex/concave for univariate components, and to be convex/concave for bivariate components. Fitting is done by new sparse implementations of the dense interior point (Frisch-Newton) algorithms already available in the R package `quantreg`.

Koenker, R., and I. Mizera, (2004). "Penalized triograms: Total variation regularization for bivariate smoothing", *J. Royal Stat. Soc. (B)*, **66**, 145-163.

Koenker, R., P. Ng, and S. Portnoy (1994). "Quantile Smoothing Splines", *Biometrika*, 81, 673-680.

SciViews package

The SciViews package and its object browser

Eric Lecoutre[1] and Philippe Grosjean[2]

Abstract for the R User Conference, useR! 2004, Vienna

The SciViews package, which is in development since nearly one year, aims to provide useful graphical tools to work with R objects.

Some of those tools deals with importation/exportation of data, by example the exportation of a data frame to SAS by writing SAS code to the clipboard or to an external file.

Another class of functions defines *views* on objects. In R, we have show (S4 classes) or print (S3 classes) and summary for most objects. Here, we enlarge with other presentations of the content such as a report on missing values for instance. We use the HTML format that allows to embed graph, leading to complete output such as full reports for PCA (loadings, scores, screeplot, biplot,...). Views are written by calling functions from the R2HTML package, which use CSS styles. Thus, the look of all generated views may be changed by the user, through a custom CSS (Cascaded Stype Sheet) file. Moreover, we allow the user to easily define his own custom views.

The main feature of the package is available for Windows platform, as an add-in object browser that allows interaction with R. This object browser lists all objects within a R session, one can then filter by using predefined filters (data, data types, functions) or user-defined filters based on regular expressions. A taskCallback() function handles the autorefresh feature for the main workspace (.GlobalEnv). The GUI part of this object browser adds user friendly functionalities such as icons for objects, sorting abilities, selections and menus. Thus, the user could easily delete objects, export them, save his workspace, access to predefined views, change the display style (by changing the CSS file), etc.

This object browser is currently available for Windows plateform. The final objective is to both incorporate it in SciViews program[3] and to code a platform-independent version. Most code is either R, or HTML (for *views*), and both are platform-independent, so that only the graphical frontend interface must be rewritten to achieve this goal.

A first version of the SciViews package should be available in May 2004.

[1] Institut de statistique, Université catholique de Louvain, Belgium, lecoutre@stat.ucl.ac.be

[2] Laboratoire d'écologie numérique, Université de Mons-Hainaut, Belgium, phgrosjean@sciviews.org

[3] SciViews, http://www.sciviews.org

R2HTML package

formatting HTML output on the fly
or by using a template scheme

Eric Lecoutre
Institut de statistique, Université catholique de Louvain
<lecoutre@stat.ucl.ac.be>

Abstract for the R User Conference, useR! 2004, Vienna

We will present the R2HTML package. Basically, this package consists in a set of functions similar to base print or cat functions. Those functions allow to "export" R objects to a web page, by writing the HTML tags, as flat text. Thus, it is possible to every R user to manage complex HTML reports, without any HTML knowledge.

The package was initially designed for teaching purpose: the user may ask for an interactive output, all commands beeing automatically redirected to a HTML page. When using this functionnality, output are presented by using so called HTML frames, displaying both commands and their corresponding output at the same time. Only graphs have to be explicitely exported by using the command HTMLplot(), as there is no way to know when the user want to insert one graph (consider making a graph and adding lines on it). At the end of a training session, the student can take all the output with him - including graphs, exported as PNG og JPEG.

We will also illustrate a way of using the package to create complex reports for statistical analysis, which look can easily be modified. For that, we will use a sample analysis that returns an object (a list with a custom class), create a function that exports this object by exporting it's pieces with the functions provided by the package. Thus, analysis and reporting are separated. Finally, we will demonstrate the power of CSS (Cascading Style Sheets) to also separate formatting instructions.

Latest version of R2HTML package is available to download in CRAN and comes with a vignette explaining it's uses.

An Adventure in randomForest – A Set of Machine Learning Tools

Andy Liaw
Biometrics Research, Merck Research Laboratories
Rahway, New Jersey, USA

Abstract:

Random Forest is a relatively new method in machine learning. It is built from an ensemble of classification or regression trees that are grown with some element of randomness. While algorithmically it is a relatively simple modification of bagging, conceptually it is quite a radical step forward. The algorithm has been shown to have desirable properties, such as convergence of generalization errors (Breiman, 2001). Empirically, it has also been demonstrated to have performance competitive with the current leading algorithms such as support vector machines and boosting. We will briefly introduce the algorithm and intuitions on why it works, as well as the extra features implemented in the R package randomForest, such as variable importance and proximity measures. Some examples of application in drug discovery will also be discussed.

tuneR – Analysis of Music

Uwe Ligges

Fachbereich Statistik,
Universität Dortmund, 44221 Dortmund, Germany
e-mail: ligges@statistik.uni-dortmund.de

Abstract. In this paper, we introduce the R package **tuneR** for the analysis of musical time series.

Having done research on musical time series like "Automatic transcription of singing performances" (Weihs and Ligges, 2003) or "Classification and clustering of vocal performances" (Weihs et al., 2003), we feel there is need for a toolset in order to simplify further research. Since R is the statistical software of our choice, we are going to collect the tools (functions) we need in an R package starting with a set of functions that already have been implemented during our research mentioned above. These tools include, for example, functions for the estimation of fundamental frequencies of a given sound, classification of notes, calculations and plotting of so-called voice prints. Moreover, the package includes functions that implement an interface (Preusser et al., 2002) to the notation software *LilyPond*, as well as functions to read, write, and modify wave files.

We are planning to make the package, based on S4 classes, available to the public, and collect methods of other researchers, in order to finally provide a unique interface for the analysis of music in R.

References

Preusser, A., Ligges, U., and Weihs, C. (2002): Ein R Exportfilter für das Notations- und Midi-Programm LilyPond. *Arbeitsbericht 35, Fachbereich Statistik, Universität Dortmund.*

Weihs, C., and Ligges, U. (2003): Automatic Transcription of Singing Performances. *Bulletin of the International Statistical Institute, 54th Session, Proceedings, Volume LX, Book 2, 507–510.*

Weihs, C., Ligges, U., Güttner, J., Hasse-Becker, P., and Berghoff, S. (2003): Classification and Clustering of Vocal Performances. In: M. Schader, W. Gaul and M. Vichi (Eds.): *Between Data Science and Applied Data Analysis.* Springer, Berlin, 118–127.

Keywords

MUSIC, R, TIME SERIES

Simulating Society in R

András Löw,
Department of Statistics, Faculty of Social Sciences, ELTE
Institute of Sociology, Faculty of Humanities, PPKE
alow@index.hu

Socio-economic relationships in a society can be simulated with mathematical models. There are many special software (e.g.: Peersim, RePast, Swarm) to simulate single problems. A useful toolkit was written by Gaylord and D'Andria* in Mathematica, making use of it's rule-based programming style and graphical capabilities.

The R has almost the same tools for visualisation, however, it does not permit rule-based programming.

This above toolkit has been transformed to R. This is not only a simple translation because of the differences in the structure of the two program languages. By this way this toolkit is adaptable not only to two dimensional rectangular grid but also other topologies (e.g.: irregular grid, small world, random graph).

This package is demonstrated in the poster corresponding to the first three chapters of the book *Simulating Society — A Mathematica Toolkit for Modeling Socioeconomic Behavior.* The following two groups of social phenomena are simulated:

1) How do people come to have shared values based on ideas, beliefs, likes and dislikes, or attitudes? One possible mechanism for the spreading of values through population is through a sort of contagious process, occurring as individuals come into contact with one another and interact. This interaction results in a form of imitative behaviour sometimes referred to as cultural transmission or social learning.

   In the package several models of the change of values in mobile society are considered. Especially a simpler version of the model is analysed in which social status determines the direction of meme transmission.

2) Why are people generally honest in their dealings with others, even in the absence of a central authority to enforce good behaviour? The role of ostracism is modelled as a tool for discouraging bad behaviour, and thereby encouraging good behaviour. Two cases are demonstrated in which people have the ability to remember or learn about other people's bad behaviour. In the first situation, people remember every individual who has done them wrong in a previous encounter and they refuse to interact with such a person again. In the second situation, good guys use word-of-mouth or

*Gaylord, Richard J. and D'Andria, Louis (1998): Simulating Society - A Mathematica Toolkit for Modeling Socioeconomic Behavior; New York, NY: TELOS/Springer Verlag ISBN 0-387-98532-8

> gossip in addition to personal experience to learn who are the bad guys, and they avoid interacting with a person with a bad rep even once.

In R program language there is no need to generate large data files because files can be analysed during the simulation process and only the necessary statistics need to be saved. This method is memory efficient compared to the post processing technique and faster than using two interacting programs.

MCMCpack: An Evolving R Package for Bayesian Inference

Andrew D. Martin[*] Kevin M. Quinn[†]

February 12, 2004

MCMCpack is an R package that allows researchers to conduct Bayesian inference via Markov chain Monte Carlo. While MCMCpack should be useful to researchers in a variety of fields, it is geared primarily toward social scientists. We propose to discuss the design philosophy of MCMCpack, the functionality in the current version (0.4-7) of MCMCpack, and plans for future releases of MCMCpack. We also hope to use the useR! forum to learn what features current and potential MCMCpack users would like to see in future releases.

MCMCpack is premised upon a five point design philosophy: a) widespread, free availability; b) model-specific, computationally efficient MCMC algorithms; c) use of compiled C++ code to maximize computational speed; d) an easy-to-use, standardized model interface that is very similar to the standard R model fitting functions; and e) compatibility with existing code wherever possible.

MCMCpack currently offers model fitting functions for 14 models. Some of these models are quite common (linear regression, logistic regression) while other are more specialized (Wakefield's baseline model for ecological inference, a factor analysis model for mixed ordinal and continuous responses). In addition, MCMCpack makes use of the coda library for posterior analysis and has a number of helper functions that are useful for manipulating the MCMC output.

In future releases we hope to: add support for additional models, allow researchers to specify a wider range of prior distributions, add an instructional module, improve the documentation, and to include a number of C++ and R template files that will help researchers write code to fit novel models.

We hope to use useR! to learn more about the preferences and goals of the (potential) MCMCpack user-base. In addition, we hope to learn how new features of R (such as namespaces and S4 classes) can be exploited to improve MCMCpack.

[*] Assistant Professor, Department of Political Science, Washington University in St. Louis, Campus Box 1063, St. Louis, MO 63130. admartin@wustl.edu

[†] Assistant Professor, Department of Government and CBRSS, 34 Kirkland Street, Harvard University, Cambridge, MA 02138. kevin_quinn@harvard.edu

LSD Plots: Some Aspects of Their Implementation in R

Ivan Mizera, University of Alberta

In the contribution, we discuss our experience with the implementation of a new graphical tool for exploratory data analysis, the LSD plot proposed by Mizera and Muller (2004). LSD plots are plots of the contours of location-scale depth, a data-analytic construct in the vein of general theory of Mizera (2002), whose origins may be traced back to Tukey (1975) and Rousseeuw and Hubert (1999). The Lobachevski geometry structure of its most feasible variant, the Student depth, creates a link to multivariate location halfspace depth, which enables to utilize the recent algorithmic advances in the field - like those of Struyf and Rousseeuw (2000) or Miller et al. (2003). The LSD plots can be used for checking distributional assumptions about the data, in a fashion similar to that of quantile-quantile plots; they exhibit the similar incisive nature of the latter as well.

