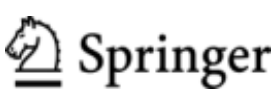
Book of Abstracts

The R User Conference 2006

2nd International R User Conference
June 15–17 2006, Vienna, Austria



Contents

Robust Estimation for Circular Data using R

Claudio Agostinelli *
Dipartimento di Statistica
Università Ca' Foscari di Venezia, Italia

February 23, 2006

Abstract

In this work we study the problems arising when there are outliers in a data set following a circular distribution.

To obtain robust estimation of the unknown parameters the methods of Weighted Likelihood and Minimum Disparity are used. The methods are defined for a general parametric family of circular data.

We investigate the class of Power Divergence and the related Residual Adjustment Function in order to improve the performance of the introduced methods.

The robust behavior and the performance of Weighted Likelihood and Minimum Disparity are studied for the Von Mises (circular normal) distribution and for the Wrapped Normal distribution. Some computational aspects are illustrated. Two examples based on real data set and the results of a Monte Carlo study are presented.

The implementation and the use in `R` of these robust methods (available in package `wle`) together with plot and print functions (available in package `circular`) is also illustrated.

Keywords: Circular data, Disparity measures, Kernel density estimation, Outliers in circular data, Residual Adjustment Function, Robust estimation, Weighted likelihood.

*Dipartimento di Statistica, Università Ca' Foscari di Venezia, San Giobbe, Cannaregio 873, 30121 Venezia, Italia, email: claudio@unive.it

R Modules for Accurate and Reliable Statistical Computing, Perturb package

Micah Altman and Jeff Gill and Michael McDonald

Most empirical social scientists are surprised to find that low-level numerical issues in the software they use can have deleterious effects on the estimation process. In fact, statistical software that appears to be performing in a perfectly adequate fashion can be heneously wrong with revealing such problems. This article is intended to further raise awareness of such issues and to provide tools for detecting and correcting such problems. We develop a set of set of `R` modules that provide two general tools for improving accuracy: a way to measure the *sensitivity* of the results of one's statistical analyses to measurement error and numerical problems; and a method of detecting errors in data that occur in the translation of statistical data from another format. These modules also provide specific methods for checking the integrity of standard estimation techniques in a general way.

R algorithms for the calculation of markers to be used in the construction of predictive and interpolative biplot axes in routine multivariate analyses.

M. Rui Alves[1,2] and M. Beatriz Oliveira[1]

[1]REQUIMTE, Serviço de Bromatologia, Faculdade de Farmácia, Universidade do Porto
R. Aníbal Cunha, 164, 4099-030 PORTO, Portugal
[2]ESTG / Instituto Politécnico de Viana do Castelo, Av. Atlântico, s/n, 4901-908 Viana do Castelo, Portugal. mruialves@clix.pt

Multivariate analyses are used to search for the main data structures of data sets, and whenever possible provide the means to display those structures as graphs. However, for inexperienced users of statistics, such outputs are sometimes difficult to interpret.
Predictive biplots[1] can be used to carry out interpretations in relation to initial values and variables instead of latent variables, without loosing the benefits of the multivariate modulation. Once multivariate graphs are produced and printed, the interpolative biplots[1] can be used in routine laboratory practice to position new samples in the graph.
To achieve the objectives of biplots, one biplot axis for each initial variable must be drawn in the graphs. First one decides on convenient scale values and calculate the coordinates of each scale value in the multidimensional subspaces; then, one marker is drawn in the graphs to indicate the position of the scale value; afterwards, all scale values are linked together by a straight line to represent the variable's axis, all markers are labelled with the original scale values, and the axis is also labelled to identify the variable in question.
Biplots can be used with advantage in the solution of many multivariate problems[2] and also in complex situations where different multivariate techniques are coupled together[3].

Fig. 1: Predictive biplot

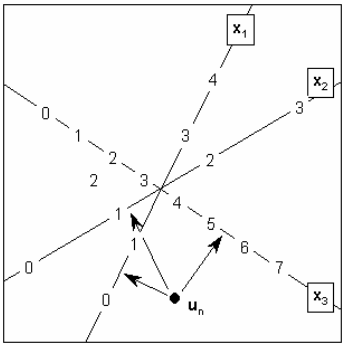


Fig. 2: Interpolative biplot

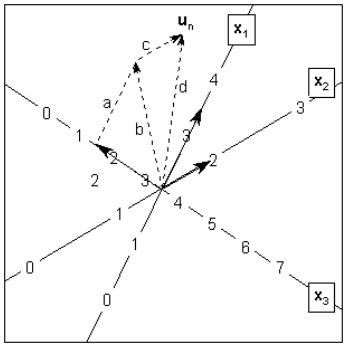


Fig. 1 shows how a predictive biplot is used to read off from the graph the sample $\mathbf{u}_n$ initial values: $\mathbf{x}_1 \approx 0.5$, $\mathbf{x}_2 \approx 1.1$ and $\mathbf{x}_3 \approx 5.4$.

Fig. 2 shows how a given sample $\mathbf{u}_n$ with coordinates $\mathbf{x}_1=3.5$, $\mathbf{x}_2=2.0$ and $\mathbf{x}_3=1.3$ is interpolated in the two-dimensional space by the complete parallelogram's method.

This paper presents complete algorithms in R, adapted from our algorithms initially written in Genstat 5.3.1, to carry out principal components and canonical variate analyses, and produce the markers for predictive and interpolative biplots, exemplified by applications to real laboratory data. The graphical outputs are still being studied in order to obtain interactive ways of adding/deleting biplot axis to achieve the final biplot representations.

[1] Gower, J.C.; Hand, D.J. *Biplots*. Chapman and Hall: London, 1996.
[2] Alves, M.R.; Oliveira, M.B. (2004). Predictive and interpolative biplots applied to canonical variate analysis in the discrimination of vegetable oils by their fatty acid composition. *Journal of Chemometrics*, **18**: 393-401.
[3] Alves, M.R.; Cunha, S.C.; Amaral, J.S.; J.A. Pereira; Oliveira, M.B. (2005). Classification of PDO Olive Oils on the Basis of Their Sterol Composition by Multivariate Analysis. *Analytica Chimica Acta*, **549**: 166–178.

aRT: R-TerraLib API

Pedro Ribeiro de Andrade Neto
Paulo Justiniano Ribeiro Junior
{pedro,paulojus}@est.ufpr.br

Statistical spatial data analysis and Geographical Information Systems (GIS) can act together in order to understand and model spatially distributed data. Geoprocessing operations can equip statistical models with relevant information which on their hand can be used to better understand main features of usually noisy and multidimensional data. Therefore integration between GIS and statistical software can be highly beneficial for both sides.

There are some pieces of work in this direction within the scope of the R project as part of the R-Spatial Task View. We present here the implementation of an R package named aRT to access a GIS library called TerraLib. TerraLib is a set of C++ classes that offers functions and data structures for building customized geographical applications. TerraLib is an open source and free software, and its main objective is to provide a powerful environment for GIS development in a new generation of GIS, once it incorporates space-time support to conventional Database Management Systems (DBMS), for instance MySQL and PostgreSQL.

The package encapsulates C++ classes into S4, therefore the user can manipulate TerraLib objects directly in memory using the implemented wrappers. aRT can manipulate spatial data using the data structures of the sp package, reading and writing *Spatial* data in the database.

Some functionalities already implemented in the package are:

- spatial predicates, such as *touches*, *within*, *contains*, *crosses* and *overlaps*;
- polygons operations, as *union*, *intersection*, *difference* and *simplification*;
- manipulation of temporal tables, and temporal slicing, given a time interval.

aRT is available as source code and also a cross-compiled Windows binary, at `http://www.est.ufpr.br/aRT`.

Using R for teaching statistics to nonmajors: Comparing experiences of two different approaches.

Thomas Baier, Richard Heiberger,
Erich Neuwirth, Wilfried Grossmann

Temple University currently teaches an introductory statistics course based on RCommander and RExcel, and the University of Vienna teaches a course strongly based on a web site using Rpad with interactive examples. The Vienna students additionally use plain R and RExcel in the course labs.

We will show typical examples from the course and discuss the strengths and weaknesses of both approaches. We will also discuss the technical and administrative infrastructure required to teach courses following these two different models.

Implementation of robust methods for locating quantitative trait loci in R

Andreas Baierl and Andreas Futschik

Abstract:

One approach to QTL-mapping is to regress a quantitative trait of interest (e.g. height, yield, tumor count) on the observed genotypes at various positions on the genome (markers) in order to detect locations that influence the trait. We investigate additive and non-additive (epistatic) effects of markers, even if the corresponding main effects are not included in the model. As pointed out by Broman and Speed (2002), the task of selecting the correct markers can be treated as a model selection problem.

Since the distribution of the quantitative trait is often reported to be non-normal, we investigate the application of robust methods (M-estimators, L1-regression) and compare the results to least square regression to estimate the likelihood of the model.

We adapt the modified Bayesian Information Criterion (BIC) proposed by Bogdan et al (2004) that controls the overall type I error of detecting additive effects and pair wise interactions by an additional penalty term. Markers are chosen by an extended forward selection procedure with a backward elimination step.

The performance of the different methods is investigated by an extensive simulation study applying various error-distributions and genetic setups. We have implemented programs carrying out the proposed QTL-mapping approach in R. The robust regression estimators have been calculated using derived the procedures *rlm* from the *MASS*-package and *rq* from the package *quantreg*.

References:

Bogdan, M., J. K. Ghosh and R. W. Doerge, 2004. Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. Genetics, **167**: 989-999.

Broman, K. W. and T. P. Speed, 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. J Roy Stat Soc B, **64**: 641-656.

R in clinical practice - summarizing pharmacological data

P. A. Beitinger*, R. Beitinger†, S. Fulda*, and T. C. Wetter*

28. Februar 2006

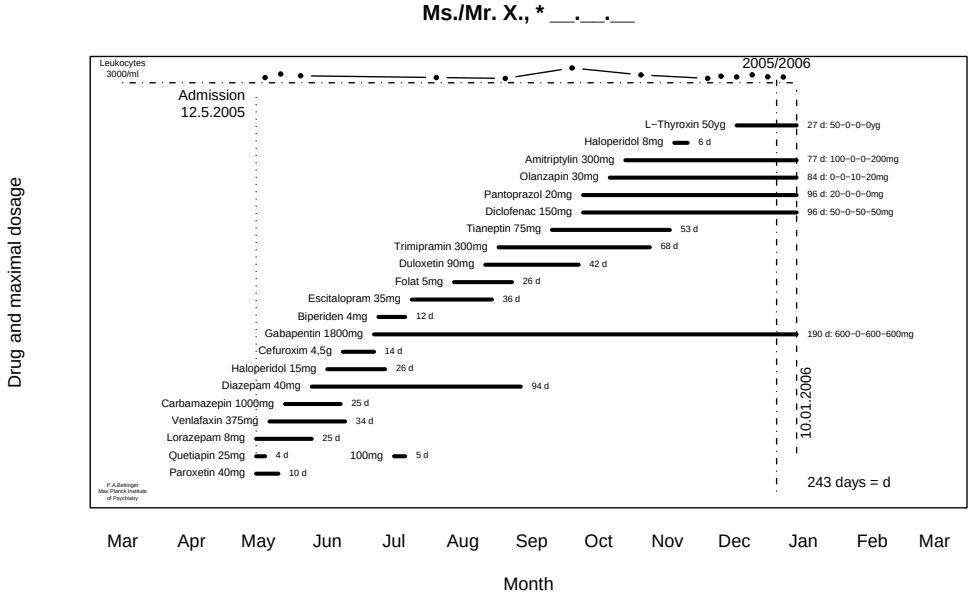
Knowledge of physicians about previous medications is an important basis for making decisions concerning the next therapeutic steps.

In 1994 Powsner and Tufte developed a complete, challenging and well reasoned system for presenting information of patients with about 11000 spreadsheet cells [1, 2]. In contrast to this very comprehensive and elaborate method, we focussed on a clear and effective presentation of relevant key information such as pharmacotherapy and laboratory findings.

Healthcare professional has just to key in each medication including dosage, first and last day of treatment in a concise matrix. With this compact database our R function visualizes the documentation as a timeline graph. This overview not only includes the time information, but also enables the physician to evaluate the past treatment, to display ongoing ineffective medication and develop remaining therapeutic options.