While the computational experiments with the new methodology were done predominantly in MATLAB, a transition to more user-oriented implementation in R posed several technical problems. In particular, the availability of (somewhat) interactive statistical environment, like iPlots developed by Urbanek and Theus (2003), seems to be paramount for the routine use of the technique.

Miller, K., Ramaswami, S., Rousseeuw, P., Sellares, A., Souvaine, D., Streinu, I., and Struyf, A. (2003). Efficient computation of location depth contours by methods of computational geometry, Statist. and Comp. in press.

Mizera, I. (2002). On depth and deep points: A calculus. Ann. of Statist. 30, 1681-1736.

Mizera, I. and Muller, Ch. H. (2004). Location-scale depth, J. Amer. Statist. Assoc. in revision.

Rousseeuw, P. J. and Hubert, P. (1999) Regression depth (with discussion). J. Amer. Statist. Assoc. 94, 388-402.

Struyf. A. and Rousseeuw, P. J. (2000). High-dimensional computation of the deepest location. Comput. Statist. and Data Anal. 34, 415-426.

Tukey, J. W. (1975). Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians, Vol 2., Vancouver, B. C., 1974, Canad. Math. Congress, Quebec, 523-531.

Urbanek, S. and Theus M. (2003). iPlots: High interaction graphics in R. Proceedings of the 3rd International Workshop in Distributed Statistical Computing, March 20-22, Vienna, Austria (Kurt Hornik, Friedrich Leisch, and Achim Zeileis, eds.).

Orientlib: Using S Version 4 Objects

D.J. Murdoch[*]

April 15, 2004

Abstract

R supports two systems of object oriented programming. The S version 3 objects work by a naming convention. Chambers (1998) introduced the S version 4 object system. Here a registration system connects objects to their methods. In this presentation, I will describe Orientlib, an R package using S version 4 objects.

Orientlib (JSS, 2003) is an R package to facilitate working with orientation data, i.e. elements of $SO(3)$. It provides automatic translation between different representations of orientations, including rotation matrices, quaternions, Euler angles and skew-symmetric matrices; it also has functions for fitting regression models and displaying orientations. This presentation reviews simple properties of orientation data and describes Orientlib.

Keywords: S version 4 objects, SO(3), orientation data, Euler angles, quaternions, rotation matrices

[*] University of Western Ontario

Comparative analysis of the oestrogen-responsive gene expression profile of breast cancer cells with three different microarray platforms

Margherita Mutarelli(1),(2), Walter Basile(3), Luigi Cicatiello(3), Claudio Scafoglio(3), Giovanni Colonna(2), Alessandro Weisz(3) Angelo Facchiano(1),(2)
[1] Istituto di Scienze dell'Alimentazione CNR, via Roma 52A/C, 83100 Avellino
[2] Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università degli Studi di Napoli, via de Crecchio, 80138 Napoli, Italy
[3] Dipartimento di Patologia Generale, Seconda Università degli Studi di Napoli, via de Crecchio, 80138 Napoli

The DNA microarray is a technique which makes it possible to analyze the expression patterns of tens of thousands genes in a short time. The widespread use of this technique and the rapidly improving different technologies available by commercial and academic providers has led to the publication of thousands of results, extremely heterogeneous in terms of technology used and analysis subsequently applied to data. This leads to a difficulty in collaborating and exchange data between groups with common research interest, whereas collaborations would be extremely useful due to the high cost of this techniques and to the consideration that an experiment carefully designed could bring results relevant to different groups, each focusing on a different aspect of a main biological problem. So the awareness for the need of common standards or, at least, comparable technologies is emerging in the scientific community, as shown by the effort of the Microarray Gene Expression Data (MGED) Society, which is trying to set up at least experimental methodology, ontology and data format standards.

In addition, it is important the ability of being able to compare newly produced data with preceding experiments, so to ensure of keeping high the value of results produced with equipment of the old generation.

We thus started this work with the aim of evaluating the technical variability between three commonly used microarray platforms, such to adapt the first part of the analysis to the peculiarity of each technique, and the feasibility of a common subsequent analysis path, thus taking advantage of the different data-extraction abilities of the three. The chips used to study the gene expression profiles of hormone-responsive breast cancer cells with and without stimulation with estradiol are:

i) the Incyte 'UniGEM V 2.0' microarrays, containing over 14,000 PCR-amplified cDNAs, corresponding to 8286 unique genes, spotted at a high density pattern onto glass slides;

ii) the Affymetrix technology, based on 25 nucleotide-long oligonucleotides directly synthesized on a GeneChip® array, representing more than 39,000 transcripts derived from approximately 33,000 unique human genes;

iii) the Agilent 'Human 1A Oligo' Microarray consisting of 60-mer, *in situ* synthesized oligonucleotide probes for a total of about 18000 different genes.

The same samples were used to generate fluorescent targets to be hybridized on the different slides, with balanced dye swap when applicable for competitive hybridizations.

Ackowledgements

Research supported by: Italian Association for Cancer Research (AIRC, Investigator Grants 2003), Italian Ministry for Education, University and Research: FIRB Post-genomica (Grant RBNE0157EH_001/_003), European Commission (Contract BMH4-CT98-3433), Ministry of Health (Progetti Speciali 2000 and RF02/184), Second University of Naples (Ricerca di Ateneo 2002-2003). M. Mutarelli and C. Scafoglio are PhD students, supported by EU funds, of Dottorato di Ricerca in "Biologia Computazionale" and in "Oncologia medica e chirurgica ed Immunologia clinica", respectively, at the Second University of Naples.

Taking into account spatial dependence in multivariate analysis

Sébastien Ollier, Anne-Béatrice Dufour, Jean Thioulouse, Daniel Chessel

Laboratoire de Biométrie et de Biologie Evolutive. UMR CNRS 5558, Villeurbanne, France.

Abstract

R is a valuable tool to develop statistical methods using its package structure and the open sources. Indeed, it allows combining classes and methods from different packages in order to generalize their own capabilities. The present topic deals with the relationship between classes and methods from both packages.**spdep** and **ade4**, which implement methods for spatial data analysis (mainly lattice/area style) and multivariate analysis respectively. Combining objects from both packages leads to the development of a new statistical method to take into account spatial dependence in multivariate analysis.

spdep is a package defined by Bivand as “a collection of functions to create spatial weights matrix **W** from polygon continuities, from point patterns by distance and tessellations for permitting their use in spatial data analysis”. Two classes of objects have been established: “nb” a list of vectors of neighbour indices, and “listw” a list containing a “nb” member and a corresponding list of weights for a chosen weighting scheme. The package contains in addition many functions to test spatial autocorrelation and estimate spatial autoregressive model but it only supports univariate data analysis.

ade4 is precisely a package dedicated to multivariate analysis of Ecological and Environmental Data in the framework of Euclidean Exploratory methods. It is a collection of functions to analyse the statistical triplet (**X**,**Q**,**D**) where **X** is a data set of variables or a distance matrix; **Q** and **D** are diagonal matrices containing column and row weights, respectively. The singular value decomposition of this triplet gives principal axes, principal components and, column and row coordinates. All these elements are gathered in a list defining the **du**ality **di**agram class called “dudi”. Each basic statistical method corresponds to the analysis of a particular triplet. For instance, **ade4** implement the principal component analysis on correlation matrix *via* a function called “dudi.pca”. In that case, **X** is a table containing normalized quantitative variables, **Q** is the identity matrix $\mathbf{I}_p$ and **D** is equal to $\frac{1}{n}\mathbf{I}_n$.

An illustration of the relationship between **spdep** and **ade4** is shown through the “multispati” function. This function combines objects of class ”listw “ with objects of class “dudi” to analyse the statistical quadruplets (**X**,**Q**,**D**,**W**). It allows first, the generalisation of the univariate Moran test and second, to take into account spatial constraints in Euclidean Exploratory methods. During this talk, the implementation of multivariate spatial correlations analysis proposed by Wartenebrg (1985) will be described as a particular use of the “multspati” function.

Wartenberg, D. E. 1985. Multivariate spatial correlations: a method for exploratory geographical analysis. Geographical Analysis **17**:263-283.

R-parse: A Natural Language Processing Application of R

Over the last several years, the fields of Natural Language Processing and Computational Linguistics have shifted their focus toward statistical methods, emphasizing probabilistic grammars, Expectation Maximization, Maximum Entropy, and Log-linear modeling frameworks. These developments make R an attractive environment in which to explore the development of new computational linguistic models. One such model involving a straightforward application of R is Latent Semantic Analysis/Indexing (LSA/LSI), which employs Singular Value Decomposition (SVD) to perform a principal components-style analysis of term frequencies in documents. This application is readily handled by writing wrapper functions that perform specialized handling of the built-in matrix analysis and plotting functions of R.

Yet, an impediment to the widespread use of R in NLP/CL research is the lack of good facilities for the handling of traditional symbolic processing tasks such as language parsing and generation. The fact that R is a member of the LISP family of programming languages suggests that this could be done, since most symbolic NLP work has used declarative languages like LISP and Prolog. At the same time, it remains to be seen whether R itself is flexible enough to permit useful parsers to be written without resorting to external language extensions.

This paper reports on efforts to develop a simple but flexible chart-based parser in R called R-parse. Chart parsers are preferred in NLP applications, as they minimize the redundant work of the parser by employing a data structure for recording attempted parses. R-parse is implemented as a recursive function that returns a parsed string in the form of a chart. Charts and grammar rules are both implemented as data frames, hence both parser output and grammar are amenable to statistical analysis using functions such as `glm()` (e.g. for log-linear modeling of the parsed output). Furthermore, a probabilistic generator function R-gen accepts the same form of grammar as R-parse. Together, the parser and generator allow one to model and study all aspects of the relationship between probabilistic grammars and the languages they generate entirely within the resources of R. The properties of R-parse and R-gen can presently be demonstrated on modest but realistic grammars and language training corpora.

An obstacle for the development of R-parse has turned out to be R's relatively impoverished meta-programming facilities. Such features are necessary for implementing exotic execution control schemes needed in parsers (e.g. backtracking) without resorting to an entirely redesigned stack. Much of this work might be handled with some form of environment-passing, but unfortunately, in its present implementation, R's environment-passing is not powerful enough as it provides no way to selectively prevent the modification of variables in an environment. Hence intervening computations can "spoil" the outcome of a computing on an environment that has since been inadvertently modified. An improved meta-programming environment in R, perhaps made possible by exposing more of R's internal workings in the fashion of `delay()`, might permit more elegant solutions to these problems. In the mean time, a working version of R-parse exists which correctly records in the chart all possible parses of a string, including failed and incomplete parses, for both context-free and regular grammars.

Implementation and Simulation of Ecological Models in R: an Interactive Approach

Thomas Petzoldt[*]

February 9, 2004

In recent years, the R system has evolved to a mature environment for statistical data analysis, the development of new statistical techniques, and, together with an increasing number of books, papers and online documents, an impressing basis for learning, teaching and understanding statistical techniques. Moreover, due to the advantages of the S language and the OpenSource availability of the R software and its libraries, this system is also suitable to implement simulation models, which are common in ecology.

However, there remains one problem. When implementing models of different type, e.g. differential equations or individual-based models, the result may be a lot of different simulation programs and the interactive ease of a command driven R-system gets lost. The need to edit parameter files or the sourcecode directly is a serious disadvantage, particularly when sharing such models.

As a first step towards an intercheangable but easy to use format I propose a standardized list-structure to allow users to run, modify, implement, discuss and share ecological models. Each simulation model can be implemented as a list (simecol simulation model object) with an intentionally simple structure. The list has very few mandatory components, namely `equations`, which contains the model equations, rules or arbitrary program code, `params` with the constant model parameters and `init` which is used to define the initial state as vector, matrix, list or via a special initialization function.

For dynamic models the vector `times` is used to define simulation interval and time steps and `inputs` for variable data (e.g. time dependend environmental forcings). The list component `solver` defines the solver to be used to simulate the model. It can be either `iteration` for discrete event models, an ODE-solver (e.g. `lsoda`) for ODE models or a user-provided algorithm. Simulation results are stored within the simecol object for later use, e.g. plotting or statistical analysis.

In this way, a simecol object contains the entire dataset, that is needed to run a basic simulation simply by entering the model object via `source()` or `data()` from the harddisk or the internet and then to run and plot the model via `plot(simulate(simecol.object))`. As an additional advantage R-helpfiles and package vignettes can be used as usable and standardized methods to overcome the documentation dilemma.

Interface functions are provided to get or set model parameters, time steps and initial values, but it is also possible to modify the components of simecol objects directly, e.g. the model equations. Compared to a strictly object oriented approach, this might seem dangerous and may lead to an inconsistent state, but on the other hand the list-based approach makes implementation and share of new models extremely simple. Using this, several applications with different types of ecological models will be demonstrated to show the interactive potential of this approach.

[*] Institute of Hydrobiology, Dresden University of Technology, 01062 Dresden, petzoldt@rcs.urz.tu-dresden.de

useR! – at an investment bank?

Bernhard Pfaff [*]

Frankfurt, January 29, 2004

Abstract

This paper describes the demands placed on the statistical and/or econometric software utilised by an investment bank. It outlines the reasons and considerations for choosing R instead of alternative programs/programming environments as *the* statistical software environment at Dresdner Kleinwort Wasserstein (henceforth referred to as DrKW). The features of a typical workflow process as well as examples of financial econometric models and their implementation are also discussed. The experiences and insights described in this paper have been gained from day-to-day usage of R since mid-2002.

Introduction

Since the beginning of the 90s not only academic interest in econometric modelling of financial markets has increased but also in todays financial industry the role and necessity of quantitative research econ

Pecularities

Econometric Modelling

Conclusion

[*] The views expressed in the paper are the author's own and do not necessarily reflect those of Dresdner Kleinwort Wasserstein or any associated legal entity. This note was prepared for: "useR! 2004 – The R User Conference", May 20-22, 2004, Vienna, Austria. Adress correspondence to: Bernhard Pfaff, Dresdner Kleinwort Wasserstein, Corporates & Markets, Capital Markets, Global Research, Theodor–Heuss–Allee 44–46, 60301 Frankfurt am Main, Germany, email: bernhard.pfaff@drkw.com

Taking into account uncertainty in spatial covariance estimation for Bayesian prediction

Jürgen Pilz[1], Philipp Pluch[1] and Gunter Spöck [1]

[1] Department of Mathematics
University of Klagenfurt
A-9020 Klagenfurt, Austria

Keywords: kriging, Bayesian predictive distribution, Matern covariance function, variogram parameter uncertainty

Abstract

In practice, spatially varying phenomena are modelled using second order random fields, given by their mean function and covariance function. For estimation of the random field the so called Kriging predictor is used, which is known as the best linear unbiased predictor. But the optimality just holds on the assumption that the covariance function of the underlying random field is exactly known. The estimated covariance function, however, may lead to an underestimation of the prediction errors.

We take into account this uncertainty by developing a robust Bayesian predictor which applies to the whole family of plausible covariance functions. We get this family by means of a simulation strategy. Instead of getting a single predictor, we calculate the whole predictive densities at the points to be predicted.

This poster shows how bad plug in prediction as used in all facets of kriging can be. After giving details of the derivation of the Bayesian predictor and its calculation we give an example. The data set used deals with radioactivity measurements 10 years after the Chernobyl accident.

useR!2004 Conference Abstract

R at NCAR

Matthew Pocernich, Eric Gilleland and Doug Nychka

The need to understand the complex physical, biological and societal relationships of the Earth system will pose a rich set of statistical problems. We are using our perspective at the National Center of Atmospheric Research to adapt the R environment for the geophysical community. Problems include dealing with very large datasets and working with models that use both observational data and data derived from numerical models. NCAR's mission is in part to "... foster the transfer of knowledge and technology for the betterment of life on Earth." R plays an important and growing role in fulfilling this mission. This is due to R's philosophical commitment to sharing, its acceptance within the statistical community and the increasing sophistication, stability and breadth of applications created by the community.

This presentation will consider the current use of R within the scientific community at NCAR and possible future efforts. In particular, examples of packages created at NCAR will be presented. Limitations (real and perceived) for using R at NCAR will also be discussed. An overview of the several packages that have been developed at NCAR and are (or will be soon) available on CRAN will be presented. These packages include the following.

extRemes is a package developed to assist in teaching extreme value statistics. Developed by Eric Gilleland and others, this packages provides a graphical user interface to the **ismev** package developed by Stuart Coles and implemented in R by Alec Stephenson.

fields was developed by the Geophysical Statistics Project (GSP) at NCAR. This package focuses on curve and function fitting, with a special emphasis on spatial problems. Major topics addressed by functions in the package include thin plate spline regression, kriging, space-filling designs and other functions related to working with spatial data.

RadioSonde presents and summarizes data from radiosondes and drop-sondes. These devices are attached to balloons or dropped from planes in order to gather atmospheric information.

verify was developed internally to standardize some routine model and forecast verification diagnostics. While created with weather forecasts in mind, it is written to be readily applied to other types of forecasts and models. By making algorithms more readily available, we hope to extend discussions of their usefulness from the theoretical to the applied.

While R is universally used by the statisticians at NCAR, far fewer atmospheric scientists and software engineers use it. The following actions and activities could expand the number of R users at NCAR and will be briefly discussed.

- The NCAR Command Language (NCL) is a programming language designed for the analysis and visualization of data. NCL is particularly useful in handling and processing very large datasets stored in netcdf, grib and binary files. Presently, NCL has somewhat limited statistical capabilities. Graphs are produced using a low level graphics language. Allowing users of NCL to use R's statistical and graphics function could greatly help this community.

- Prototyping by software engineers is commonly done in Matlab and so many Matlab libraries exist. The ability to use these Matlab libraries would be key in getting more software engineers to proto-type in R.

- R's GPL license could be a benefit in attracting some software engineers to R. There have been instances when developers wanted to offer their code to an external audience but were prohibited because of Matlab's license.

- Some scientists use Excel to do graphs and simple analyses. When the analysis become too complicated or the datasets too large, software engineers often provide solutions either in Matlab or in a compiled code. Instruction and support for scientists who want to use R will help R gain acceptance amongst these users. NCAR currently has a users group, but it is not as active as it might be. A more active user community will help extend the sues of R at NCAR.

Replacing Cluttered Scatterplots With Augmented Convex Hull Plots

Maja Pohar, Gaj Vidmar
Institute of Biomedical Informatics
University of Ljubljana
maja.pohar@mf.uni-lj.si

Ljubljana, April 14, 2004

Abstract

The presentation of multi-group scatter plots can be often obscured by the visual clutter. A way to get more information from such plot is to replace the scatterplot points with convex hulls. Thus space is gained for vizualization of descriptive statistics with error bars or confidence ellipses within the convex hulls. Bivariate density plots might be used instead of convex hulls in the presence of outliers. An informative addition to the plot is calculation of the area of a convex hull divided by corresponding group size — a bivariate dispersion measure weighting all deviations from the center equally. Marginal distributions can be depicted on the sides of the main plot in the established ways.

The limited possibilities for producing such plots in existing software have led us to implement these graphs in R - we introduce the function *chplot* that automatizes this kind of plots. Following the standard plotting options we have devised a function that makes the calculations needed to produce a sensible plot of the data, but at the same time allows the user to change any of the options at his will. We include some examples of the usage based on the Iris dataset and daily statistical consulting practice.

Doing Customer Intelligence with R

Jim Porzak,
Director of Analytics, Loyalty Matrix, Inc., San Francisco, CA, USA
jporzak@loyaltymatrix.com

The goal of customer intelligence (CI) is to transform behavioral and motivational customer data into business insights that can change an organizations marketing strategies and tactics. CI, when done well, delivers improved customer satisfaction and loyalty. CI can only be done well, however, by merging rigorous data techniques with the skills of knowledgeable business analysts — truly is a combination of art and science.

Existing CI analytical tools are designed for the largest businesses. They are pricy and complex requiring significant commitment. Our goal, at Loyalty Matrix, is to deliver the benefits of CI to midsize organizations to enable them to compete with the large enterprises. To achieve our goal, we must be extremely efficient. We achieve efficiency by first standardizing our data housing and analytic techniques and then using tools that are economic and well understood.

Our MatrixOptimizer® platform handles data housing, basic analysis, and client presentation tasks. It is based on standard Microsoft technology: SQL Server, Analysis Services (OLAP) and .Net. What has been missing is strong presentation graphics, exploratory data analysis (EDA), rigorous basic statistics and advanced data mining methods. R promises to fill this gap.

For the last six months, Loyalty Matrix has been using R primarily for EDA, ad hoc project specific statistics and some modeling. We will present the following case studies:

1) Analysis of restaurant visits: visit intervals, travel distance, mapping diner travel patterns, seasonality by region.

2) Automotive purchase patterns: purchase intervals, survey analysis, randomForest modeling of segments, brand loyalty.

3) Retail loyalty survey analysis: visual survey analysis, competitive shopping patterns.

4) Direct marketing campaign design and evaluation: prospect targeting, predicting lift, campaign response evaluation.

Based on these learnings and our past work with numerous clients, our next step is to integrate CI methods written in R into the MatrixOptimizer making rigorous analytics, presentation quality charts, and some data mining methods directly accessible by our business analysts.

At useR! 2004, Loyalty Matrix will announce our sponsorship of an "Open CI" project. We intend to publish one or more packages specifically targeted at CI and, more broadly, at the use of R in quantitative marketing. We see great benefit in making core CI analytics open, peer reviewed, and extensible. We will be encouraging other groups to join us in the project.

Teaching applied finance using R

John Randal Peter Thomson

February 11, 2004

Abstract

Volatility estimation is an ever expanding part of the financial literature. This topic in applied finance is the theme of a first year graduate paper taught by the authors to students whose typical background not only involves little or no programming experience, but also limited mathematics. We describe the challenges involved in teaching such a course to this "unskilled" audience.

We outline the general course structure, and also the structure of the parallel lab-based course in R programming and practice. We indicate how implementation in R of the techniques taught in the class, many of them current research, allowed the students to not only master the techniques themselves, but also to develop familiarity with a powerful tool. Further, we have found use of R, in particular for simulation and graphical analyses, can compensate for relatively weak mathematical skills, and allow students to grasp ideas and make progress that would otherwise be prohibitively costly in the scope of the course.

In other courses, the students were typically use EViews, however indications from the course evaluation are overwhelmingly supportive of the use of R, despite the students having only 12 weeks to overcome their "poor" programming backgrounds, and to realise the merits of such a package.

Option pricing using R

John Randal

February 11, 2004

Abstract

R has a contributed package RQuantLib which provides access to some of the QuantLib libraries, and common, practitioner-oriented option pricing tools. These functions are fairly limited in their scope and applicability from a research perspective, in the sense that they are based around the standard Black-Scholes assumptions.

We review extensions to the Black-Scholes pricing equation, in particular the constant elasticity of variance, and the compound option pricing models, focussing on problems relating to computability of their closed form solutions. In each case, R offers probability tables that are otherwise fairly obscure, in particular, the non-central chi-squared distributions, and the multivariate normal distributions available in the package mvtnorm.

We present a further extension, the extended compound option pricing model, which also has a closed form solution, and which has applications in stock, stock option, and debt pricing. The model can be implemented in R, however evaluating the pricing formula presents some challenges. These challenges are discussed.