Although the computing and visualisation task could have been done using other programming languages, "R" provided flexible and high end output with minimal and simple coding. Thereby this tool is qualified to increase the therapeutic and time effectiveness in medical long-term treatment.

*Max Planck Institute of Psychiatry, Kraepelinstr. 10, 80804 Munich, Germany
†Department of Psychiatry and Psychotherapy, Technische Universität München, Ismaningerstrasse 22, 81675 Munich, Germany



Literatur

[1] S. M. Powsner and E. R. Tufte. Graphical summary of patient status. *The Lancet*, 344(8919):386–389, 1994.

[2] S. M. Powsner and E. R. Tufte. Summarizing clinical psychiatric data. *Psychiatr Serv*, 48(11):1458–1461, 1997.

Statistical Learning for Analyzing Functional Genomic Data

Axel Benner
Dept. of Biostatistics (C060)
German Cancer Research Center
Im Neuenheimer Feld 280
D-69120 Heidelberg
Email: benner@dkfz.de

An important topic concerning the statistical analysis of functional genomic data is multivariable predictive modelling, where the best prediction of a given outcome variable is sought. Since in microarray studies the number of predictor variables is much larger than the number of observations, standard statistical model building does not work properly. Statistical learning is a new approach to develop prediction models allowing the inclusion of all available data. Selection methods like boosting and regularization methods like penalized regression have been recognized as important statistical learning methods which can control for complexity.

Validation of the fitted models by using independent test samples, bootstrap resampling or cross validation is another important issue. The methods presented above enable for adaptive model selection by tuning their parameters and the set of variables included. Variable selection and choice of parameters is often done by minimization of the cross validated error rates. To estimate the prediction error at least double cross validation is necessary.

We illustrate and compare the different approaches using a data set on predicting survival for patients with acute myeloid leukemia. The results will be compared with respect to the prediction error and interpretability of the results.

Comparison of approaches for fitting generalized additive models

Harald Binder
Institut für Medizinische Biometrie und Medizinische Informatik,
Universitätsklinikum Freiburg, Germany

Generalized additive models are popular when a regression model is to be fitted to a non-normal response where the influence of single covariates is assumed to have unknown non-linear shape. One important difference of the procedures available within the R environment for fitting such models is how covariates and effective degrees of freedom for each covariate are selected. We evaluate the optimization approach of the recommended R package `mgcv`, a stepwise approach, and a mixed model approach. For comparison we offer a new fitting procedure `GAMBoost` – based on boosting techniques – which is built to perform implicit selection of covariates for high-dimensional problems. Its implementation is made available as a new R package. For comparison we focus on simulated data with a small signal-to-noise ratio and/or a large number of covariates. For the underlying true structure simple linear models as well as models incorporating non-linear covariate effects are used. The former allow for comparison of the R packages in situations where the class of models offers too much flexibility while the latter require fitting of complex structure with a limited amount of data. Performance of the procedures is evaluated with respect to prediction performance as well as with respect to the identification of influential covariates. In addition settings are identified where many procedures do not to return any fit at all and only `GAMBoost` provides a viable alternative.

Using Grid Graphics to produce linked micromap plots of large financial datasets

Gordon Blunt
CACI Ltd
gblunt@caci.co.uk

This paper describes how linked micromap (LM) plots, drawn using the flexibility of Grid Graphics' multiple coordinate systems, can be used to develop graphical summaries of large financial services datasets. A typical LM plot can contain sixty, seventy or more individual plot elements, which need to be arranged carefully on a page, with particular regard to alignment of the elements. The Grid Graphics package allows precise control over layouts, and so the exact placement of each plot element within the layout. Further, use of the different coordinate systems in Grid allows the flexible specification of the different plot elements – micromaps, labels, statistical graphics and text – with ease.

This flexibility comes at a price, however, in that Grid Graphics provide only low level graphical functions rather than high level functions that can produce complete plots. However, once these low level functions are understood, complex plots can be constructed with with relatively little effort. This is well illustrated by the Lattice package, which uses Grid to render its plots.

LM plotting can be thought of as a method that combines the techniques of exploratory data analysis and statistical graphics, but which maintains the spatial context of the data. Many financial services datasets have pronounced geographic differences for which traditional statistical summaries – in isolation – are inadequate; on the other hand, choropleth maps on their own cannot convey the range of data to be presented. Linked micromaps are one solution to this problem, and have the following characteristics.

- They display several sequences of panels that are linked by position. These panels contain names, maps and graphical summaries of the data. The latter can be any form of statistical summary; for example histograms, time series or dot plots, among many others.
- Data are sorted by the variable(s) of interest, to improve perception between sequences of panels.
- The dataset is partitioned into these relatively small panels to allow attention to be focused on small areas at a time.
- They draw on principles from a number of disparate disciplines, including statistics (in particular exploratory data analysis), cartography and psychology to produce convenient, yet revealing, summaries of large data sets which preserve most of the important elements of the data.

Another advantage of LM plots is that they provide insights into the structure of a dataset at many levels. This is a consequence of the range of statistical graphics it is possible to use, which are able show extremes as well as measures of central tendency. LM plots are flexible enough to be used for this further analysis, and can thus be thought of as a flexible graphical data mining tool. Thus, they allow data sets with millions of observations to be summarised in a convenient – and revealing – graphical way.

LM plots have been used mainly in the fields of epidemiology, ecology and official statistics but have not, as yet, been applied to financial datasets.

The ritools package: Tools for Exact and Randomization Inference

Jake Bowers and Ben Hansen

Tests based on randomization (or permutation) are very attractive for the conceptual simplicity of their foundations and for the paucity of assumptions they require of the data analyst. However, such tests have not had widespread acceptance among political scientists, sociologists, and economists, despite their long heritage in statistics. One impediment to wide use of such tests has been that they are seen as computationally burdensome. This problem is disappearing at nearly the rate of Moore's law. Other reasons for avoiding these tests have arisen from essentially their user interface: compared to the unified framework of the generalized linear model, the tests of Fisher; Cochran, Mantel, and Haenszel; McNemar; and Wilcox seem confusingly unconnected (and unruly). Adaptations of the linear model seem more conceptually elegant than choosing from a grab bag of named techniques. Finally, another reasons for the lack of engagement between social scientists and randomization inference has been the scarcity of datasets generated using random assignment in the research design, which is the design that provides the firmest foundation to randomization inference techniques. Recently, however, more and more random assignment is being done in the social sciences, and new developments in the analysis of observational studies are allowing randomization inference to compete with other modes of testing and estimation.

Our presentation will discuss the design and use of an R package containing Randomization Inference Tools ("ritools"), which aims to overcome a few of the problems mentioned above — namely the user interface problem, and some of the computational complexity problems. We will describe the package and some of the challenges that we've overcome in its creation. And we will show it in operation in two modes: (1) testing for balance during the application of optimal, full matching [using the "optmatch" package], and (2) estimating treatment effects for binary responses using Rosenbaum's "attributable effects" framework (Rosenbaum 1991).

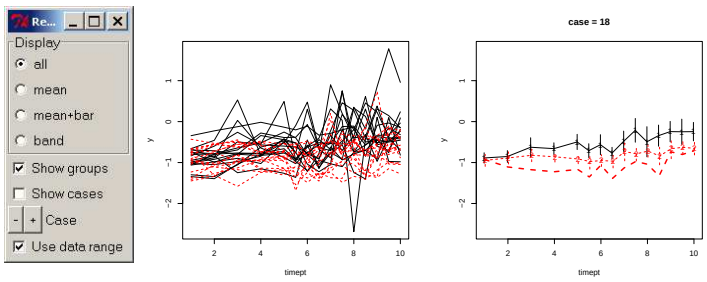
An abstract submitted to the useR! 2006 conference

rpanel: simple interactive controls for R functions using the tcltk package

Adrian Bowman & Ewan Crawford
Dept. of Statistics, The University of Glasgow, Glasgow G12 8QQ

In a variety of settings it is extremely helpful to be able to apply R functions through buttons, sliders and other types of graphical control. This is particularly true in plotting activities where immediate communication between such controls and a graphical display allows the user to interact with a plot in a very effective manner. The tcltk package described by Dalgaard (2001) provides extensive tools for this and the aim of the rpanel package is to provide simple and well documented functions which make these facilities as accessible as possible, particularly for those who have reasonable familiarity with R but are less confident in embarking on a general exploration of the *Tcl/Tk* system. In addition, the global variables which are the means of communication in tcltk are also managed by rpanel so that the user need not be aware of them. A standard, flexible form of parameter passing is also used to communicate with the plotting functions written by the user. Among other advantages, this allows users to call a particular function repeatedly, to produce simultaneous and independent copies of the same graphical method without causing control difficulties.

The basic design of the software will be described and illustrated on a variety of examples of interactive control of graphics. Part of the aim of the presentation is also to outline the range of uses to which these facilities can be put. This includes situations in data analysis where interactive control gives important and convenient insight, such as in dynamic graphics. The figure below shows an example with repeated measurements data where a panel allows easy movement between plots of the raw data, means and standard errors or individual profiles. This provides a very helpful means of identifying unusual profile shapes in a manner which is not possible by plotting all the data, due to the difficulty in identifying individual cases. A further example involving the animated display of three-dimensional shape data, employing the facilities of the rgl package (Adler, 2005), will also be described. Other examples will be drawn from teaching, where animation in particular can communicate some concepts in a much more effective manner than static plots. Interaction with more general images is also a potentially rich facility.



The current version of the rpanel package, and a full description of its aims (Bowman, Crawford, Alexander & Bowman), are available at

www.stats.gla.ac.uk/~adrian/rpanel.

Applied Asymptotics in R

Alessandra R. Brazzale

Institute of Biomedical Engineering
Italian National Research Council

The likelihood function represents the basic ingredient of many commonly used statistical methods for estimation, testing and the calculation of confidence intervals. In practice, much application of likelihood inference relies on first order asymptotic results such as the normal approximation to the distribution of the standardized maximum likelihood estimator. The approximations can, however, be rather poor if the sample size is small or, generally, when the average information available per parameter is limited. Thanks to the great progress made over the past twenty-five years or so in the theory of likelihood inference, very accurate approximations to the distribution of statistics such as the likelihood root have been developed. These not only provide modifications to well-established approaches, which result in more accurate inferences, but also give insight on when to rely upon first order methods. We refer to these developments as *higher order asymptotics*.

The purpose of this presentation is to show the importance and range of application of higher order asymptotics in statistical practice. We will do this by presenting a selection of examples. These range from elementary one-parameter models, chosen to illustrate the potential accuracy of the procedures, to more specialized examples. All calculations are carried out using R. One intriguing feature of the theory of higher order likelihood asymptotics is that relatively simple and familiar quantities play an essential role. Furthermore, many classes of models can be treated in a modular way, so higher order quantities can be expressed using a few elementary building-blocks. These are the key to efficient numerical implementation of higher order asymptotics. An example is the `hoa` package bundle.

`alessandra.brazzale@isib.cnr.it` `www.isib.cnr.it/~brazzale/`
`www.isib.cnr.it/~brazzale/CS/`

A fixed effects approach to GLMs with clustered data

Göran Broström

Department of Statistics
Umeå University
SE–901 87 Umeå, Sweden

1 Introduction

In situations where a large data set is partitioned into many relatively small groups, and you want to test for group differences, the number of parameters tend to increase with sample size. This fact causes the standard assumptions underlying asymptotic results to be violated. There are (at least) two possible solutions to the problem, first, a random intercepts model, and second, a fixed effects model, where asymptotics are replaced by a simple form of bootstrapping.

In the `glmML` package, both these approaches are implemented. In this paper, only the fixed effects approach is considered.

2 The fixed effects model

In the fixed effects model, testing is performed via a simple bootstrap. Under the null hypothesis of no grouping effect, the grouping factor can be randomly permuted without changing the probability distribution. This is one basic idea in the estimation of the p-value by simulation. The direct parametric aproach is to draw bootstrap samples from the estimated probability distribution.

We first show how to write down the log-likelihood function and all the first and second partial derivatives. Then we introduce the profiling approach which reduces an optimizing problem in high dimensions to a problem consisting of numerically solving several one-variable equations and optimization in low dimensions. The profiling cannot be done explicitely, but it is possible via *inplicit differentiation*.