Using R for producing a monthly Agrometeorological report

R. Rea, G. Toller, E. Eccel

Unità operativa Agrometeorologia e Clima,
Istituto Agrario San Michele all'Adige, Italy.

Corresponding author address: Roberto Rea, Istituto Agrario San Michele all'Adige, Unità operativa Agrometeorologia e Clima, via Mach 1, I-38010 S.Michele all'Adige (Trento - Italy). E-mail: roberto.rea@ismaa.it

In the present work a procedure is described for publishing a monthly Agrometeorological report for the province of Trentino (Italy).
Two types of report are generated: a general report which shows the agrometeorological development of the previous month all over the province and a specific one which describes the situation at every site where a weather stations is present.
The input data for this report are collected by the Agrometeorological Observation Network of the Istituto Agrario San Michele all'Adige, which actually manages about 80 weather stations.
The data recorded by the stations are stored twice a day in a MySQL Database.
The R-procedure automatically starts when the data of the previous month are complete. The data of different variables such as air temperature, precipitation, soil temperature, day degrees, humidity and leaf wetness are extracted by using the library *RMySQL*. Daily and monthly tables reassuming the agrometeorological situation of the previous month are generated using base *R* commands and finally written in .tex format by using the library *xtable* .
Various plots that give a picture of the wheather situation in the previous month over the region and in the various stations are generated by using *base package* and the *lattice* library. Maps of mean monthly temperature and total precipitation to be included in the general report are built using *gstat* library for spatial interpolation.
The *R* procedure writes a .tex file in which all the graphics and tables are included and formatted. At the same time notes and captions are added.
Finally the .tex file is compiled by using a latex compiler and reports are generated in .pdf format. Further developments include the generation of weekly and daily reports.

Bioassay analysis using R

Christian Ritz & Jens C. Streibig[1]

[1] Department of Natural Sciences & Department of Agricultural Sciences, The Royal Veterinary and Agricultural University, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark, E-mail: ritz@dina.kvl.dk

Background

Herbicides are designed to kill plants and their selective use in crops is based on rates that do not harm the crop but effectively control the weeds. Use of dose-response curves and subsequent definitions of relative potency, no observable effect level, estimation of selectivity index are pivotal for the safe and cost-effective use of herbicide in agriculture.

Various dose-response curves have been used to describe plant response. The objective of this abstract is to present an R application that makes it easy to analyse bioassays and subsequently performs tests of biological relevance not only in herbicide research and development, but also in other branches of the biological sciences, e.g. toxicology and ecotoxicology.

R application

In order to automate the fitting and testing of various hypotheses of the dose-response curves we have developed the package "drc" which can handle most of the situations common for bioassay work:

- Simultaneously fitting of numerous curves.
- Testing for similar parameters among some or all simultaneously fitted curves.
- Reducing numbers of parameters using likelihood ratio tests.
- Calculating relative potencies (comparison of parameters).

The non-linear least squares estimation is performed using the `base` function `optim`. The resulting fit is an object of class "drc". In addition to some more specialised functions, the package provides `anova`, `plot` and `summary` (`print.summary`) methods for objects of class "drc".

The framework can easily be extended to other situations of non-linear regression with parameters depending on some factors. It requires that the non-linear mean function and (preferably) an accompanying self start function are supplied.

References

Finney, D. J. (1979). Bioassay and the Practice of Statistical Inference. *Int. Statist. Rev.* **47**, 1–12.

Ritz, C. and Streibig, J. C. (2004). Bioassay analysis using **R**. *Manuscript.*

Fitting the Generalized Pareto Distribution to Threshold Exceedances of Stock Index Returns In Bull and Bear Periods

Angi Rösch*
FOM University of Applied Sciences
Munich, Germany

Abstract

The investigation of phenomena of joint extreme returns on financial assets is crucial to financial risk assessment. We focussed on joint threshold exceedances of daily returns on stock indices, threshold meaning the lower, respectively upper 10% quantile of the return distribution. Exceedances in bull and bear periods were treated separately.

In a first step, we fitted the generalized Pareto distribution to the exceedances of several stock indices, applying the elemental percentile method of Castillo and Hadi. This provided the basis for the second step: Pairs of marginal distributions of threshold exceedances were coupled, assuming a logistic dependence structure for returns, in which the degree of dependency is reflected by a single correlation parameter. This led to bivariate distributions of threshold exceedances of returns.

Comparing this method to one which ignores the dependency among return exceedances, we found that financial risk would then be seriously underestimated. The degree of dependency tended to be stronger during bear periods.

All computations were carried out in the statistical computing environment *R*.

*e-mail: angi.r@t-online.de

klaR: A Package Including Various Classification Tools

Christian Röver, Nils Raabe, Karsten Luebke, and Uwe Ligges

Fachbereich Statistik,
Universität Dortmund, 44221 Dortmund, Germany

Abstract. Classification is a ubiquitous challenge for statisticians. The R-package **klaR** includes functions to build, check and visualize classification rules. Since the package is being developed at the Dept. of Statistics at the University of Dortmund, the name of the package is based on the German word "Klassifikation" and R. In this paper, some of the functions included in the package are described and presented in the context of classification of the German economy's business cycle phases (Heilemann and Münch, 1996).

Among these functions, there is a wrapper to call SVMlight (by Thorsten Joachims, `http://svmlight.joachims.org/`) as well as an implementation of an Regularized Discriminant Analysis (*RDA*; Friedmann, 1989) which may also optimize the parameters needed for an RDA.

Another function, `stepclass()`, performs "stepwise classification", i.e. it selects variables of the data that are used for the classification method by minimizing the cross-validated error. The classification result for any two variables can be visualized by drawing the partitions of the data.

In order to check the results various performance measures like the one described by Garczarek (2002) can be calculated. If the data consists of 3 or 4 classes, the membership values of different classifiers can be compared by visualization in a barycentric coordinate system.

Moreover, part of the package are functions for *EDAM*. In this approach a set of vectors is visualized in a fixed solution space by an "Eight-Directions-Arranged-Map" (EDAM; Raabe, 2003). If the visualized vectors are equal to the centroids of a set of clusters, the result can directly be compared to that of a Self-Organizing-Map.

References

Friedman, J.H. (1989): Regularized Discriminant Analysis. *Journal of the American Statistical Association, 84*, 165–175.

Garczarek, U.M. (2002): *Classification rules in standardized partition spaces.* Phd Thesis, University of Dortmund.

Heilemann, U. and Münch, J.M. (1996): West german business cycles 1963-1994: A multivariate discriminant analysis. *CIRET-Conference in Singapore, CIRET-Studien 50.*

Raabe, N. (2003): *Vergleich von Kohonen Self-Organizing-Maps mit einem nicht-simultanen Klassifikations- und Visualisierungsverfahren.* Diploma Thesis, Department of Statistics, University of Dortmund.
`http://www.statistik.uni-dortmund.de/de/content/einrichtungen/lehrstuehle/personen/raabe/Diplomarbeit.pdf`.

Keywords

Classification, Visualization

crossdes — A Package for Design and Randomization in Crossover Studies

Oliver Sailer

Fachbereich Statistik
Universität Dortmund, 44221 Dortmund, Germany

February 13, 2004

Design of experiments for crossover studies requires dealing with possible order and carryover effects that may bias the results of the analysis. The classical approach to deal with those effects is to use designs balanced for the respective effects. Almost always there are effects unaccounted for in the statistical model that may or may not be of practical importance. A convenient way of addressing those effects is randomization. Different kinds of models, however, require different methods of randomization. For many types of experimental designs it is not known whether a certain method of randomization is conformable with a certain model.

The construction of balanced designs is often based on latin squares. The package *crossdes* contains some functions to construct designs balanced for carryover effects like the ones promoted in Wakeling and MacFie (1995). Simulation functions are provided that test whether some basic randomization procedure may be applied to a given design so that one-way or two-way block models give unbiased estimates for mean and variance of treatment contrasts. Here an approach similar to that of Kunert (1998) and Kunert et al. (2002) is used. The simulations done in **R** help to assess the use of experimental designs that are not fully balanced for carryover or period effects in crossover studies (Kunert and Sailer, 2004).

Supplementary functions for randomization and checking for balance of any given design are also provided. The beer testing experiment presented in Kunert (1998) will be discussed as an example.

References:

KUNERT, J. (1998): Sensory experiments as crossover studies. *Food Quality and Preference, 9, 243–253.*

KUNERT, J., MEYNERS, M. and ERDBRÜGGE, M. (2002): On the applicability of ANOVA for the analysis of sensory data. In: *7ème Journées Européennes Agro-Industrie et Méthodes Statistiques.* Proceedings. 129–134.

KUNERT, J. and SAILER, O. (2004): On nearly balanced designs for sensory trials. *In preparation.*

WAKELING, I.N. and MACFIE, H.J.H. (1995): Designing consumer trials balanced for first and higher orders of carry-over effect when only a subset of k samples from t may be tested. *Food Quality and Preference, 6, 299–308.*

Keywords:

NEIGHBOUR BALANCED DESIGNS, CROSSOVER STUDIES, RANDOMIZATION, SIMULATION

An empirical Bayes method for gene expression analysis in R

Michael G. Schimek and Wolfgang Schmidt

Keywords: Empirical Bayes, Bioconductor project, microarray, multiple testing, R, sparse sequence.

Revised abstract for oral presentation

In recent years the new technology of microarrays has made it feasible to measure expression of thousands of genes to identify changes between different biological states. In such biological experiments we are confronted with the problem of high-dimensionality because of thousands of genes involved and at the same time with small sample sizes (due to limited availability of cases). The set of differentially expressed genes is unknown and the number of its elements relatively small. Due to a lack of biological background information this is a statistically and computationally demanding task.

The fundamental question we wish to address is differential gene expression. The standard statistical approach is significance testing. The null hypothesis for each gene is that the data we observe have some common distributional parameter among the conditions, usually the mean of the expression levels. Taking this approach, for each gene a statistic is calculated that is a function of the data. Apart from the type I error (false positive) and the type II error (false negative) there is the complication of testing multiple hypotheses simultaneously. Each gene has individual type I and II errors. Hence compound error measures are required. Recently several measures have been suggested ([1]). Their selection is far from trivial and their calculation computationally expensive.

As an alternative to testing we propose an empirical Bayes thresholding (EBT) approach for the estimation of possibly sparse sequences observed with white noise (modest correlation is tolerable). A sparse sequence consists of a relatively small number of informative measurements (in which the signal component is dominating) and a very large number of noisy zero measurements. Gene expression analysis fits into this concept. For that purpose we apply a new method outlined in [5]. It circumvents the complication of multiple testing. More than that, user-specified parameters are not needed, apart from distributional assumptions. This automatic and computationally efficient thresholding technique is implemented in R.

The practical relevance of EBT is demonstrated for cDNA measurements. The preprocessing steps and the identification of differentially expressed genes is performed using R functions ([4]) and Bioconductor libraries ([3]). Finally comparisons with selected testing approaches based on compound error measures available in `multtest` ([2]) are shown.

References

[1] Benjamini, Y. and Hochberg, Y (1995). *Controlling the false discovery rate: A practical and powerful approach to multiple testing.* J. Royal Statist. Soc.,

B 85, 289 – 300.

[2] Dudoit, S. and Ge, Y. (2003). *Bioconductor's multtest package.* Report, `http://www.stat.berkeley.edu/∼sandrine`.

[3] Dudoit, S. and Hornik, K. (2003) The Bioconductor FAQ. `http://www.bioconductor.org/`.

[4] Ihaka, R. and Gentleman, R. (1996). *R: A language for data analysis and graphics.* J. Computat. Graph. Statist., **5**, 299 – 314.

[5] Johnstone, I. M. and Silverman, B. W. (2004). *Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences.* To appear in Annal. Statist.

Acknowledgement: The second author acknowledges funding from Gen-AU project "A Comprehensive Disease Bank for Functional Genomics".

Affiliation: Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, Auenbruggerplatz 2, A-8036 Graz, Austria, michael.schimek@meduni-graz.at.

Multiple Imputation and Model Selection in Cox Regression: Estimating the Risk of Dementia in Elderly

M.Schipper, M.M.B. Breteler, Th. Stijnen
Dept. of Epidemiology & Biostatistics,
Erasmus MC, Rotterdam, the Netherlands

Missing values in the observed predictor variables complicate the derivation of an adequate prognostic time to event model. We suggest an approach with multiple imputation via van Buurens MICE library followed by stepwise model selection. In every step we pool the coefficients (and its variance-covariance matrix) of the Cox proportional hazard models, each based on one of a (small) number of imputed datasets. A generalized Wald-test by Rubin gives suitable diagnostics for the next step in the model selection procedure.

Once we know how to do the model selection on a series of imputed datasets using a proportional hazards approach, we can repeat this procedure any number of times in a bootstrap setting. In this way it is possible to assess the calibration and discrimination abilities of our final model following the approach of Harrell et al. The bootstrap also enables to estimate the shrinkage factors for the final model to get a better calibration.

Because of its flexibility and extent R is an excellently suitable environment to program the whole model selection procedure. Choosing an object-based approach, we can even use default R model selection functions.

In an example we apply this modeling strategy to derive a model that assesses the risk of developing dementia over time. The example is based on The Rotterdam Study.

A population based prospective cohort study in Rotterdam of about 8,000 people of 55 years and above. During 10 years of follow up, over 400 cases of dementia were recorded. Initially there are 39 covariates of interest. Although only 9missing, deletion of incomplete cases leads to a loss of over 70the cohort. Therefore multiple imputation and application of the aforementioned model selection procedure seems an appropriate approach here.

References:

[1] Harrell FE Jr, Lee KL, Mark DB. Tutorial in Biostatistics. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Stat Med 15:361-387, 1996

[2] Rubin DB. Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York, 1987

[3] van Buuren S, Oudshoorn CGM. Flexible multivariate imputation by MICE. Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054, 1999

Modelling water flux

Martin Schlather
Université du Luxembourg, Luxembourg

9th February 2004

The modelling of water infiltration in soil is satisfactorily solved only at laboratory scale, using PDEs, especially the so-called Richards' equation. At lysimeter scale, i.e. small field scale, PDEs present various difficulties, such as high computer costs, vaguely known initial conditions and sparse information about the spatial variation of material properties. Consequently, various ways to attack the problem are pursued. The more traditional one is to replace the unknown parts, especially the spatially variable conditions, by stochastic models. On the other hand, simpler, stochastic models that do not try to describe the physical processes in all details are also sought.

The scope of the R package SoPhy is to provide tools for the analysis and the modelling of infiltration experiments at lysimeter scale including both physically based methods and novel stochastic approaches. In the talk, some of the features are presented. The interactive surface for physically based modelling incooperates deterministic information such has atmospherical data and material constants, but also stochastic models for the spatial variability of the material properties and for the distribution of the roots and the stones, see Fig. 3 for a snapshot. Fig. 1 and Fig. 2 show simulation results of a simple, purely stochastic model. The model consists of two components, the availability of water near the surface and the tortuosity of the paths. The former is modelled by a two-dimensional random field, whereas the latter is constructed from two independent random fields. The depth of the water front is then a function of the tortuosity and the water availability. Great depths are reached if locally, the tortuosity is small and the availability of water is high.

SoPhy V1.0 is based on the contributed R packages RandomFields V1.1 and SoPhy V0.91. The latter is essentially an R wrapper for SWMS_2D by Šimunek et al. (1994), a Fortran program for solving the Richards' equation using the Galerkin finite element method. SoPhy V1.0 will appear in summer this year.

References

J. Šimunek, T. Vogel, and M. Th. van Genuchten. The SWMS_2D code for simulating water flow and solute transport in two-dimensional variably saturated media, version 1.21. Technical Report 132, U.S. Salinity Laboratory, Riverside, California, February 1994.

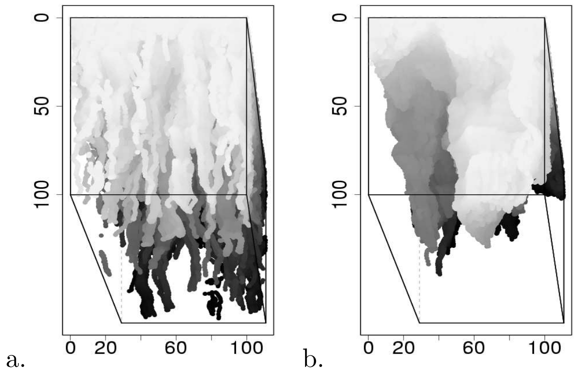


Figure 1: Influence by the scale parameter for the spatial dependence structure of the water availability: a. low value, b. high value for the scale parameter.

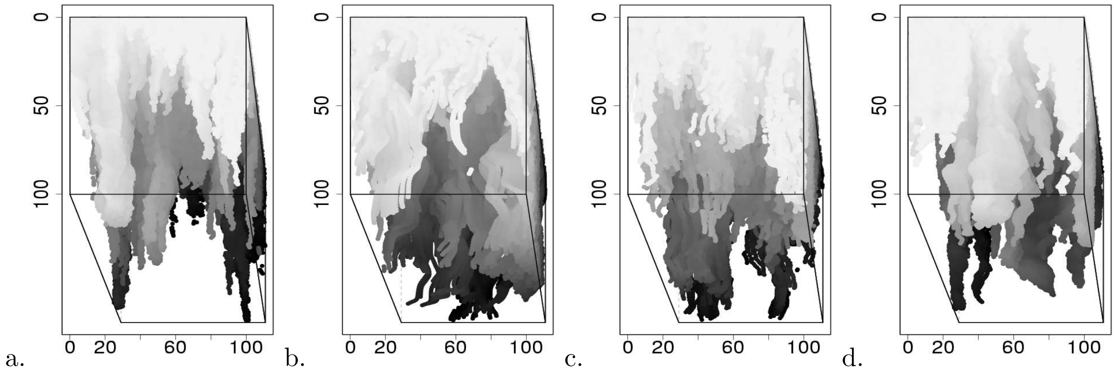
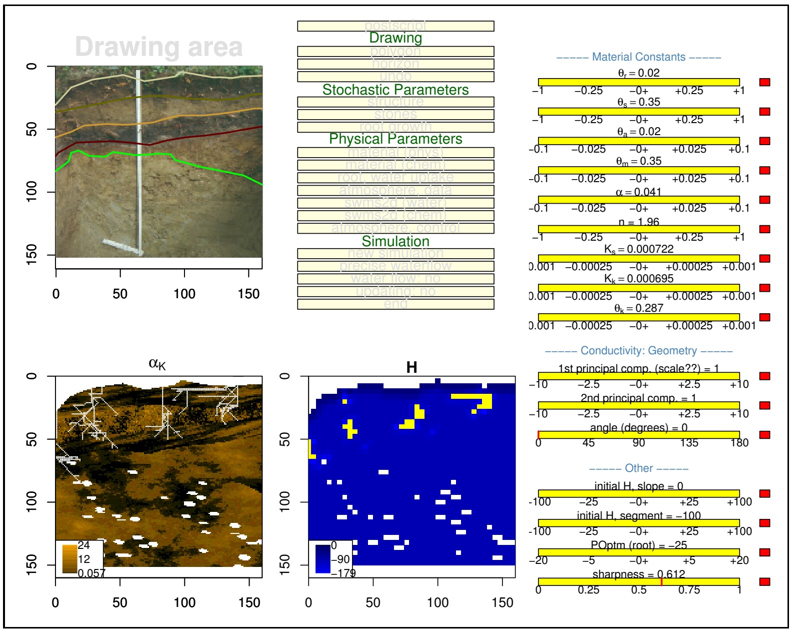


Figure 2: Influence by the scales for horizontal and vertical dependence structures of the tortuosity: In a. and b. the value for the vertical scale parameter is varied; in c. and d. the value for the horizontal scale parameter; a. and c. low values; b. and d. high values.



$\theta_r = 0.02$

$\theta_s = 0.35$

$\theta_a = 0.02$

$\theta_m = 0.35$

$\alpha = 0.041$

$n = 1.96$

$K_s = 0.000722$

$K_k = 0.000695$

$\theta_k = 0.287$

1st principal comp. (scale??) = 1

2nd principal comp. = 1

angle (degrees) = 0

initial H, slope = 0

initial H, segment = −100

POptm (root) = −25

sharpness = 0.612

Figure 3: Snapshot of the interactive surface for physical water flux modelling; upper left: the invetsigated profile; bottom left: simulation of the material including stones and roots; bottom center: calculated matrix potential; top center: inactive main menu; right: active submenu.

Determining Extreme Returns on Stock Indices

Harald Schmidbauer[*]
Istanbul Bilgi University
Istanbul, Turkey

Abstract

Consider a time series of daily returns on a stock index. How can we tell if the return on an arbitrary day is an unexpectedly high gain (or loss)? There is a number of reasons why the answer to this question is important. For example, it serves as the basis for an investigation if investors overreact or underreact in the short run — for example, if, in the given series, a huge gain (loss) is usually followed by a significant loss (gain), the investors may be suspected to overreact, which might even lead to a point where the market is no longer efficient.

It is not meaningful to regard exceedances of a quantile of the entire return series (say, those days with returns above the 95% quantile or below the 5% quantile) as unexpectedly high gains or losses. Obviously, there are good reasons to formulate investors' expectations about tomorrow's returns in terms of a time series model fitted to the series up to today, without using future data. This necessitates the estimation of a sequence of time series models, whose parameters are allowed to evolve in order to reflect changing investor experience and expectation.

If parameter estimation is entirely carried out in the statistical computing language R, we will be faced with the problem of long computing time. This is mainly due to the evaluation of the likelihood function, which has to be computed recursively. Therefore, we propose to call a C routine to compute the likelihood and to do everything else in R. Results are presented for several models, including the usual GARCH, a threshold GARCH, and an exponential GARCH.

[*] e-mail: harald@bilgi.edu.tr

Rho: An Integrated System for R-based Web Applications

Chad Shaw, Nathan Whitehouse and Andrew Young
Department of Molecular and Human Genetics, Baylor College of Medicine
Houston, Texas USA

The central benefits of R analysis include (a) maintanance of state across sessions (b) rapid development of customized software and (c) the extensive collection of community available Rpackages. Unfortunately, the full benefits of R are often unavailable in web-deployment of R analysis. To bring the full power of R to end users we have created Rho – an integrated system for web deployment of R analyses. Our system is servlet based, supports interactive graphics, and can maintain user state. Because we maintain state across sessions, the system supports "dual use"--- where the same analysis objects can be manipulated in "pure R" (through the command line interpreter) as well as "automated R" over the web interface. A single servlet instance can support a variety of applications and concurrent users. The system manages projects, users and data sources. The implementation of these enterprise scale features is mirrored within R itself as well as through the web. The system extends the R compilation machinery to allow immediate deployment of Rpackages. Rho is currently in use through a number of web applications related to bioinformatics, yet the infrastructure is general enough to support any Rpackage. We have experimented with many client interfaces including java applets, SOAP services, and pure HTML. We present a sample application based on microarray analysis and a web-service for analysis of genelist content.