The procedure is implemented in `glmmML` for the *Binomial* and *Poisson* families of distributions. Some comparisons with alternative approaches are made by simulation.

oligoExpress – exploiting probe level information in Affymetrix GeneChip expression data

Jan Budczies, Joachim Grün

Oligene GmbH, Schumannstr.20/21, Campus Charité Berlin-Mitte, 10117 Berlin, Web: www.oligene.de, Email: budczies@oligene.de

We present oligoExpress, a database system for management and analysis of Affymetrix gene expression data. Often, the evaluation of Affymetrix expression data starts with summary of the probe level measurements to a matrix of expression values (e.g. GCOS signals) that is used as input for all further analyses. However, a lot of other useful information can be extracted from the probe levels measurements. Examples are detection p-values, signal log ratios (SLRs), and change p-values, all introduced by the chip manufacturer (cf. Affymetrix, 2002, Statistical Algorithms Description Document). As an exhaustive exploitation of probe level information allows conducting some statistics on the chip, it can be helpful to keep the number of expensive external replications small. Three different kinds of summaries can be generated from the probe level data: measurements of a single chip, results of pairwise chip comparison, and results of comparisons between biological sample groups. We have developed an `R` application that copes with the enormous amount and the different kinds of the summary data and collects them in a database.

Our software integrates data processing and annotation in an automated workflow that uses the raw data files (cel-files) and a sample annotation file as input. The sample annotation file is an Excel table with names of the cel-files as row names and an arbitrary number of attribute columns. Each attribute corresponds to a (biological) property of the samples and has a value or is “not applicable” for each of the samples. Data processing starts with absolute analysis (signals, detection calls) of each of the chips and comparative analysis (SLRs, change calls) of each possible pair of chips. For absolute analyses the implementation of Affymetrix algorithms in the package `affy` is employed. For comparative analysis we have implemented the corresponding algorithms ourselves, as they have not been available under `R` up to now. As final step of data processing, analysis of biological groups, as they are stored in the annotation file, is performed. For probe set and gene annotation we make use of the package `annaffy`. All data are stored in a relational database that is defined and accessed via the RODBC interface. Microsoft Access is employed as test target database in our first applications, but usage of all other databases with ODBC interface (e.g. MySQL, Oracle) is straightforward.

One of the most common goals of DNA microarray experiments is the detection of differentially expressed genes between two states or types of cells, for example cells from healthy and diseased tissues. As an application of the oligoExpress database system, we have evaluated different procedures for the detection of differential gene expression. The Latin Square data set was downloaded from the Affymetrix homepage and prepared as oligoExpress database. The Latin Square data consist of 3 technical replicates of 14 hybridizations of 42 spiked transcripts in a complex human background. Different procedures including summary statistics for signals, detection calls, signal log ratios and change calls were checked for their performance. Results, recorded in terms of sensitivity and specificity, demonstrated the power of summary statistics based on signal log ratios for the detection of differential transcripts.

Using R to Evaluate Trading Strategies

Patrick Burns

26th February 2006

Abstract

R is arguably the best environment in which to evaluate trading strategies. The strong programming language and rich data structures make this difficult task easier. A particular trading strategy is explored and the R code used for the analysis is outlined.

Studies on financial time series analysis

L. J. Carbonaro
University of Oxford

This talk will examine two of the most important volatility models in finance: ARCH and Stochastic Volatility models. Then, it will describe some aspects of neural networks and the concepts of penalised- and quasi-likelihood methods. In respect of these considerations, a novel model will be introduced together with a new estimation procedure for financial time series analysis based on a penalised quasi likelihood (PQL). Results will be provided by using foreign exchange market data and simulation methods. The talk concludes with the measuring of market risk in accordance to the recommendation of the Basel Accords. All calculations are implemented in R.

KEYWORDS: Time-Varying Volatility, Stochastic Volatility, Recurrent Neural Networks, Penalised Quasi-Likelihood, Foreign Exchange Market, Basel Accords.

A History of S and R

(with some questions for the future)

John M. Chambers

June 15, 2006

Just thirty years ago, a group of statistics researchers at Bell Telephone Laboratories (as it was then) met to plan a software system suited to their own needs. The descendants of that system (in particular the open-source R software) are now the dominant software for implementing and communicating new statistical methods. This talk reviews some of the main events that got us from there to here. We will see some important continuities, as well as equally important coincidences and unintended consequences.

We will also consider where "the peaceful collision of statistics and computing" (John Tukey) might want to go from here, and what innovations may be needed to get there?

GEAR: GNU Econometric Analysis with R*

Christine Choirat, Paolo Paruolo, Raffaello Seri†

The GEAR project aims at providing a free, advanced and extensible set of standardized R packages for econometric analysis that can be used as: (i) a GUI (*i.e.* Graphical User Interface) program (when performing standard tasks) in the spirit of EViews (see `http://www.eviews.com`) and GiveWin/PcGive (see `http://www.oxmetrics.com`); (ii) a set of libraries oriented towards econometrics (for more advanced analysis) in the spirit of the routines availables in Gauss (see `http://www.aptech.com`) and Ox (see `http://www.doornik.com`).

GEAR is entirely written in R (except for computer-intensive tasks which are coded as DLL's). Obviously, R can already be used for econometrics (for a review, see Cribari-Neto and Zarkos, 1999, Racine and Hyndman, 2002, Farnsworth, 2006 and A. Zeileis' CRAN task view for computational econometrics). However, only a partial list of methods has already been implemented (moreover by independent authors so that many econometric methods are lacking whilst others are redundant). The GUI of GEAR is implemented using the `tcltk` package of P. Dalgaard and some well-known Tk extensions (especially BWidgets and TkTable) that can be found in the ActiveTcl bundle and are also available separately. Even if Tk is not the most modern and pleasant-looking available GUI tool, it allows GEAR to be a really cross-platform application that requires very little configuration on the part of the user. GEAR has been tested on several MS Windows and Linux/Unix versions and Mac OS X (both with X11 and Aqua).

GEAR is organized in a modular way. Each module (which in practice takes the form of an R package) is meant to correspond to a particular class of econometric models (*e.g.* linear regression, univariate time series, VAR, panel data, etc.) and is constructed around the same steps (*i.e.* model definition, model estimation, display of the output, diagnostic tests and graphics). This helps the user find his way through the model-elaboration strategy.

The main package `gear-main` has already been implemented. It features a tabbed output window, a calculator and a spreadsheet. The cross-section regression package `gear-crossreg` has already been written too. It features estimation, tests and graphical output. More packages (in particular advanced cointegration analysis) are under active development.

*Further information and screenshots are available from the authors.

†Università degli Studi dell'Insubria, Dipartimento di Economia, via Monte Generoso 71, 21100 Varese, Italy. E-mails: {cchoirat, pparuolo, rseri}@eco.uninsubria.it. Homepages: http://www.eco.uninsubria.it/webdocenti/{cchoirat, pparuolo, rseri}.

Computing Weighted χ^2 Distributions and Related Quantities

Christine Choirat, Raffaello Seri*

It is well known that the asymptotic distribution of degenerate $U-$ and $V-$statistics is, in general, an (infinite) weighted sum of χ^2 random variables. The behavior of the statistic in terms of asymptotic distribution and power is strictly linked to the eigenvalues and the eigenfunctions of an integral operator. We provide an algorithm for the numerical approximation of these quantities, and of the cdf of a weighted sum of χ^2 random variables. The algorithm can be used to approximate (as precisely as needed) the power of the test statistics, and to build several measures of performance for tests based on $U-$ and $V-$statistics. The algorithm uses the Wielandt-Nyström method for the approximation of the solution of integral operators.

The algorithm previously exposed has been implemented in an R package. The computation of the eigenvalues and of the eigenfunctions can be performed using a Monte-Carlo method, a quasi-Monte Carlo method based on the Halton sequence or on the Hammersley point set, the trapezium rule, the Gauss-Legendre quadrature rule and the Clenshaw-Curtis quadrature rule. The approximation of the cdf uses a routine recently written by Robert B. Davies in C (indeed, a new version, available from the Internet at `http://www.robertnz.net/ftp/qf.tar.gz` or `qf.zip`, of the 1980 program that was originally written in Algol). The performance of the method can be analyzed through a Berry-Esséen bound. On the basis of extensive experimentation, we advocate the use of the Wielandt-Nyström method based on the Clenshaw-Curtis quadrature rule.

*Università degli Studi dell'Insubria, Dipartimento di Economia, via Monte Generoso 71, 21100 Varese, Italy. E-mails: `{cchoirat, rseri}@eco.uninsubria.it`. Homepages: `http://eco.uninsubria.it/webdocenti/{cchoirat, rseri}`.

Missing Data, PLS and Bootstrap: A Magical Recipe?

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

Alexandra Machás
ESCS/IPL, Departamento de Economia
amachas@netcabo.pt

M. Manuela Neves
ISA/UTL, Departamento de Matemática
manela@isa.utl.pt

February 10, 2006

Abstract

The problem of missing data or incomplete data is frequently found in many data bases. The amounts of missing data create difficulty in statistical analysis because the techniques used are not designed for them. Therefore, missing data reduces statistical power because the statistical methods presume that the data bases has information on all variables.

In order to solve the problem of missing data we use some missing data techniques or data imputation algorithms for reconstructing the incomplete data to a complete data set. These algorithms fill out the missing data values, by examining the range of probable values for each variable and calculates many future values randomly. So, using these methods we end up with a credible data set and the results often produce more accurate estimates.

Ensuring the good quality of data, methods as Structural Equation Models (SEM) and Customer Satisfaction Models (CSM) can be considered a true strategic instrument for the organizations and the base for the definition of action marketing planning.

The main objective of this work is to bring up in discussion a problem that could affect the quality of estimators and the validation of the models- the missing data. This issue will be applied to a CSM using data from a market survey conducted for the mobile telecommunication sector in Portugal. Using these results we intend to obtain estimates for the missing data and also to achieve better results combining this procedure with the PLS, reducing the biases of estimators. An extensive computer work is perform and a large number of estimates are calculated using R Software and their packages. Overall, it was concluded that for a higher non-response rates(50%) bootstrap is the best method to be adopted in case of missing data completely at random.

Author Keywords: PLS; Bootstrap; Missing Data; Missing Data methods;

plm : linear models for panel data

Yves Croissant
Laboratoire d'Economie des Transports
Institut des Sciences de l'Homme
14, avenue Berthelot
F-69363 Lyon cedex 07
yves.croissant@let.ish-lyon.cnrs.fr
33 4.72.72.64.49

`plm` is a package that implements the main estimators and tests used in econometrics for panel data.

Panel datas have an individual and a temporal dimension. `plm` provides specific functions for reading and applying special transformations to panel data sets, and for estimating and testing linear models.

reading data `pdata.frame` takes as main argument a `data.frame`. It returns a `data.frame` with further arguments usefull for panel datas, such as the number of individuals and time observations.

special functions this includes `plag` and `pdiff`, which computes lags and differences of series, `pmean` which computes the mean of a serie, conditionnal on the individual or the time index,

estimation `plm` is a general function which implements the main panel data estimators. The basic usage of `plm` consist of estimating four models :

- **pooling** the ordinary least squares estimator applied to raw observations,
- **within** the ols estimator applied to observations measured as deviations from individual (or time) means,
- **between** the ols estimator applied on individual (or time) means,
- **random** the random effect, a generalized least squares estimator which is a wheighted average of the `within` and the `between` estimator.

`plm` returns by default an object of class `plms`, which is a list of the four models previously described, which are objects of class `plm`. `plms` and `plm` objects have `print` and `summary` methods. These estimators deals with oneway (individual or time) effects or twoways and with unbalanced panel. Different instrumental variable estimators are also available (for example the Hausman and `Taylor` estimator)

tests different tests of model specification are provided :

- **pFtest** a simple test for the presence of individual (or/and time) effects based on the comparison of the `pooling` and the `within` models,
- **plmtest** a set of likelihood ratio tests for the presence of individual (or/and time) effects based on the comparison of the `random` and the `pooling` model,
- **phausman** a hausman test for the correlation between explanatory variables and individual (or/and time) effects, based on the comparison of the `within` and the `random` models.

Further developments planed for `plm` include :

system estimation seemingly unrelated regression and three stage least squares estimators, using the `systemfit` package,

robust covariance matrix using the `sandwich` package,

autoregressive models Arrelano and Bound general method of moments estimator.

All the functions of `plm` have been tested using the examples provided in the book of B. Baltagi "Econometric analysis of panel data". The data sets used in this book are provided in packages `Ecdat` and `plm`.