Using R to Analyze Space Use by Wildlife

Alan K Swanson, Jennifer W. Sheldon and Robert L. Crabtree
Yellowstone Ecosystem Research Center, Bozeman, Montana, USA.

Wildlife researchers have used a variety of "black-box" computer programs to generate bivariate pdf's (known as utilization distributions in the wildlife literature) from a spatial point pattern of animal relocations. These competing programs often give conflicting results when used to estimate parameters of wildlife space use such as home range area. The authors demonstrate the use of two-dimensional kernel smoothing functions available in R to generate the pdf's, and then show how these pdf's may be used to estimate various parameters. Estimates made using gaussian kernels are compared to those derived using quartic kernels, and the effect of sample size is examined. Various statistics for comparing two bivariate pdf's are presented. Bandwidth selection is discussed.

JGR: a unified interface to R

Martin Theus, Markus Helbig
Augsburg University

There is no native interface to R other than the command line. Since the majority of the users want to have a bit more comfortable interface to R, various, platform dependent GUIs have been developed. They are all different, support different features and are usually insufficient for the expert user.

JGR (speak jaguar) is a Java Gui for R. It addresses primarily the non-developer R-user. JGR features a build in editor with syntax highlighting and direct command transfer, a spreadsheet to inspect and modify data, an advanced help system and an intelligent object navigator, which can, among other things, handle data set hierarchies and model comparisons. JGR also allows to run parallel R sessions within one interface. Since JGR is written in Java, it builds the first unified interface to R, running on all platforms R supports.

Using meta-programming to make BRugs from BUGS and R

Andrew Thomas

WinBUGS is a successful Bayesian statistical modeling package using Markov Chain Monte Carlo methods. WinBUGS has been written as a series of small cooperating components using the Oberon programming language. R is a statistical and data analysis package which can be interfaced to C code using dynamic link libraries. We have developed a dynamic link library, incorporating a component loader, that allows R to make use of WinBUGS components. Meta-programming is used to communicate between the R and Oberon software.

RNGstreams — An R package for multiple independent streams of uniform random numbers

Günter Tirler, Pierre L'Ecuyer and Josef Leydold

The core R implementation of uniform random number generation uses a global procedure that can be used for random generation purposes. It can be changed by means of the RNGkind function. However, this approach is not always convenient for simulation. In this contribution we propose a new approach that uses classes where independent instances of uniform random number generators can be created. Random streams can then be generated by methods of this class. Moreover these streams can be independently reset to their starting values. It is also easy to generate common or antithetic random variates as they required for variance reduction techniques. Another important feature is its ability to use this approach for parallel computing, as this usually requires independent streams of uniform random numbers on each node. In a first implementation these instances can be used together to replace the build-in RNGs. But with such a concept it would be possible to run random routines with different RNGs when the instance of the random stream is given as an argument. Yet we have implemented an interface to two sources of uniform random number generators: the rngstreams library by P. L'Ecuyer (http://www.iro.umontreal.ca/lecuyer) and O. Lendl's implementation of various types random number generators (LCG, ICG, EICG), see http://statistik.wu-wien.ac.at/prng/.

Signal noise decomposition of financial data: An infrequent trading analysis

Helgi Tomasson

The observed transaction prices on a stock market at discrete time points are assumed to be a sample from a continuous time-value process. The theory of an efficient market is used as motivation for a random-walk type model. The fact that bid-ask spread and other microstructure phenomena exist is accounted for by adding a noise term to the model. Models for elementary detrending based on stochastic processes in continuous time are set up. For empirical analysis, they are formulated in state-space form, and calculations are performed with the Kalman-filter recursions. The result is an analytical way of decomposing the observed transaction price change into a value innovation and a market noise component. The respective innovation standard deviations and market noise standard deviations are easily interpreted. Some alternative stochastic structures are considered and applied to data from the Iceland Stock Exchange. Algorithm is implemented in the statistical language R combined with Fortran subroutines.

Keywords: Diffusion processes, state-space models, financial data, infrequent trading

Rough Set based Rule Induction Package for R

Shusaku Tsumoto and Shoji Hirano
Department of Medical Informatics,
Shimane Medical University, School of Medicine,
Enya-cho Izumo City, Shimane 693-8501 Japan

Abstract

Rough set theory is a framework of dealing with uncertainty based on computation of equivalence relations/clases. Since a proability is defined as a measure of sample space, defined by equivalence classes, rough sets are closely related with probabilities in the deep level of mathematics. Furthermore, since rough sets are closely related with Demster-Shafer theory or fuzzy sets, this theory can be viewed as a bridge between classical probability and such subjective probabilities. Also, this theory is closely related with Baysian theories.

The application of this theory includes feature selection, rule induction, categorization of numerical variables, which can be viewed as a method for categorical data analysis. Especially, rough sets have been widely used in data mining as a tool for feature selection, extracting rules (if–then rules) from data. Also, this theory includes a method for visualization, called "flow graphs."

This paper introduces a rough set based rule induction package for R, including: (1) Feature selection: rough sets call a set of independent variables "reducts." This calculation is based on comparisons between equivalence classes represented by variables with respect to the degree of independence. (2) Rule Induction: rough sets provide a rule extraction algorithm based on reducts. Rules obtained from this subpackage are if–then rules. (3) Discretization (Categorization of Numerical Variables): discretization can be viewed as a *sequential* comparison between equivalence classes given in a dataset. (4) Rough Clustering: calculation of similarity measures can be also viewed as that of comparisons between equivalence classes. Rough clustering method gives a indiscernbility-based clustering with iterative refinement of equivalence relations. (5) Flow Graph: this subpackage visualizes a network structure of relations between given variables. Unlike bayesian networks, not only conditional probabilies but also other subjective measures are attached to each edge. (6) Rule Visualization with MDS: this subpackage gives a visualization approach to show the similar relations between rules based on multidimensional scaling. The usage of R gives the following advantages: (a) Rough set methods can be easily achieved by fundamental R-functions, (b) Combination of rough set methods and statistical packages are easily acheived by rich R-packages. In the conference, several aspects of this package and experimental results will be presented.

Keywords: Data Mining, Rough Sets

Application of R to Data Mining in Hospital Information Systems

Shusaku Tsumoto and Shoji Hirano
Department of Medical Informatics,
Shimane Medical University, School of Medicine,
Enya-cho Izumo City, Shimane 693-8501 Japan

Abstract

Hospital information systems have been introduced since 1980's and have stored a large amount of data of laboratory examinations. Thus, the reuse of such stored data becomes a critical issue in medicine, which may play an important role in medical decision support. On the other hand, data mining from the computer science side emerged in early 1990's, which can be viewed a re-invention of what statisticians, especially those on exploratory data analysis, had proposed in 1970's. The main objective of the present data mining is to extract useful patterns from a large mount of data with statistical and machine learning methods. Especially, it has been reported in medical field that a combination of these two methodologies are very useful: Machine learning method is useful for extracting "hypotheses" which may not be significant from the viewpoint of statistics. After deriving these hypotheses, statistical analysis is used for its validity. Thus, it has been expected that combination of these two methodologies will play an important role in medical decision support, such as intra-hospital infection control, detection of risk factors.

This paper report an application of R to data mining in hospital information systems. As a preliminary tool, we developed a package for data mining in medicine, including the following procedures: (1) Interface for Statistical Analysis, which is based on Rcmdr. (2) Rule Induction, which supports associaton and rough set-based rule induction method. (3) Categorization of Numerical Variables: detection of cut-off point is very important in medical diagnosis. Several methods proposed in data mining and medical decision making are implemented. (4) Clustering, based on R packages. Also, this package supports rough clustering, which gives a indiscernbility-based clustering with iterative refinement of equivalence relations in a data set. (5) Temporal Data Mining (Multiscale Matching and Dynamic Time Warping): these methods have been introduced for classification of long-term time series data. These methods output a distance between two sequences, which can be used for clustering methods. The usage of R gives the following advantages: (a) Rough set methods can be easily achieved by fundamental R-functions, (b) Combination of rough set methods and statistical packages are easily acheived by rich R-packages. In the conference, several aspects of this package and experimental results will be presented.

Keywords: Data Mining, Hospital Information System, Time Series Analysis

Using R in other Applications
Various practical ways to integrate an own software and R

Simon Urbanek
University of Augsbung

R provides a flexible, computational framework for statistical data analysis. Due to its modularity the spectrum of available functions is being continuously extended. This functionality is typically used from the R command shell where the user performs an analysis. On the other hand many applications provide a different interface to the user, mainly to reach different target groups of users, but are in need of the functionality R provides. Instead of re-implementing such functions, it is more efficient to integrate R in the application or vice versa. In this talk we present the various methods of integration: linking own code to R, linking R into an own application, using inter-process communication or other packages offering specific interfaces to R. The focus is placed on practical examples using the various techniques, as well as explanation of the implementation and limitations of each approach. In particular the methods are illustrated on a web server application using R and graphical interfaces to R.

Exploiting Simulation Features of R in Teaching Biostatistics

Zdeněk Valenta

European Center for Medical Informatics, Statistics and Epidemiology, Institute of Computer Science AS CR, Prague, Czech Republic

Conveying or illustrating some fundamental statistical principles may become a challenge when teaching students who do not possess a sufficient theoretical background. More often than not they do rely on a hands-on experience in understanding different theoretical concepts.

One of such concepts in statistics in general is, for example, that of confidence intervals. For example, the idea that the true mean of a normal distribution, a fixed and unknown constant μ, will rest within some interval with random endpoints with a pre-specified probability, say 95%, may to medical students at large represent an equally remote concept such as that of average relative frequency of covering the true population mean μ by such intervals constructed during the process of repeated random sampling from the target population.

Our experience is that explaining similar concepts to medical or other interdisciplinary students using R seems quite rewarding. Students can each perform their own simulations, discuss and compare the corresponding results in the class and observe that though not being identical the results are all consistent with the probabilistic statements they wondered about. It suddenly becomes a relevant learning experience to many of them.

In our classes we endeavoured to illustrate the concept of confidence intervals using a simulated birth weight data set that might actually represent the birth weights of children born in the Czech Republic. An example of the real birth weight data for children born in Boston City Hospital, which served as an inspiration to our simulated data sets, is discussed with regard to introducing the concept of confidence intervals e.g. in Rosner[1].

A sample R code is used to illustrate the meaning of confidence intervals. We wrote an R function counting the relative frequency of occurrences in which the hypothetical true population mean is actually being "covered" by the random confidence intervals during the process of repeated sampling. Examples are shown based on 1000 simulations and a sample size of 100 where the sample birth weights are drawn from a normal population with the mean of 3.2 kg and standard deviation 0.6 kg.

The examples document versatility of R environment which serves very well not only for the research purposes of biostatisticians, but also in teaching biostatistics to medical students and medical professionals, where the hands-on experience may be essential for grasping more difficult statistical concepts.

References

[1] Rosner B.: Fundamentals of Biostatistics, 4th Edition, Duxbury Press 1994, Wadsworth Publishing Company, 10 Davis Dr., Belmont, CA 94002, U.S.A., ISBN 0-531-20840-8

Using R and SPSS simultaneously in teaching statistics: A tricky business

Vandewaetere, M., Van den Bussche, E. and Rosseel, Y.
Department of Data Analysis
Ghent University
Henri Dunantlaan 1, B-9000 Gent, Belgium.

Applied statistics and data analysis are crucial courses in a psychology department. The authors are involved in teaching introductory statistics and data analysis to both undergraduate and graduate students. Naturally, statistical software packages play an important role in our courses, aiding students in understanding and applying frequently used statistical issues. In most course material, theoretical sections describing a particular topic (say, ANOVA) are immediately followed by practical examples, with detailed instructions (even screenshots) on how the procedure should be carried out by a computer program, and how the output should be interpreted.