Repeated measures tools for multivariate linear models

Peter Dalgaard

A set of methods which extend preexisting methods for objects of class "mlm" was introduced in R versions 2.1.0-2.2.0. These methods deal with linear models with multivariate response.

The new methods allow model reduction tests based on multivariate normal theory. However, multivariate models are often employed as a first step in the analysis of repeated measurements data. In such data, the individual coordinates of the response measure fundamentally the same thing, but one could be unsure of the correlation structure.

In a repeated measurements context, the concepts of sphericity and the Greenhouse-Geisser and Huynh-Feldt adjustments to standard F tests become important. A further aspect is that you generally need to work with transformed response vectors, e.g. contrasts or averages within subjects, and there is a need for a structured specification of such transformations.

Spatial and statistical modelling of phenological data using 'R'

Daniel Doktor*
Imperial College

This paper analyses the spatio-temporal patterns of Land Surface Phenology (LSP) in Germany. LSP is the study of the spatio-temporal development of vegetated land surface as revealed by synoptic sensors, whether space borne remote sensors or in situ observational networks. LSP provides a critical window on the local consequences of global change.

The phenological data used for this study come from the phenological network of the German Weather Service and is managed by the relational database system 'Oracle'. The packages **'DBI'** and **'ROracle'** (64 bit application) were utilized for communicating and interfacing between 'R' and the database. The package **'gstat'** was applied for geostatistical modelling purposes such as Variogram estimation and Kriging. Spatial interpolation of phenological ground observations was carried out either using Detrended Kriging (referring to average elevation gradients) or applying External Drift Kriging. Both interpolation methods performed on a similar level with estimation errors between 3-9 days (using cross-validation). The results (d_{BB}) were visualised using the packages **'lattice'** and **'maptools'**.

The interpolated budburst dates from ground observations were compared to satellite derived dates of green-up (d_{GU}). Mean, modus and median of the frequency distribution of $d_{BB} - d_{GU}$ indicate that satellite derived green-up preceded observed tree budburst dates on average by 3 days.

In order to quantify the influence of cold spells on the temporal evolution (pace) of spring time phenology we used Gaussian Mixture Modelling. Using this methodology it was possible to quantitatively characterise the frequency distributions of observed budburst dates. Mixture components could be identified either via Expectation-Maximisation (EM) or via an optimisation algorithm. The EM algorithm was initialised by hierarchical clustering (package **'mclust'**) for parameterised Gaussian Mixture Models. The number of clusters and the clustering model is chosen to maximise the Bayesian Information Criterion (BIC). Secondly, an optimisation algorithm (**'base'**: optim) was applied on the minimisation of several (maximum four) Gaussian Mixture Functions. Based on Akaike's Information criterion it was decided which optimised function was the most appropriate. The identified mixture components also formed the methodological base for a new outlier detection algorithm to be applied within huge phenological databases.

Additionally, space-time correlations of the phenological ground observations were analysed. Every phenological station was compared to all others provided the respective station pair had at least 20 identical years of observations. Dependent on the geographical distance of the pair correlation coefficients were assigned to certain distance categories for each station. Consequently, it was possible to determine how phenological time series correlate over space and which stations showed noticeably different trends when compared to the Grand Mean.

It is the authors intention to include phenologically related functionalities into the 'R'-package **pheno**.

*Correspondence: Daniel Doktor
d.doktor@imperial.ac.uk
Prince Consort Road, RSM Building
SW7 2BP London

Integrating R in an Advanced Building Control System

David Lindelöf

We use the R environment to perform core calculations in an automatic building control system. The system controls electrical lighting and venetian blinds in one- or two-person offices and relies on a Bayesian algorithm to evaluate the users' visual comfort. At the centre of this algorithm lies the estimation, after every recorded user action, of the density of certain physical variables before and after the action. These variables are those that are thought to correlate with the subjective impression of visual comfort, and that are relatively easy to model, such as the illuminances at different points in the field of view.

Here we describe the implementation of this algorithm in a high-level language and how that language uses calls to R for the most mathematically difficult steps.

Some experiments on Statistical Matching in the R environment

Marcello D'Orazio, Marco Di Zio, Mauro Scanu
(madorazi@istat.it, dizio@istat.it, scanu@istat.it)

ISTAT, Via Cesare Balbo 16, 00184 Roma, Italy

Key Words: Data Analysis, Data Processing.

In the last years, interest on Statistical Matching problems has increased (Rässler, 2002, D'Orazio *et al.* 2006). This is due to the large amount of data-sets available and, at the same time, to the need of timely and not costly information. Statistical Matching techniques aim at combining information from different sources. In particular, it is assumed that the two sources (e.g. two samples) do not observe the same set of units, so that neither merging nor record linkage techniques can be applied.

In order to explore the properties of matching techniques and therefore apply them to real data problems, a series of matching experiments have been carried out in the R environment. R has already been used for the definition of some statistical matching algorithms (Rässler 2002). In D'Orazio *et al* (2006) more algorithms have been translated in R. The codes will be available on the web page `http://www.wiley.com/go/matching`.

At first an extensive simulation study has been carried out taking into account two separate cases: (i) all the variables are continuous and (ii) all the variables are categorical. In the presence of continuous variables most of the experiments aim at the evaluation of the performances of matching techniques based on regression methods. As far as categorical variables are considered, we worked in the direction of exploring the uncertainty of the results typical of a statistical matching application. We show how the usage of some basic auxiliary information, in the form of logical constraints involving values of different variables, can reduce the uncertainty. An application of statistical matching to real data is also presented. We tried to estimate the Social Accounting Matrix by means of the fusion of the Household Balance Survey conducted by the Bank of Italy and the Household Expenditure Survey conducted by the Italian National Statistical Institute. A Social Accounting Matrix is a system of statistical information containing economic and social variables in a matrix formatted data framework. The matrix includes economic indicators such as per capita income and economic growth.

References

D'Orazio M, Di Zio M, Scanu M (2006) *Statistical Matching: theory and practice*, Wiley.

Rässler S (2002). *Statistical Matching: a frequentist theory, practical applications and alternative Bayesian applications*, Springer.

R in Psychometrics and Psychometrics in R

Jan de Leeuw
UCLA Statistics

In psychometrics, and in the closely related fields of quantitative methods for the social and educational sciences, R is not yet used very often. Traditional mainframe packages such as SAS and SPSS are still dominant at the user-level, Stata has made inroads at the teaching level, and Matlab is quite prominent at the research level.

In this paper we define the most visible techniques in the psychometrics area, we give an overview of what is available in R, and we discuss what is missing. We then outline a strategy and a project to fill in the gaps. The outcome will hopefully be a more prominent position of R in the social and behavioral sciences, and as a result less of a gap between these disciplines and mainstream statistics.

The use of R as part of a large-scale information management and decision system

Joris DE WOLF, Koen BRUYNSEELS, Rindert PEERBOLTE and Willem BROEKAERT
CropDesign

February 28, 2006

This presentation describes the successful use of R as a part of an information system in a industrial setting.

An information management and decision system has been developed for TraitMill™, a highly automated plant evaluation platform allowing high-throughput testing of the effect of the introduction of transgenes on agronomically valuable traits in crop plants. The screening is based on plants grown in-greenhouse in specific experimental layouts, imaged at weekly intervals, and harvested on an individual plant basis. About one hundred thousand plants are screened annually. The measurements are automated to a large extent and a vast amount of data is stored directly in a central relational database. This database is used to manage the information flows but also gives input to a decision support system that assists in detecting interesting genes in the test population.

The relational database is built in Microsoft SQL Server and accessed through a Java fat client or a web based front-end, all running on a Linux platform. Additional components were needed to carry out formal statistical analysis and inference as well as for producing graphs. These components had the following requirements: (i) be highly versatile and programmable in-house, (ii) be able to perform complex statistical analysis (linear mixed models, survival analysis, non-linear curve fitting among others) in a reliable and automated manner, and (iv) exchange data and results with the database swiftly and reliably. The graphic component had to (v) produce graphs for integration in websites and for high-quality publication.

R has been chosen to perform both the analytical and graphical tasks. It fulfilled all requirements mentioned above. For the core of the statistical analysis, off-the-shelf R packages and functions proved to be sufficient and performed these tasks swiftly enough. The accessibility of the code and the relative simplicity of the language made the development of scripts for specific goals straightforward. The only low-end adaptation that needed to be developed was the connection between R and MS SQL Server. The latter functionality has been put in the open source domain by CropDesign.

Currently R is used in two ways in this system. First, batches of scripts are run unsupervised in the background, using data from the database and storing results back into the database and graphical output into file systems. Second, R is accessed interactively by the user via the Java interface. For the latter Rserve is used.

A drawback of R is the high turnover of new releases and the problems or suspicion of backward incompatibility that this may bring along.

Despite this downside, R has proven in the last three years of high-througput operation of TraitMill™ to be a valuable ressource for information management.

Asterias: an example of using R in a web-based bioinformatics suite of tools

Ramón Díaz-Uriarte[1,2], Andrés Cañada[1], Edward R. Morrissey[1], Oscar Rueda[1], Andreu Alibés[1], David Casado[1]

[1]Bioinformatics Unit, Spanish National Cancer Center (CNIO), Melchor Fernández Almagro 3, Madrid, 28029, Spain

Abstract

Asterias (`http://www.asterias.info`) is an integrated collection of freely-accessible web tools for the analysis of microarray gene expression and aCGH data. Six of the applications use R (and many R packages) for the computations and graphics, and four of those rely heavily on Rmpi and/or snow for parallelization (Asterias runs on a computing cluster with 60 CPUs). R has shown, once again, that it is an ideal system to "turn ideas into software, quickly and faithfully" (Chambers, 1998, *"Programming with data"*), but setting up and maintaining the system up and running has presented several challenges. In this talk we will discuss some of the features of our set-up, including load-balancing and high-availability, the combination of R and Python for the web-based applications, checking the status of MPI and launching MPI-dependent applications in a web-serving context, and automated testing of web-based applications.

[2]Corresponding author. Email: `rdiaz@ligarto.org`

Regression rank-scores tests in R

Jan Dienstbier[(1)] and Jan Picek[(2)]
(1) Department of Statistics, Charles University, Prague
(2) Technical University of Liberec, Czech Republic

R. Koenker a G. Basset (1978) proposed the regression quantiles as any generalization of usual quantiles to linear regression model. They characterized the regression quantile as the solution of the linear program. Gutenbrunner and Jurečková (1992) called the componenents of the optimal solution of dual problem as the regression rank scores. They showed that many aspects of the duality of order statistics and ranks in the location model generalize naturally to the linear model.

Gutenbrunner and Jurečková (1992) proposed some tests based on regression rank scores generated by truncated score functions. A general class of tests based on regression rank scores, parallel to classical rank tests as the Wilcoxon, normal scores and median, was constructed in Gutenbrunner et al. (1993). The tests of the Kolmogorov-Smirnov type were proposed by Jurečková (1992) and the tests of homoscedasticity in the linear model based regression rank scores were proposed by Gutenbrunner (1994). The concept of regression rank scores was extended in Koul and Saleh (1995), as autoregression rank scores, to the autoregressive (AR) model. The tests of the linear hypothesis on the AR parameter based on the autoregression rank scores were constructed in Hallin and Jurečková (1997) and in Hallin and el. (1997). A nonparametric test of independence of two autoregressive time series was considered Hallin and el. (2001). Goodness-of-fit tests in the model with nuisance regression and scale parameters were constructed by Jurečková, Picek and Sen (2003).

The purpose of this presentation is to show the implementation of above mentioned tests in R.

References

[1] Gutenbrunner, C. and J. Jurečková (1992). Regression rank scores and regression quantiles. *Ann. Statist.* **20**, 305-330.

[2] Gutenbrunner, C., J. Jurečková, R. Koenker and S. Portnoy (1993). Tests of linear hypotheses based on regression rank scores. *J. Nonpar. Statist.* **2**, 307-331.

[3] Hallin, M. and J. Jurečková (1997). Optimal tests for autoregressive models based on regression rank scores.

[4] Hallin, M., T. Zahaf, J. Jurečková, J. Kalvová and J. Picek (1997). Non-parametric tests in AR models, with applications to climatic data. *Environmetrics* **8**, 651-660.

[5] Hallin, M., Jurekov, J., Picek, J. and Zahaf, T. (1999) Nonparametric tests of independence of two autoregressive time series based on autoregression rank scores. J. Statist. Planning Inference 75, 319-330.

[6] J. Jurečková, J. Picek and P.K. Sen (2003). Goodness-of-fit test with nuisance regression and scale. *Metrika* 58, 235–258.

[7] Koul, H.L. and A.K.Md.E. Saleh (1995). Autoregression quantiles and related rank scores processes. *Ann. Statist.* **23**, 670-689.

A Package for Inference about Ratios of Normal Means

Gemechis Dilba, Frank Schaarschmidt, Ludwig A. Hothorn

Faculty of Natural Sciences, Teaching Unit Biostatistics
University of Hannover, Germany

Abstract

Inferences concerning ratios of means of normally distributed random variables arise in a variety of problems in the biological sciences. For example, in tests for non-inferiority of two or more treatments, it is often easier to define and also to interpret the non-inferiority margins as percentage changes. This talk aims to introduce an R package called *MRatio* which can perform inferences about one or more ratio parameters. For two-sample problems, the package is capable of constructing Fieller confidence interval and perform Sasabuchi test when the variances are homogeneous or heterogeneous. The package can also construct simultaneous confidence intervals for multiple ratios, performs multiple tests, and calculates the sample sizes required in many-to-one comparisons based on ratios. The functionality of the package will be demonstrated using real data examples.

Enterprise Automatons with R

Zubin Dowlaty, Vice President Decision Sciences, InterContinental Hotels Group
Dean Mao, Advanced Computing Analyst, InterContinental Hotels Group
Simon Urbanek, AT&T Labs - Research

Abstract

Modern competitive enterprises today need software that enables the creation, persistence and scheduling of analytical processes or workflows. As we store data in ubiquitous databases, the equivalent persistence and query mechanism is needed for processes. Further, to enable analytics to become more accessible within the enterprise, the capability for easy integration with existing heterogeneous systems, extensible architecture and a visual modeling metaphor will increase the likelihood for adoption and success. As we develop and mature along the analytics path, automation of existing processes will likely become an important objective, enabling the analyst to continue to innovate on new and novel applications, rather than serve the current and past processes that are needed to be maintained.

Presently within the software industry their exists a genre with the label of Business Process Modeling or BPM, which is an attempt to create a generic framework for modeling workflow processes. Best of breed commercial vendors like Tibco and Webmethods are playing in this space. Further, SAS with Enterprise Miner, SPSS with the Clementine product, and S with Insightful Miner have also introduced a process driven approach focusing more on statistical applications. When we scan the open source landscape, there presently did not exist an enterprise capable analytics solution that leverages these BPM concepts. R as a statistical language is extremely robust, we feel R coupled with an enterprise quality open source BPM engine and visual client that can be used to model, persistent and schedule analytics workflows, this combination would elevate R into many new and unique applications within the enterprise and beyond.

The purpose of our talk today is to discuss our solution, demonstrate examples utilizing the actual software we have developed and release the codebase to the open source community.

Subselect 0.99: Selecting variable subsets in multivariate linear models

A. Pedro Duarte Silva and Jorge Cadima and Manuel Minhoto and Jorge Orestes Cerdeira

The subselect package combines, in a unified framework, a set of search routines that look for k-variable subsets that are good surrogates for a full p-variable data set. In version 0.8, presented at User!2004, no assumptions were made about the intended use of the data, and the criteria implemented measured the quality of the surrogates through different functions of the original covariance or correlation matrices.
The new version, 0.99, extends the package incorporating criteria that are more relevant when it is known that the data was collected with a particular type of analysis in view. Different kinds of statistical methodologies are considered within the framework of a multivariate linear model X = A Ψ + U, where X is the (n*p) data matrix of original variables, A is a known (n*q) design matrix, Ψ an (q*p) matrix of unknown parameters and U and (n*p) matrix of residual vectors. The new criteria are several descriptive indices, related to traditional test statistics, that measure the contribution of each subset to an "effect" characterized by the violation of a linear hypothesis of the form C Ψ= 0, where C is a known coefficient matrix of rank r. All these indices are functions of the r positive eigenvalues of a product H T-1 where H and T are matrices of "effect" and "total" squared and cross-product deviations associated with X.
Important cases within this framework include traditional canonical correlation analysis, in which the columns of A are observations on a fixed set of variables related to X, and C a (q*q) identity. If A consists on a single (dependent) variable, then the problem reduces to the traditional selection problem in linear regression analysis, using the R^2 coefficient as comparison criterion. Variable selection in generalized linear models can also be easily accommodated by specifying H and T in terms of appropriate Fisher information matrices. Linear Discriminant Analysis can be addressed by making A a matrix of group indicators, Ψ a matrix of group specific population means and the hypothesis C Ψ = 0 equivalent to the equality of all population means across groups. Searches for the variable subsets that best characterize a multivariate effect found by a multi-way MANOVA or MANCOVA analysis can also be easily addressed.
All the previous features and options of the subselect package are applicable to these new problems and criteria. In particular, for a moderate number of original variables, say less than 20 or 30, it is often possible to conduct an exhaustive search through all subsets using efficient adaptations of the classical Furnival and Wilson algorithm for variable selection. For larger data sets, several effective meta-heuristics are provided through reliable and updated implementations of simulated annealing, genetic and restricted local improvement algorithms. Furthermore, it is possible to forcibly include or exclude variables from the chosen subsets, specify the number of solutions to keep in each dimension, control several tuning parameters in the random search routines, specify a limit for the time spent in an exhaustive search and so on.

Key words: Variable selection algorithms. Heuristics. Linear Models. Generalized linear models. Discriminant analysis. Canonical correlation analysis

Using R for the Analysis of BeadArray Microarray Experiments

Mark Dunning, Natalie P Thorne, Michael Smith, Isabelle Camilier, Simon Tavaré
Department of Oncology, University of Cambridge, England

The development of diseases such as cancer is caused by fundamental changes in the function and morphology of cells in an organism. These changes are governed by the regulation of proteins produced in the cell, which are in turn regulated by the amount of expression of particular genes. Therefore it is of great interest to medical researchers to be able to compare the expression levels of genes between different conditions or samples. In recent years, the technology of microarrays has made it feasible to measure the gene expression levels of many thousands of genes in cells taken from different samples. The sheer volume of data produced even by a simple microarray experiment has led to a new inter-disciplinary subject within Bioinformatics which uses the expertise of biologists, computer scientists and mathematicians to be able to manipulate, analyse and draw meaningful biological conclusions from microarray experiments.

Illumina have created an alternative microarray technology (BeadArray) based on randomly arranged beads, each of which carries copies of a gene-specific probe. Random sampling from an initial pool of beads produces an array containing, on average, 30 randomly positioned replicates of each probe type. This degree of replication makes the gene expression levels obtained using BeadArrays more robust whilst spatial effects do not have such a detrimental effect as they do with conventional arrays, where there is often little or no replication of probes over an array. BeadArrays are already being used in a number of high-throughput experiments (eg www.hapmap.org).

Until now, analysis of BeadArray data was carried out by using Illumina's own software package and therefore did not utilise the wide range of Bioinformatic tools already available via the Bioconductor website (www.bioconductor.org). Also, the data output from this software only gives a single measurement for each bead type on an array, thus losing information about the replicates. The intention of our project was to create an R package (*beadarray*) for the analysis of BeadArray data incorporating ideas from existing microarray analysis R packages. We aimed to provide a flexible and extendable means of analysing BeadArray data both for our own research purposes and for the benefit of other users of BeadArray technology. The *beadarray* package also gives users access to the full dataset for each array.

We will describe the methods available in our R library for the reconstruction and analysis of BeadArray data. We will demonstrate the low variability and high reproducibility of data generated by BeadArray experiments along with methods for quality control. An important step in the analysis of microarray data is normalisation in which data from separate experiments are made comparable by removing any systematic variation between arrays. We will present our results on investigations in comparing and assessing the performance of various normalisation approaches (including the different preprocessing and background correction steps) for BeadArray data.

The latest version of the *beadarray* package is available now at
http://www.bioconductor.org/packages/bioc/1.8/html/beadarray.html

Data Analysis System with Graphical Interface

R. Dutter[1]

[1] Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria

Keywords: Computer Program System R, Robustness, DAS Data Analysis System, Geochemical Analysis.

Abstract

The tcl/tk package is used to establish a user-friendly data analysis system with special emphasis on geochemical analysis data. The starting point was the Data Analysis System DAS which methods and many more ideas are collected in the system DAS+R. Robust methods are a central point.

We present a program report and examples with spatial geochemical data.

References

J. Fox (2004). Getting Started with the R Commander: A Basic-statistics Graphical User Interface to R. *'useR 2004' Conference*, May 20-22, 2004, Vienna University of Technology, Austria.

C. Reimann, M. Äyräs, V. Chekushin, I. Bogatyrev, R. Boyd, P. de Caritat, R. Dutter, T.E. Finne, J.H. Halleraker, Ø. Jæger, G. Kashulina, O. Lehto, H. Niskavaara, V. Pavlov, M.L. Räisänen, T. Strand, and T. Volden. (1998). *Environmental Geochemical Atlas of the Central Barents Region. Geological Survey of Norway (NGU), P.O. Box 3006, Lade, N-7002 Trondheim, Norway.*

Use R fifteen different ways: R front-ends in Quantian

Dirk Eddelbuettel
edd@debian.org

Abstract submitted to useR! 2006

Quantian (Eddelbuettel, 2003) has become one of the most comprehensive environments for quantitative and scientific computing. Within Quantian, the R language and programming environment has always had a central focus: Quantian provides not only R itself, but numerous add-ons, interfaces as well as essentially all packages from the CRAN and BioConductor repositories.

With release 0.7.9.2 of Quantian[1], the list of 'interfaces' (where the term is used in a fuzzy and encompassing way) has increased considerably. The paper will briefly discuss and summarize all of the interfaces to R that are now included and ready to be deployed immediately:

1. R can of course be used the traditional way from the command-line or shell as `R`;
2. R can be used via the cross-platform graphical user interface when started as `R --gui=Tk`;
3. R can be invoked as `R --gui=gnome` using the Gnome/Gtk GUI available on Unix platforms;
4. Emacs Speaks Statistics permits to launch R via the `M-x R` combination from within XEmacs;
5. The Rcmdr GUI by John Fox is available from R via `library(Rcmdr)`;
6. The Rpad web interface by Tom Short is available at `http://localhost/Rpad`[2] – and the smaller Rcgi package is also available as an alternative;
7. The award-winning Java Gui for R (JGR) can be launched from the command-line via `JGR`;
8. R is one of several mathematical languages that can be launched and used directly from Texmacs;
9. An early version of the Rkward GUI is also available from the command-line under its name;
10. The headless Rserve R network service is available via `R CMD Rserve`;
11. R can be invoked from Python using the Rpy module;
12. Similarly, RSPerl permits R to be driven from Perl[3] – and the other way around;
13. R can also be embedded into other applications: one such example is provided by the PI/R procedural R language for the PostgreSQL RDBMS;
14. rJava permits R to be used from Java programs;
15. SNOW provides a high-level interface to distributed statistical computing, and the underlying components Rmpi and Rpvm are also available directly.

Quantian 0.7.9.2 also contain 877 packages providing a complete collection of R code. This set comprises essentially all Unix-installable packages from CRAN, the complete BioConductor relase 1.7, a few Omegahat packages as well as packages from J. Lindsey and T. Yee.

Lastly, several related software projects such as Ggobi, Mondrian, Weka or GRASS are available as well to further complement Quantianfor particular scientific communities.

Time permitting, we plan to demonstrate each of the fifteen different user interfaces from a running Quantianinstallation.

References

Dirk Eddelbuettel. Quantian: A scientific computing environment. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 2003. URL `http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/Eddelbuettel.pdf`. ISSN 1609-395X.

[1] This version should be released in early March 2006.

[2] An initial `/etc/init.d/apache start` is required.

[3] Environment variables have to be set. The first example can be started as `cd /usr/local/lib/R/site-library/RSPerl/examples; R_HOME=/usr/lib/R LD_LIBRARY_PATH=../libs/Perl perl -I../share/lib/perl/5.8.8 test.pl`.