The statistical software package of choice in our department is, and will probably always remain, SPSS. However, given the growing popularity of R, we wanted to give our students the opportunity to use this non-commercial, free and open-source alternative. For various reasons, it was not possible to replace SPSS by R. Consequently we opted for the simultaneous use of SPSS and R in our computer practica, giving the students the choice of using the computer program they prefer.

We have been using R in this fashion for a few years now, and we have encountered major advantages, but also many difficulties in the simultaneous use of R and SPSS. In most cases, these difficulties are due to subtle statistical differences between the two packages, obvious to experienced statisticians, but often hard to explain to students in a psychology department with little mathematical background. Typical issues are the following:

- The '`anova`' command in R produces sequential (or Type I) sum of squares, while SPSS uses Type III sum of squares per default.
- Testing a linear hypothesis of the form $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$ (where $\boldsymbol{\beta}$ is a vector of model parameters, and $\mathbf{L}$ is a hypothesis matrix), especially in a linear model with categorical predictors is a tricky business for the unexperienced user:
  - a custom command is not availabe in R (per default); it is available in several packages (eg the '`car`' package)
  - in the 'GLM' procedure in SPSS, redundant parameters (typically corresponding to the last level of a factor) are fixed to zero but still included in the parameter vector; in R, redundant parameters are dropped from the parameter vector. As a result, the parameter vectors differ in length, and the $\mathbf{L}$ matrix has a different number of columns in R and SPSS.
- Testing a linear hypothesis of the form $H_0 : \mathbf{L}\boldsymbol{\beta}\mathbf{M} = \mathbf{C}$ in the context of MANOVA or (the multivariate approach to) repeated measures is currently not possible in R without writing your own code.
- Several basic procedures (calculating the a priori power for t-tests, getting a correct one-sided p-value for a t-test, etc. ...) are currently not possible in SPSS.

The discrepancies between the two programs make it hard for us teachers to create concise and comprehensive course material integrating both software packages (SPSS and R). During the presentation, we will illustrate these issues with some practical examples, and discuss various ways of how we have dealt with them.

Merging R into the web: DNMAD & Tnasas, web-based resources for microarray data analysis.

Juan M. Vaquerizas, Joaquín Dopazo and Ramón Díaz-Uriarte
Bioinformatics Unit, CNIO (Spanish National Cancer Centre)
Melchor Fernández Almagro 3, 28029 Madrid, Spain.

April, 2004

Abstract

DNMAD and Tnasas are two web-based resources for microarray data analysis (normalization and class prediction), that implement R code into a web server configuration. These tools are included in GEPAS, a suite for gene expression data analysis [4].

The first tool we present is called DNMAD that stands for "**D**iagnosis and **N**ormalization for **M**icro**A**rray **D**ata". DNMAD is essentially a web-based interface for the Bioconductor package limma [7]. DNMAD has been designed to provide a simple and at the same time robust and fast way for normalization of microarray data. We provide global and print-tip loess normalization [9]. We also offer options for the use of quality flags, background adjustment, input of compressed files with many of arrays, and slide-scale normalization. Finally, to enable a fully featured pipeline of analysis, direct links to the central hub of GEPAS, the Preprocessor [5] are provided. The tool, help files, and tutorials are available at http://dnmad.bioinfo.cnio.es.

The second tool we present is called Tnasas for "**T**his is **n**ot **a** **s**ubstitute for **a** **s**tatistician". Tnasas performs class prediction (using either SVM, KNN, DLDA, or Random Forest) together with gene selection, including cross-validation of the gene selection process to account for "selection bias" [1], and cross-validation of the process of selecting the number of genes that yields the smallest error. This tool has mainly an exploratory and pedagogical purpose: to make users aware of the widespread problem of selection bias and to provide a benchmark against some (overly) optimistic claims that occasionally are attached to new methods and algorithms; Tnasas can also be used as a "quick and dirty" way of building a predictor. Tnasas uses packages e1071 [2], randomForest [6], and class (part of the VR bundle [8]). The tool, help files, and tutorials are available at http://tnasas.bioinfo.cnio.es.

Both tools use a Perl-CGI script that checks all the data input by the user and, once checked, sends it to an R script that performs the actions requested by the user. To merge the Perl-CGI script with the script in R we use the R package CGIwithR [3]. Source code for both applications will be released under the GNU GPL license.

References

[1] C. Ambroise and G.J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA*, 99: 6562–6566, 2002.

[2] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingess. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.
http://cran.r-project.org/src/contrib/PACKAGES.html#e1071

[3] D. Firth. CGIwithR: Facilities for the use of R to write CGI scripts.
http://cran.r-project.org/src/contrib/PACKAGES.html#CGIwithR

[4] J. Herrero, F. Al-Shahrour, R. Díaz-Uriarte, Á. Mateos, J. M. Vaquerizas, J. Santoyo, and J. Dopazo. GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res*, 31:3461–3467, 2003.

[5] J. Herrero, R. Díaz-Uriarte, and J. Dopazo. Gene expression data preprocessing. *Bioinformatics*, 19:655–656, 2003.

[6] A. Liaw, M. Wiener, with original Fortran code from L. Breiman and A. Cutler. randomForest: Breiman's random forest for classification and regression.
`http://cran.r-project.org/src/contrib/PACKAGES.html#randomForest`

[7] G. K. Smyth, M. Ritchie, J. Wettenhall, and N. Thorne. limma: Data analysis, linear models and differential expression for microarray data.
`http://www.bioconductor.org/repository/release1.3/package/html/limma.html`

[8] B. D. Ripley and W. Venables. VR: Functions and datasets to support Venables and Ripley, 'Modern Applied Statistics with S' (4th edition). `http://cran.r-project.org/src/contrib/PACKAGES.html#VR`

[9] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30:e15, 2002.

Mining some medical information in a fee database

Pablo E. Verde [*][†], Christian Ohmann [‡],
Christof Kugler [§] and Max Geraedts [¶]

February 10, 2004

Abstract

Billing collection systems of medical insurance institutions generate a massive amount of data, which are used for money transaction and for further administrative purposes. Recently, public health researchers highlighted the possibility that these databases could contain useful information for health-care evaluation. In this work, the R system is used to analyze a snapshot of 5 years of information of a fee database containing the clinical history of 38,153 heart illness patients. Two main problems were investigated. The first one involves learning about the dynamic of the course of patients clinical interventions. The second problem concerns assessing the effect of the first clinic of intervention. This analysis illustrates statistical data mining with R. The course of patients interventions is reduced to a multivariate contingency table with transition frequencies in the cells. Log-linear modelling is applied to analyze a Markov dependency structure and to assess baseline effects. This analysis shows that stationary Markov models, commonly used in health-care evaluation could be an oversimplified modelling approach.

KEYWORDS: data mining, health-care evaluation, Markov chains, log-linear models, over-dispersion.

[*] Coordination Center for Clinical Trials, Heinrich-Heine University Duesseldorf

[†] E-mail: pabloemilio.verde@uni-duesseldorf.de

[‡] Coordination Center for Clinical Trials, Heinrich-Heine University Duesseldorf

[§] Institute for Integrative Care in Medicine - University of Giessen

[¶] Public Health Program, Heinrich-Heine University Duesseldorf

Graph Data in R, a User's Perspective

Chris Volinsky

February 14, 2004

Data which can be represented by a graph (ie. as a collection of nodes and edges between those nodes) have been generating a lot of interest lately. Graphs have become an increasingly popular way of representing data in many different domains, ranging from web connectivity to disease transmission to relationships between monks in closed communities, and many more. Traditionally, the tasks of graph creation, layout, manipulation, modelling and rendering might be handled by several different tools. In the last few years the R community has developed packages for integrating many of these tools into one cohesive environment. They include:

- **graph**: The R **graph** library is a toolbox of graph creation and manipulation tools that allow a graph to be created from already existing R structures. The representation of the graph can be as a list of nodes and edges, or as a distance matrix, depending on the nature and size of the graph. **graph** contains many useful tools for graph manipulation and modelling, including boundary calculations, graph distance, and random graph probability measures.
- **Rgraphviz**: The **graphviz** library is a commonly used graph layout tool. The **Rgraphviz** package supports two major graph layout algorithms, dot (a hierarchical layout) and neato (a radial layout).
- **Rggobi**: **Ggobi** is a powerful interactive data visualization tool which has developed some nice features for graph rendering and manipulation. **Rggobi** was developed as an interface for R to interact with Ggobi in a seamless manner, and allows the user to do brushing and spinning and many more interactive data tasks.

Using these three tools together provides a powerful toolbox for data analysis with graph data. An R object can be turned into a graph object using the **graph** library, **Rgraphviz** is then used to generate a snazzy layout, which is then passed to **Rggobi** tools to render the plot so that the user can interact with the data.

I have used these tools to help in analyzing telecommunications data at AT&T. I look at "Communities of Interest", or COI, which are small subgraphs of our massive callgraph, which we use as a network signature for the purposes of fraud detection. In this talk, I will present some COI data in the context of a fraud application, and show how I have used the graph tools above to learn about, analyze and model these graphs.

A stateful R web-interface, leveraging XML and XSLT

Richard P. Waterman
Analytic Business Services
And Department of Statistics, The Wharton School
University of Pennsylvania

Thomas A. Hurley
Analytic Business Services [*]

February 15, 2004

We will present work in progress that integrates R as the engine behind a stateful web friendly analytics package.

The work is funded by an NIH/SBIR grant to create user friendly software that will guide users through propensity score analysis for causal modeling (2). As such the likely users will come from a variety of backgrounds, for example epidemiology and marketing, where a command line driven interface will not work for most people. Further, providing a web browser based front end will reduce the ownership costs, both technologically and materially to many users.

One of the key design decisions behind the project is to only use open standards, and the project makes full use of the XML functionality now available in R (1) through the StatDataML representation of R objects.

In order to provide for a session memory a lightweight server termed the clientBrain (implemented in JAVA) has been created. The role of this server is to accept requests from the client (via a TomCat servlet engine) and pass these requests onto R, by way of an additional servlet termed the R-server.

Communication through to R is by socket connections, and we make full use of R's various connections; pipes and sockets as well as simple file reading and writing.

The presentation layer is governed by extensive use of XSLT, the Extensible Stylesheet Language Transformations. XSLT is used in two places – first the StatDataML representations are transformed into a canonical representation, and then this canonical representation is transformed into a browser friendly format, for example HTML or SVG.

These transformed XML documents are kept in memory on the clientBrain, so that additional requests from the browser for transformation based requests, for example sorting, are accomplished via dynamic XSLT edits, rather than resubmitting requests to R.

The rationale behind the two stage transform, first from StatDataML to a canonical XML form is that R is potentially one of a number of plug-in components to the architecture. Another natural one are user driven SQL queries to a database, the responses to which would not be in the StatDataML format.

[*] NIH/SBIR Grant 1 R44 CA094611-01

We will share some of the difficulties that we have encountered and describe further possibilities for this architecture, in particular with regard to our business, which is an internet based quantitative market research company.

In summary, the philosophy of this implementation is to only use R for it's analytic capabilities, with all representation and presentation decisions relying on the XML/XSLT paradigm.

References

[1] D. Meyer, F. Leisch, T. Hothorn, and K. Hornik, *StatDataML: An XML format for statistical data*, Computational Statistics, (2004, Forthcoming).

[2] P. Rosenbaum and D. B. Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika, 70 (1983), pp. 41–55.

Use R for Teaching Statistics in the Social Sciences?!