How Much Can Be Inferred From Almost Nothing? A Maximum Entropy Approach to Fundamental Indeterminacy in Ecological Inference With an Application to District-Level Prediction of Split-Ticket Voting

Martin Elff*, Thomas Gschwend†, and Ron Johnston‡

Ecological inference aims at reconstructing data and data-generating processes at the individual level based on aggregate data. A typical application is the estimation of the probability of split-ticket voting on the individual level based on voting results on the level of electoral districts. Therefore, ecological inference has to rely on crucial assumptions about data generating processes at the level of the individual that cannot be tested by classical means of statistical hypothesis testing. This leads to a fundamental modelling indeterminacy: It is impossible to make sure that the statistical model of the data-generating process on which an ecological inference procedure relies is appropriate.

Focussing on the estimation of general $R \times C$ voting matrices of individual electoral districts, we address this problem of modelling indeterminacy. Our approach at solving this problem builds on the Principle of Maximum Entropy. We show how this principle can be used to determine models that exploit all information about the available aggregate data but are structurally independent from unavailable individual-level data. Further, we consider the apparent paradox that such entropy maximizing models imply very restrictive assumptions about the data generating process and explain how to reconcile the need of such restrictive assumptions with the abovementioned fundamental modelling uncertainty.

By way of extensive simulation studies conducted with R, we show that entropy maximizing models lead to unbiased predictions of district-level voting matrices. With these simulation studies we show further, how uncertainty about these predictions, to which fundamental modelling indeterminacy leads, can be assessed based on the Principle of Maximum Entropy. Finally, we apply this approach to the prediction of split-ticket voting at the level of electoral districts during the 1992 General Election of New Zealand.

*Dept. of Social Sciences, University of Mannheim (corresponding author)
†MZES, University of Mannheim
‡School of Geographical Sciences, University of Bristol

SparcMats and Generalized Pairs Plots

John Emerson[1], Walton Green, Avi Feller, and John Hartigan

We would like to present several graphical innovations and/or improvements. Although we hope others will find these tools interesting and useful, we are particularly enthusiastic about the power and flexibility of R for teaching and research. These graphical tools were motivated by research problems and developed in a classroom environment.

First, the proliferation of data in the computer age has made the search for methods of compressing higher dimensions onto a flat, two-dimensional graphical display more difficult and important. We illustrate a graphical, exploratory approach using a generalization of Edward Tufte's "sparklines," that can be used to represent data of continuous variables distributed in space and time. Using this and other graphical tools, we identify certain peculiarities of seven 4-year-long monthly time series of climate data on a 24 by 24 raster covering an area of South and Central America.

Second, we propose a generalized pairs plot, recognizing the fundamental importance of the roles of categorical and quantitative variables. Others have produced pairs plots for categorical variables (with each tile consisting of a small mosaic plot). We introduce a new "barcode" plot, envisioned by John Hartigan (having similarities to side-by-side boxplots and histograms), and we generalize the pairs plot to include (a) scatterplots for pairs of quantitative variables, (b) mosaic plots for pairs of categorical variables, and (c) boxplots and barcode plots for pairs consisting of one categorical and one quantitative variable. Additional options are provided.

[1] Corresponding and contributing author.

Cluster Analysis: Past, Present and Future

Brian S. Everitt

Cluster analysis is a generic term for a wide range of numerical methods for examining data with a view to detecting, uncovering or discovering groups or 'clusters' of objects or individuals that are (1) homogeneous and (2) separate. Many of the clustering methods is use today have resulted from a crossfertilization between disciplines such as psychology, biology and psychiatry on the one hand and mathematics and statistics on the other. The result has been a considerable amount of ad hoc development and 'reinvention'—as somebody once remarked 'there may be as many clustering techniques as there are cluster analysis users'. In the last five years or so, cluster analysis has become part of the data mining industry and has found application in grouping together genes with similar patterns of expression.

This talk will look at some early papers on cluster analysis that were important to the speaker, the current state of the methodology and speculate briefly about its future development.

ZipfR: Working with words and other rare events in R

Stefan Evert, University of Osnabrück, Germany (stefan.evert@uos.de)
Marco Baroni, University of Bologna, Italy (baroni@sslmit.unibo.it)

The field of linguistics has recently undergone a methodological revolution. Whereas earlier on most linguists had relied solely on introspection, recent years have seen the rise to prominence of *corpora*, i.e. large samples of texts, as the main source of linguistic data [5]. Because of this shift, statistical analysis plays an increasingly central role in the field. However, as has been known since the seminal work of George Kingsley Zipf (e.g. [6]), standard statistical models (in particular all those based on normality assumptions) are not suitable for analyzing the frequency distributions of words and other linguistic units. Even in the largest corpora currently available (containing above one billion running words of text), word frequency distributions are characterized by a high proportion of word types that occur only once or twice. When the sample size is increased further, a non-negligible number of new types will be encountered about which the original sample did not contain any information at all. Because of these properties, often referred to as the "Zipfianness" of language data, estimation of occurrence probabilities is unreliable (even when confidence interval estimates are used, cf. [3, Ch. 4]), the central limit theorem no longer guarantees the normality of sample averages for large samples, and the number of types in the population (which has an important linguistic interpretation as the overall vocabulary size of a certain language or sub-language) cannot easily be estimated from the observed data.

In the technical literature, various equations have been proposed for modelling the probability distribution of a Zipfian population. Baayen has summarized much of this work in [1], accompanied by a software package (`lexstats`) that can be used to estimate the parameters of different population models from an observed sample, and then calculate the expected values and variances of certain sample statistics (in particular, the number of distinct types in a sample of given size, as well as the number of types occurring once, twice, etc.). However, the `lexstats` package has only found limited use among linguists, for a number of reasons: `lexstats` is only supported under Linux, its ad-hoc Tk user interface has minimal functionality for graphing and data analysis, it has extremely restrictive input options (which make its use with languages other than English very cumbersome), and it works reliably only on rather small data sets, well below the sizes now routinely encountered in linguistic research (cf. the problems reported in [4]).

Following our positive experience implementing an R library of frequency comparison tests geared towards linguists with limited mathematical background (the `corpora` library available from CRAN), we decided to develop an R-based solution as an alternative to the `lexstats` suite. Our `ZipfR` library, which integrates code from the first author's UCS project (see `http://www.collocations.de/software.html`) currently provides implementations of three population models: Zipf-Mandelbrot, finite Zipf-Mandelbrot [2] and Generalized Inverse Gauss-Poisson (cf. [1]), although we expect to add other models in the future. The `ZipfR` library features an object-oriented design, in which units that should be intuitive to linguists (such as word frequency lists and vocabulary extrapolation experiments) are treated as objects. It relies on the R standard library for special functions (e.g. `besselI`) and statistical distributions, as well as general numerical utilities (e.g. `solve` for matrix inversion in a multivariate chi-squared test and `nlm` for parameter estimation). In our current tests, `ZipfR` model estimation has proven to be robust and efficient even for the largest data sets encountered in our experiments. A set of pre-designed plots are provided for the most common types of graphs (e.g. vocabulary growth curves and frequency spectra, see [1]), while allowing experienced users to take advantage of the full range of R graphics facilities. Furthermore, we have developed a package of auxiliary Perl scripts to extract word frequency data from corpora (including sophisticated randomization options in order to test the randomness assumption underlying the statistical models, which is

often problematic for language data), and to import existing data from `lexstats` as well as a range of other formats commonly used by linguists.

We hope that the availability of a powerful, user-friendly and flexible tool such as `ZipfR` will encourage more linguists to use advanced statistical models in their work that are suitable for the Zipfian distribution of word frequencies. As a welcome side-effect, this will also familiarize them with R itself and thus make the software more widespread in the linguistic community. A first public release of our library will be available from CRAN by the time of the conference, with supplementary information to be found on its homepage `http://purl.org/stefan.evert/ZipfR`.

References

[1] Baayen, Harald (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

[2] Evert, Stefan (2004a). A simple LNRE model for random character sequences. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles*, pages 411–422.

[3] Evert, Stefan (2004b). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

[4] Evert, Stefan and Baroni, Marco (2005). Testing the extrapolation quality of word frequency models. In: *Proceedings of Corpus Linguistics 2005*.

[5] McEnery, Tony and Andrew Wilson (2001). *Corpus Linguistics*. 2nd edition, Edinburgh: Edinburgh University Press.

[6] Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Cambridge MA: Addison-Wesley.

Term structure and credit spread estimation with R

Robert Ferstl

Department of Operations Research,
Vienna University of Economics and Business Administration

Zero-coupon yield curves and credit spread curves are important inputs for various financial models, e.g. pricing of securities, risk management, monetary policy issues. Since zero-coupon rates are rarely directly observable, they have to be estimated from market data for existing coupon bonds. The literature broadly distinguishes between parametric and spline-based methods. We implement three widely-used term structure estimation procedures, i.e. the parametric Nelson and Siegel approach, the Svensson approach and the cubic splines method.

The traditional way of credit spread calculation is to subtract individually estimated zero-coupon yield curves from a risk-free reference curve. This approach often leads to twisting credit spread curves. These shapes are unrealistic and problematic if used as inputs for financial models. Therefore, we implement the existing joint estimation procedures, which return smoother credit spread curves. Goodness-of-fit tests are provided to compare the results of the different estimation methods. We illustrate the usage of our functions by practical examples with data from European and CEE government bonds, and European corporate bonds.

Outlier Detection with Application to Geochemistry

Peter Filzmoser
Department of Statistics and Probability Theory
Vienna University of Technology

Keywords: Outliers, Robustness, Multivariate methods, Extremes

Abstract

Outlier detection belongs to the most important tasks in data analysis. The outliers describe the abnormal data behavior, i.e. data which are deviating from the natural data variability. Often outliers are of primary interest, for example in geochemical exploration they are indications for mineral deposits. The cut-off value or threshold which divides anomalous and non-anomalous data numerically is often the basis for important decisions.

Many methods have been proposed for univariate outlier detection. They are based on (robust) estimation of location and scatter, or on quantiles of the data. A major disadvantage is that these rules are independent from the sample size. Moreover, by definition of most rules (e.g. mean $\pm 2\cdot$ scatter) outliers are identified even for "clean" data, or at least no distinction is made between outliers and extremes of a distribution (Reimann, Filzmoser, and Garrett, 2005).

The basis for multivariate outlier detection is the Mahalanobis distance. The standard method for multivariate outlier detection is robust estimation of the parameters in the Mahalanobis distance and the comparison with a critical value of the χ^2 distribution (Rousseeuw and Van Zomeren, 1990). However, also values larger than this critical value are not necessarily outliers, they could still belong to the data distribution.

In order to distinguish between extremes of a distribution and outliers, Garrett (1989) introduced the χ^2 plot, which draws the empirical distribution function of the robust Mahalanobis distances against the χ^2 distribution. A break in the tails of the distribution is an indication for outliers, and values beyond this break are iteratively deleted. Gervini (2003) used this idea and compared theoretical and empirical distribution function in the tails to define the proportion of outliers in the data. In a further development, Filzmoser, Garrett, and Reimann (2005) adjusted the adaptive method of Gervini (2003) to sample size and dimensionality. It turns out that the resulting outlier detection method is not very sensitive with respect to the choice of tuning parameters (Filzmoser, 2005). The method has been implemented in `R` in the package *mvoutlier*.

In an application with data from geochemistry the usefulness of the proposed method is demonstrated. Moreover, we propose a new plot for visualizing multivariate outliers of spatial data.

P. Filzmoser. Identification of multivariate outliers: a performance study. *Austrian Journal of Statistics*, 34(2):127–138, 2005.

P. Filzmoser, R.G. Garrett, and C. Reimann. Multivariate outlier detection in exploration geochemistry. *Computers and Geosciences*, 31:579–587, 2005.

R.G. Garrett. The chi-square plot: A tool for multivariate outlier recognition. *Journal of Geochemical Exploration*, 32, 319–341, 1989.

D. Gervini. A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *Journal of Multivariate Analysis*, 84, 116–144, 2003.

C. Reimann, P. Filzmoser, and R.G. Garrett. Background and threshold: critical comparison of methods of determination. *Science of the Total Environment*, 346:1–16, 2005.

P.J. Rousseeuw and B.C. Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), 633–651, 1990.

Robust Principal Component Analysis by Projection Pursuit

Heinrich Fritz and Peter Filzmoser
Department of Statistics and Probability Theory
Vienna University of Technology

Abstract: Different algorithms for principal component analysis (PCA) based on the idea of projection pursuit are proposed. We show how the algorithms are constructed, and compare the new algorithms with standard algorithms. With the R implementation *pcaPP* we demonstrate the usefulness at real data examples. Finally, it will be outlined how the algorithms can be used for robustifying other multivariate methods.

Keywords: Projection pursuit, Robustness, Principal component analysis, Multivariate methods, pcaPP

1 Introduction

Many multivariate statistical methods are based on a decomposition of covariance matrices. For high-dimensional data this approach can be computationally intensive, especially if the involved covariance matrices should be estimated in a robust way. Moreover, if the sample size is lower than the dimension, additional problems with robust covariance estimation will arise.

An alternative approach for obtaining robust multivariate methods is projection pursuit (Huber, 1985). For example, in PCA the first component is defined as that direction maximizing a measure of spread of the projected data on this direction. If a robust spread measure is considered, the resulting PC is robust. Thus, robust estimation is done only in one dimension, namely in the direction of the projected data.

A non-trivial task is finding the direction which maximizes an objective function, like a robust spread measure for robust PCA. In this context, Croux and Ruiz-Gazen (2005) suggested to use each observation for the construction of candidate directions. We will extend this idea and introduce other algorithms. In a straightforward manner we can also obtain other (robust) multivariate methods.

2 Extensions of the Algorithm of Croux and Ruiz-Gazen (2005)

Croux and Ruiz-Gazen (2005) suggest to use as candidate directions for the first PC all directions from each data point through the center of the data cloud, estimated e.g. by the L_1-median. Subsequent PCs are estimated in a similar way, but the search is done in the orthogonal complement of the previously identified PCs. However, due to its construction, this algorithm may not be very precise for data sets with low sample size n or where n/p is low, with p being the number of variables. And there is yet another problem: By construction, the direction is determined by one of the data points. When the data are projected to the orthogonal complement, the projection of this data point is zero. This can lead to implosion of the scale estimator if p is sufficiently high.

To avoid these drawbacks one can add an updating step which is based on the algorithm for finding the eigenvectors. The drawbacks of the algorithm of Croux and Ruiz-Gazen (2005) can also be avoided by taking, in addition to the n data points, other candidate directions for maximizing the objective function. These directions are randomly generated: Generate n^+ data points with p-dimensional multivariate standard normal distribution, and project the data to the unit sphere. The directions of each generated data point through the origin are the new random directions, and by definition they have norm one.

3 Grid Algorithm

The optimization is always done in a plane rather than in the p-dimensional space. The first step is to sort the variables in descending order according to the largest scale. Then the optimization is done in the plane spanned by the first two sorted variables, where the candidate directions are constructed by dividing the unit circle into a regular grid of segments. A second approximation of the projection direction is then found by maximizing in the plane formed by the first and the third sorted variable. This process is repeated until the last variable has entered the optimization, which completes the first cycle of the algorithm. In a second cycle each variable is in turn again considered for improving the maximal value of the objective function. The algorithm terminates after a fixed number of cycles or when the improvement is considered to be marginal.

4 Robust Multivariate Methods

Above we described algorithms for estimating the (robust) PCs. We can use these for building a (robust) covariance matrix, which then can be plugged in into multivariate methods like factor analysis, canonical correlation analysis or discriminant analysis. On the other hand, some of the multivariate methods can be reformulated as a projection pursuit method, and the above algorithms could be applied. This approach was used for robust continuum regression (Filzmoser et al., 2006).

Croux, C., and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95, 206-226.

Filzmoser, P., Serneels, S., Croux, C., and Van Espen, P.J. (2006). Robust multivariate methods: The projection pursuit approach. In: Spiliopoulou, M., Kruse, R., Nürnberger, A., Borgelt, C., and Gaul, W. (Eds.), *From Data and Information Analysis to Knowledge Engineering*, Springer-Verlag, Heidelberg-Berlin. To appear.

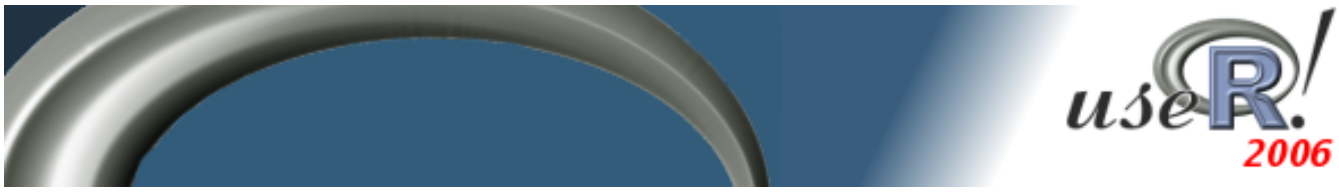
Huber, P.J. (1985). Projection pursuit. *The Annals of Statistics*, 13 (2), 435-475.

UseR! for Teaching

John Fox and Sanford Weisberg
McMaster University and University of Minnesota

The first author will describe the use of R in several classes, all for sociology students: two introductory courses (one for undergraduates and the other for graduate students), in which students interact with R through the R Commander graphical-user interface; a graduate course on applied regression and generalized linear models; and a more advanced graduate course that takes up several topics, including structural-equation models, survival analysis, and mixed-effects models for hierarchical and longitudinal data. A general theme will be the flexibility of R and the ease with which it can be customized and adapted for a variety of teaching tasks.

The second author will describe the use of R with his recently published textbook, *Applied Linear Regression, 3rd edition*, for which "computing primers" have been written for R and several other popular statistical languages. A point that will be emphasized is the difference between courses that "teach to the package" and those that view packages as secondary (or even tertiary). Both approaches have their merits. We will then turn to issues that arise in thinking about a new book project centered on nonlinear regression models. A new package in preparation to support this project will be discussed. The talk will conclude with discussion of a few things that I wish I could do in R but don't know how to do them.



Double Cross Validation for Model Based Classification

Romain François* and Florent Langrognet†

February 28, 2006

Keywords: Cross Validation, Classification, Gaussian Mixtures.

Introduction

Gaussian mixture modelling is a powerful framework for classification. The observations $x_1, \ldots, x_n$ are assumed to arise from a mixture of K normal distributed components.

$$f(.) = \sum_{k=1}^{K} p_k \Phi(.|\mu_k, \Sigma_k) \quad , \quad 0 < p_k < 1 \quad , \quad \sum_{k=1}^{K} p_k = 1 \tag{1}$$

where p_k are the mixing proportions, $\mu_k \in \mathbb{R}^d$ the mean vector of the k^{th} component, Σ_k its covariance matrix and $\Phi(.|\mu, \Sigma)$ the normal probability density function with mean vector μ and variance matrix Σ.

Celeux and Govaert (1995) proposed a decomposition of the variance matrices in terms of *volume*, *orientation* and *shape*. That decomposition yields 14 models from the simplest $[\lambda I]$ (same volume, shape and orientation with spherical variance matrices) to the standard QDA model $[\lambda_k C_k]$.

Mixmod

MIXMOD[1] is an open source C++ software for Gaussian mixture modelling with EM-like algorithms. MIXMOD proposes all the 14 models from Celeux and Govaert (1995) and model selection criteria based on penalized likelihood (BIC, ICL, NEC) or quality of prediction via cross validation. MIXMOD is originally interfaced with MATLAB or SCILAB and has been ported to R recently. The R package *mixmod* should be available on CRAN soon.

```
out  <-mixmod(iris[,1:4], nbCluster=3)
plot(out, type="zones") # produces fig 1 : left
rgl(out, contours=TRUE, obs=TRUE) # produces fig 1 : right
```

Double Cross Validation

The cross-validated error rate may be used to select a model between several candidats when quality of prediction matters. However, that error rate is too optimistic as it does not take into account the uncertainty of the selection procedure.

The *double* cross-validated error rate that we propose makes use of the cross validation methodology at two stages in order to estimate the overall error rate of the procedure *"model the observations by a Gaussian mixture with one of the 14 structures"*. It proceeds as follows :

*INRIA Futurs, projet SELECT. Corresponding Author. R. François, Université Paris SUD, Bat. 425, 91405 Orsay Cedex (France). email : Romain.Francois@inria.fr

†UMR 6623 CNRS, Université de Franche-Comté

[1]The MIXMOD program and its documentation are available freely on the internet: http://www-math.univ-fcomte.fr/mixmod/

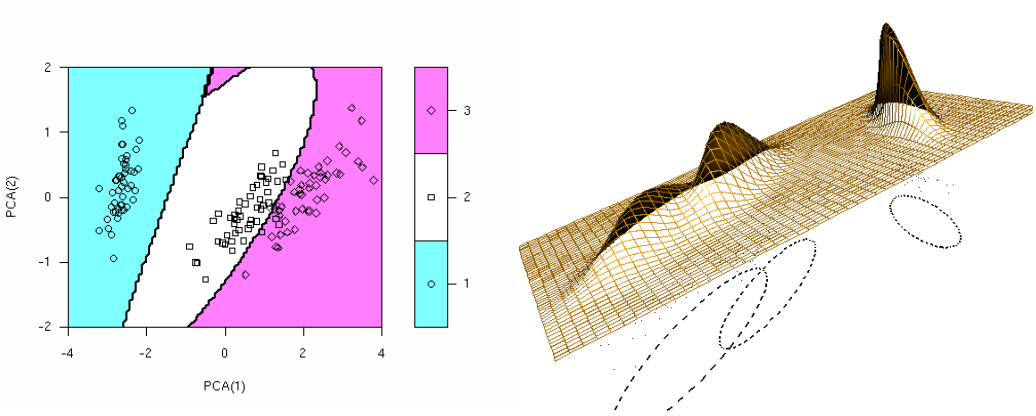


Figure 1: Some MIXMOD graphics. The one on the left uses `filled.contour` to display the classification zones in the two first PCA axes. The graphic on the right uses `rgl` to display the density mixture and the cluster's 95% iso-density ellipses.

- Random split the sample S in V sub-samples $S^{(1)}, \ldots, S^{(V)}$
- For $v = 1, \ldots, V$, do :
  - Merge $V-1$ subsamples into $S^{(-v)} = S - S^{(v)}$
  - For each candidate model $m \in \mathcal{M}$, compute the discrimination rule $\theta_m^{(-v)}$ and select the best model regarding the cross-validated error-rate :

$$m_v^\star = \underset{m \in \mathcal{M}}{\operatorname{argmin}}\ CV_m^{(-v)}$$

  - Evaluate the error rate t_v of $m_v^\star$ on the test sample $S^{(v)}$
- Average the V error rates $t_1, \ldots, t_V$

Double cross validation : Sketch of the algorithm

The double cross-validated error rate gives a good point estimate of the error rate and a measure of its variability. Moreover, the frequencies of the models inside the V winners $m_1^\star, \ldots, m_V^\star$ may be used to qualify the stability of these models. A model selected frequently has good chances to be well adpated to the problem.

```
R> out  <- mixmod(iris[,1:4], crit="DCV", lab=as.numeric(iris[,5]))
R> out$modelOutput$DCV
[1] 0.0333333
```

References

Birenacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics & Data Analysis*, to appear.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–93.

François, R. (2006). *Sélection d'un modèle de mélange pour la classification. Double validation croisée. Application aux données de biopuce.* INRIA Futurs. Mémoire de stage ISUP. `http://addictedtor.free.fr/rapportStage`.

Capturing Unobserved Heterogeneity in the Austrian Labor Market Using Finite Mixtures of Markov Chain Models

Sylvia Frühwirth-Schnatter and Christoph Pamminger

Johannes Kepler Universität Linz
Institut für Angewandte Statistik (IFAS)

Wage mobility in a labor market describes the chances but also the risks of an individual to move between wage categories. Wage mobility is usually measured on an aggregate level for the whole labor market and little work is done to capture unobserved heterogeneity across individuals. The present study tries to obtain some results on unobserved heterogeneity for the Austrian Labor Market.

For each individual, wage mobility is described through a first order Markov chain which is heterogenous across the individuals. We will compare two approaches for capturing unobserved heterogeneity. The first approach is based on using a finite mixture of Markov chains models moving with different speed. In this approach it is assumed that within each hidden group no more heterogeneity is present. This leads to a rather large number of groups which are not easily interpreted. The second approach is based on using finite mixtures of multinomial-Dirichlet distributions. In this approach it is assumed that within each hidden group heterogeneity in wage mobility is still present which may be described through a Dirichlet distribution with an unknown dispersion parameter.