Adalbert Wilhelm
School of Humanities and Social Sciences,
International University Bremen,
P.O. Box 750 561, D-28725 Bremen, GERMANY
e-mail: a.wilhelm@iu-bremen.de

April 15, 2004

Abstract

Teaching statistics to students in the social sciences is a challenging task. Typically, the students lack enthusiasm to enter into the formal world of statistical models and mathematical derivations. On the other hand, empirical studies are ubiquitous in the social sciences and students learn quickly that they need active knowledge on statistical methods to successfully accomplish their courses and develop expertise for their future jobs. Introducing statistical concepts by working with real life data sets and stressing the interpretation of statistical results is the method of choice to make statistics attractive to social science students. Quite naturally, computers and statistical software are used to ease necessary calculations and to familiarize students with the relevant computer output. What are important points in choosing a statistical software to be used in courses for social sciences?

In many places, SPSS is the software of choice for this purpose. Although SPSS has itself established to be the standard statistical software package for the social sciences, it has some drawbacks as a teaching tool. SPSS as many other software comes with a graphical user interface (GUI) with nice pull-down menus from which the appropriate analysis methods can be easily chosen. While GUI's can be straightforwardly used by anyone who is able to read, command-line interfaces constitute a barrier for the beginners because they require knowledge of a particular syntax. However, managing your way through a sequence of pull-down menus and pop-up windows to start the analysis you aim at is a major effort when using a GUI. Once you are familiar with the software, you'll easily find your way, but beginners often loose track and are bound to learn by trial and error. In the classroom, even when you have access to modern multi-media teaching labs, it is extremely difficult for students to keep track of the instructor clicking her path through the jungle of options and parameter choices. Hence there are a variety of instances for which the use of a command-line program is preferable.

In this paper, we have a close look at three different statistical programs, SPSS, Data Desk and R, and judge their usability for teaching purposes in the social sciences. From various aspects, e.g. costs, functionality, didactical strength, we shed a light on these packages and assess their capabilities.

USE R FOR PEPTIDE MASS FINGER PRINTING.

E. W. Wolski[1 2], T. Kreitler[1], H. Lehrach[1], J. Gobom[1], K. Reinert[2]

Abstract. We use R in order to generate a greater specificity and sensitivity of protein identification by Peptide Mass Fingerprinting (PMF). This is achieved through analysis, detection and correction of measurement errors, by filtering of MS data prior to database searches and by analysis of the search results. These functions are implemented as an add-on packages to the freely-available statistical software, R.

1. Introduction

Protein identification using PMF data is performed by comparing experimentally determined peptide masses of a protein cleaved by a specific protease to in silico generated PMFs of known protein sequences[4]. Mass lists are generated by assigning m/z values to each monoisotopic signal in the mass spectra. The mass lists are then used for protein identification by search engines.

Using R[2] we

- analyze the measurement error and calibrate the masses
- find contaminants and remove them
- submit peak-lists to the identification software Mascot[5], and analyse the search result.

For this tasks we use functionality provided by R e.g.: spline functions (`R/modreg`), linear regression, functions for matrix computation, descriptive statistic functions (`R/base`) and clustering algorithms (`R/mva`). In addition we implemented functions for pairwise protein sequence comparison in the package `R/pairseqsim` and for communication with web servers in the package `R/httpRequest`.

2. R Packages for Peptide Mass Fingerprinting

R/msbase. The m/z and intensity value pairs assigned to each peak in a spectrum are stored along with the position of the sample on the MALDI sample support in the `Massvector` object. State of the art mass spectrometers can analyze several hundred samples on a single sample support. The resulting collection of `Massvectors` is modeled by the class `Massvectorlist`. The class `Massvector` provides methods for peak-list manipulation e.g. fuzzy union or fuzzy intersect. They have to be fuzzy because the masses have an measurement error. Also methods to compute distance and similarity measures on peak intensities e.g. correlation, spectral angle, similarity index, or binary measures like relative mutual information and many more, are implemented.

[1] Max Planck Institute for Molecular Genetics.

[2] Free University of Berlin, Institute of Informatics.

R/mscalib. adds methods for calibration and filtering to the `Massvector` and `Massvectorlist` classes. The masses can be calibrated by internal, external[1], pre-[6] and set based internal calibration. The data can be analyzed and filtered for abundant masses, chemical noise and significant mass differences.

R/msmascot. The calibrated and filtered peak-list can be submitted to the Mascot search software out of R. The package provides classes to store and visualize the search results. For example the class `Mascotsearch`, which stores the result of a single search, implements a plot function which visualizes the scores, number of matched peptides and the sequence coverages of fetched hits. To help to interpret the identification result, in cases when one peak-list matches multiple proteins, the summary method of the `Mascotsearch` class clusters the protein sequences and the theoretical digest. The search result obtained by submitting a `Massvectorlist` for a search is modeled by the class `MascotsearchList` which implements methods which generate summaries of e.g. the number of identification.

3. Summary

The *R/PMF* packages provide a variety of visualization, summary, analysis, filtering, and calibration methods. The combination of different calibration and filtering methods can significantly improve protein identification. Combining the implemented functionality with *Sweave*[3] enables to generate experiment and protein identification reports in high throughput. Further prospects for extension of the *R/PMF* packages include functions for peak picking in mass spectra.

References

1. J. Gobom, M. Mueller, V. Egelhofer, D. Theiss, H. Lehrach, and E. Nordhoff, *A calibration method that simplifies and improves accurate determination of peptide molecular masses by maldi-tof-ms*, Analytical Chemistry **74** (2002), no. 8, 3915–3923.
2. Ross Ihaka and Robert Gentleman, *R: A language for data analysis and graphics*, Journal of Computational and Graphical Statistics **5** (1996), no. 3, 299–314.
3. Friedrich Leisch, *Sweave and beyond: Computations on text documents*, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), 2003.
4. D. J. C. Pappin, P. Hojrup, and A. J. Bleasby, *Rapid identification of proteins by peptide-mass fingerprinting*, Curr. Biol. **3** (1993), 327–332.
5. D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, *Probability-based protein identification by searching sequence databases using mass spectrometry data*, Electrophoresis **20** (1999), no. 18, 3551–3567.
6. A. Wool and Z. Smilansky, *Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting*, Proteomics **2** (2002), no. 10, 1365–1373.

Rmetrics

www.rmetrics.org

An Environment for Teaching Financial Engineering and Computational Finance

Diethelm Würtz
Swiss Federal Institute of Technology, Zürich
Institute for Theoretical Physics, Hönggerberg
CH-8093 Zürich, Switzerland

February 2004

Rmetrics *is a collection of several hundreds of functions which may be useful for teaching "Financial Engineering" and "Computational Finance". This R port was initiated 1999 as an outcome of my lectures held on topics in econophysics at ETH Zürich. The family of the Rmetrics packages includes currently four members dealing with the following subjects:* ***fBasics*** *- Markets, Basic Statistics, Date and Time,* ***fSeries*** *- The Dynamical Process Behind Financial Markets,* ***fExtremes*** *- Beyond the Sample, Dealing with Extreme Values, and* ***fOptions*** *– The Valuation of Options.*

The package **fBasics** covers the management of economic and financial market data. Included are functions to download economic indicators and financial market data from the Internet. Distribution functions relevant in finance are added like the asymmetric stable, the hyperbolic and the inverse normal gaussian distribution function to compute densities, probabilities, quantiles and random deviates. Estimators to fit the distributional parameters are also available. Some additional hypothesis tests for the investigation of correlations, dependencies and other stylized facts of financial time series can also be found in this package. Furthermore, for date and time management a holiday database for all ecclestial and public holidays in the G7 countries and Switzerland is provided together with a database of daylight saving times for financial centers around the world. Special calendar management functions were implemented to create easily business calendars for exchanges. A collection of functions for filtering and outlier detection of high frequency foreign exchange data records collected from Reuters' data feed can also be found together with functions for de-volatilization and de-seasonalization of the data. – A new additional chapter with synonyme functions for Splus like time, date, and time series objects is scheduled for April 2004.

The package **fSeries** covers topics from the field of financial time series analysis including ARIMA, GARCH, Regression, and Feedforward Neural Network modelling. This library tries to bring together the content of existing R-packages with additional new functionality on a common platform. The collection comes with functions for testing various aspects of financial time series, including unit roots, independence, normality of the distribution, trend stationary, co-integration and neglected non-linearities. Furthermore functions for testing for higher serial correlations, for heteroskedasticity, for autocorrelations of disturbances, for linearity, and functional relations are also provided. Technical analysis and benchmarking is another major issue of this package. The collection offers a set of the most common technical indicators together with functions for charting and benchmark measurements. For building trading models functions for a rolling market analysis are available. – A new additional chapter on modeling long memory behavior including moment methods, periodgram analysis, whittle estimator, and wavelet analysis is scheduled for May 2004.

The package **fExtremes** covers topics from the field what is known as extreme value theory. The package has functions for the exploratory data analysis of extreme values in insurance, economics, and finance applications. Included are plot functions for empirical distributions, quantile plots, graphs exploring the properties of exceedences over a threshold, plots for mean/sum ratio and for the development of records. Furthermore functions for preprocessing data for extreme value analysis are available offering tools to separate data beyond a threshold value, to compute blockwise data like block maxima, and to de-cluster point process data. One major aspect of this package is to bring together the content of already existing R-packages with additional new functionality for financial engineers on a common platform investigating fluctuations of maxima, extremes via point processes, and the extremal index. – A new additional chapter on risk measures, stress testing and copulae is scheduled for October 2004.

The package **fOptions** covers the valuation of options including topics like the basics of option pricing in the framework of Black and Scholes, including almost 100 functions for exotic options pricing, including the Heston-Nandi option pricing approach mastering stochastic volatility, and Monte Carlo simulations together with generators for low discrepancy sequences. Beside the Black and Scholes option pricing formulas, functions to valuate other plain vanilla options on commodities and futures, and function to approximate American options are available. Some binomial tree models are also implemented. The exotic options part comes with a large number of functions to valuate multiple exercise options, multiple asset options, lookback options, barrier options, binary options, Asian options, and currency translated options. – A new additional chapter on exponential Brownian motion including functions dealing with moment matching methods, PDE solvers, Laplace inversion methods, and spectral expansion approaches for option pricing is scheduled for August 2004.

Two other packages are currently implemented: **fPortfolio** and **fBonds**. The first package offers functions for the cluster analysis of market data based on already available R functions and new functions for modern portfolio selection and optimization. Beyond the Markowitz approach we have implemented modern risk concepts based on conditional value-at-risk and conditional drawdown-at-risk for the investigation of hedge funds. Several graphical utility functions like heatmaps and others to display multivariate data sets are also part of this package. The second package is just at the beginning and deals with bond arithmetic, with the yield curve, with interest rate instruments, and with replicating portfolios. – The portfolio optimization functions for hedge funds will be available end of March 2004, the remaining parts of the two packages are scheduled for 2005.

The packages are documented in *User Guides* and *Reference Guides*, currently about 800 pages. The packages are made for R under the operating system MS Windows. *Why using R?* To make the software in an academic environment "free" available for everybody, we decided to implement the functions in the framework of R. This offers the possibility, that everybody can modify the existing functions and can contribute additional functions and datasets. Furthermore, in most cases the functions can be ported to SPlus. *Why using MS Windows?* In the financial community Windows is the mostly used operating system. For a broad distribution and acceptance of *Rmetrics*, we decided to develop the software under Windows 2000/XP. But nevertheless, since all source code is available it may be straightforward to adapt and compile the software for other operating systems.

Rmetrics is a collection of R functions having its source in algorithms and functions written by many authors. The aim is to bring the software together under a common platform and to make it public available for teaching financial engineering and computational finance.