Practical implementation in R is based on combing a Bayesian approach with Monte Carlo simulations based on Markov chains.

Title: **SimSurvey - an R-based E-learning tool for geo-statistical analyses**

Abstract: useR! 2006
June 15-17 2006, Wien

Authors:
Mario Gellrich[1], Rudolf Gubler[1], Andreas Papritz[1], Andreas Schönborn[1], Rainer Schulin[1]

Affiliations:
Swiss Federal Institute of Technology (ETH), Institute of Terrestrial Ecology (ITE), Universitätsstrasse 16, CH-8092, Zürich

Assessment of land pollution and soil degradation is a task that environmental scientists and engineers may face in their daily work. Geo-statistical methods are often used for such purposes. Geo-statistics lessons are part of the academic education of many environmental scientists, but as experiences show, the complexity of statistical methods is often difficult for students. We are developing an E-learning tool to complement the theoretical lessons for students. The aim of our project, 'SimSurvey', is to make the learning of geo-statistical methods easier. By creating a virtual environment (similar to the situation in a professional consulting company), students will face 'real life' problems such as the collection of soil samples, geo-statistical analyses of the collected data and the handling of financial recourses. SimSurvey consists of a 'project environment' and a statistics module. The project environment facilitates students to navigate through virtual maps, manage recourses and handle data. The statistics module consists of a graphical user interface (GUI) and a classical programming environment, which makes it a very flexible tool for the analysis of spatially explicit data. SimSurvey runs on a Linux-Server and requires Apache, Flash-Player, MySQL, PHP and R. As Flash uses XML-files to create the R-GUI, the statistics module can be easily extended and adapted without extensive programming knowledge. Single R processes run via socket connections, i.e. there is a permanent connection between the server and each R-process. This makes statistical computing faster than in 'batch mode', which is used in most existing web-based R-projects. SimSurvey is still under development. To date, the GUI allows students to explore spatial datasets graphically and by means of statistical models. Linear regression and geo-statistics functions (variograms, kriging) are implemented for multivariate data analyses and predictions.

Introduction to S programming: a teaching experience and a manual

Vincent Goulet

École d'actuariat, Université Laval, Québec, Canada

The goal of this presentation is two fold: share my experience in teaching S programming and present my manual "Introduction to S programming".

Teaching a programming language is no easy task, especially to large groups and with limited access to computer labs. In this presentation, I will share my experience in teaching R as a second programming language to first year undergraduate students. My approach is based on a compromise between time spent in class and time spent in a lab.

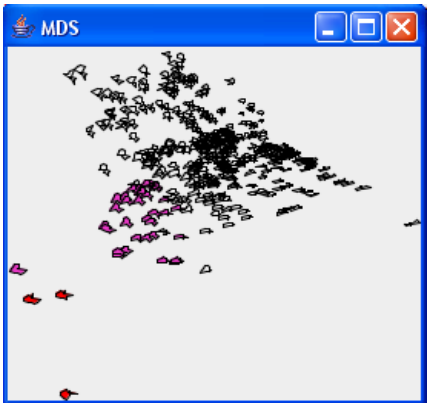
This approach is reflected in my document "Introduction to S programming", which started as class notes and later evolved into a rather complete introductory manual on S programming. There are already many documents and textbooks on S-Plus and/or R. However, most present the software within the framework of specific statistical applications. The forthcoming translation of my French document already available on CRAN rather focuses on learning the S programming language. Being published under the GNU Free Documentation License, anyone is free to reuse the manual, in whole or in part.

Interactive Glyph Analysis with R

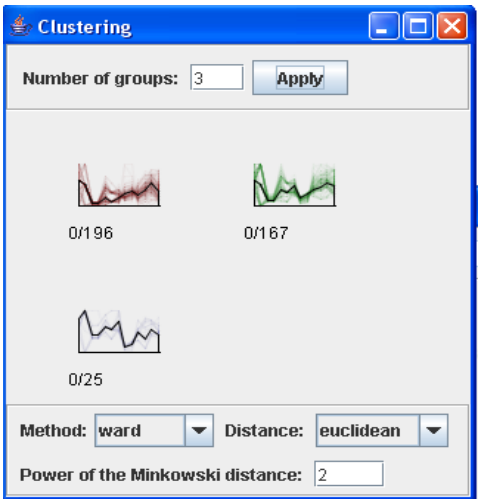
GAUGUIN (Grouping And Using Glyphs Uncovering Individual Nuances) is a project for the interactive visual exploration of multivariate data sets, developed for use on all major platforms (Windows, Linux, Mac). It supports a variety of methods for displaying flat-form data and hierarchically clustered data.
Glyphs are geometric shapes scaled by the values of dataset variables. They may be drawn for individual cases or for averages of groups or clusters of cases. GAUGUIN offers four different glyph shapes (but more could be added).
The number of data elements which can be displayed simultaneously is limited, because each glyph requirees a minimum amount of screen space to be viewed, but hierarchical glyphs can be drawn for groups of cases. Hierarchical glyphs are composed of a highlighted case representing the group and a band around it showing the variability of all the members of the cluster.
GAUGUIN also provides scatterplots and tableplots, and via Rserve is able to use R to calculate MDS views and clusters for the data. All GAUGUIN displays are linked interactively and can be directly queried.

Some plots:

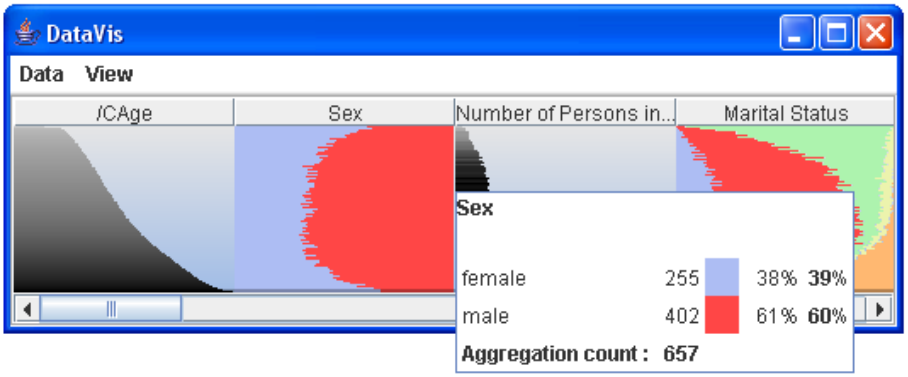
MDS (Multidimensional scaling):



CLUSTERING:



TABLEPLOT:


DataVis
Data View
/CAge
Sex
Number of Persons in...
Marital Status
Sex
female 255 38% 39%
male 402 61% 60%
Aggregation count : 657

Collaborative writing of R documentation using a Wiki

Philippe Grosjean (Numerical Ecology of Aquatic Systems,
Mons-Hainaut University, Belgium, phgrosjean@sciviews.org)

Thanks to its Open Source license, R is developed by a large community of contributors. This makes a big part of the success of R, and also, it allows to propose a wide palette of additional functions in hundreds of R packages distributed on CRAN, Bioconductor, Omegahat, etc. Documentation (both for users and for developers) is an important component of any software. R proposes a couple of formats to standardize the way functions are documented: the Rd file, a LaTeX-like syntax that can be converted into different formats (using Rdconv), and the vignette, using Sweave to compile R code embedded in LaTeX document.

Collaboration on writing these Rd files or vignettes is only possible by placing them in a CVS, or another similar collaboration system. There is no built-in collaboration tools in the R documents themselves. Currently, feedback about R documentation is done by sending bug reports (sic!), or by writing directly to the author(s) by email, thus, very basic and somehow inadequate ways to collaborate! There is also a large amount of documents (tutorials, manuals, reference cards, etc.) available on the web. A couple of web sites are really nice and propose interesting ways to share R code and R documentation, a good example being the R Graph Gallery (http://addictedtor.free.fr/graphiques/, by Romain Franois). Here again, there is little way to collaborate online on writing the documents: in the best situation, an author can submit a document to the server, and everybody has a read-only access to the page, once it is published.

A Wiki is basically a web server where pages are editable by the readers directly in their web browser. Wiki pages use a very simple syntax that allows for easy formatting of even complex pages (for most advanced Wiki engines). One of the best example of what a Wiki can do is Wikipedia, with more than 992,250 articles (English version only) written and edited by a large community of volunteers. Wikipedia was recently compared to the well-known Encyclopedia Britannica by the scientific journal Nature, and it appeared that the quality of Wikipedia articles written by volunteers was equal to those written by paid experts in the Encyclopedia Britannica. This demonstrates the power of Wiki to write high-quality documentation is a collaborative way.

There is a Wiki dedicated to R since a couple of years initiated by Detlef Steuer on http://fawn.unibw-hamburg.de/cgi-bin/Rwiki.pl. However, this Wiki has not grown as expected. We believe that it could be due to two main reasons: (1) not enough publicity about this site, and (2) the use of a simplistic Wiki engine that does not provide all R-specific features that would make R Wiki pages attractive (R code highlighting, direct link to the documentation for R functions and packages, etc.).

We present here a new Wiki dedicated to R. The Wiki engine is based on DokuWiki, a power system targeting software documentation. This engine is modified and R-specific plugins are added to make it most suitable to edit R documentation. There are plugins for syntax highlighting of R code, for direct linking to the R functions documentation, or to the home page of R packages, etc.).

As for the content, various authors have already accepted to move their documents to the new R Wiki: material from Detlef Steuer's Wiki, Paul Johnson's Rtips, Vincent Zoonekynd's Statistics with R, James Wettenhall's R tcltk examples, etc.

The structure of the Wiki site has received much attention, and many people collaborated on making it clearer, easier, more efficient, initiating the really collaborative work around it (in alphabetic order: Jonathan Baron, Damian Betebenner, Roger Bivand, Ben Bolker, Patrick Burns, Nick Drew, Jose Claudio Faria, David Forrest, Romain Franois, Gabor Grothendieck, Frank

Harrell, Paul Johnson, Martin Maechler, John Marsland, Duncan Murdoch, Tony Plate, Barry Rowlingson, Paul Sorenson, Detlef Steuer -sorry for those I forget to include in the list-). There is also a mailing list dedicate to this Wiki: 'r-sig-wiki' on https://stat.ethz.ch/mailman/listinfo/r-sig-wiki.

The R Wiki is currently in the form of a prototype, but it will be available before June. The R-Core Team has decided to support one or several R Wiki initiatives, and the final version of this Wiki will probably be available through a simple URL like http://wiki.r-project.org or http://www.r-project.org/wiki. It will run on a dedicated server for maximum performance.

Fisheries modelling in R: the FLR (Fisheries Library in R) project

Philippe Grosjean*, Richard Hillary, Ernesto Jardim, Laurie Kell,
Iago Mosqueira, Jan Jaap Poos, Robert Scott & Hunther S. Thompson

An intiative aimed at providing fisheries science with a comprehensive, open source and flexible toolbox based on R is presented here. The FLR project is organised around a core package (FLCore) that provides the basic S4 classes for many common fisheries data types and modelling tasks. The fundamental building block is a 5-dimensional array (an FLQuant) designed to accomodate the multiple spatio-temporal dimensions of fisheries data, either biological, technological or economic.

The FLCore package makes extensive use of S4 classes, representing in some cases very complex data structures, and provides a number of extensible modelling mechanisms for other packages to build on. Example classes are presented and the advantages and possible limitations of the current S4 system are analysed from our experience.

Packages have been developed that extend the basic classes and provide functions and methods useful in fisheries stock assessment, catch and abundance forecast, evaluation of fisheries management strategies, and stochastic simulation of fisheries systems. Use has been made in many cases of legacy code, and the mechanisms developed to simplify this task, i.e. C++ headers, are presented and tested.

Finally a number of suggestions for collaboration and linkage with other R packages, some of them actively ongoing, are presented.

*Numerical Ecology of Aquatic Systems, Mons-Hainaut University, Belgium, phgrosjean@sciviews.org

The optmatch package: flexible, optimal matching for observational studies

Ben B. Hansen

Matching is used both for design and analysis of case-control studies, of quasiexperiments with a treatment and a control group, and of cohort studies, among others. In conjunction with propensity scores (Rosenbaum and Rubin 1983), it offers an attractive alternative to adjustment based on regression, stably adjusting for the high-dimensional covariates that are common in social and medical science research. Greedy algorithms for pair matching are simple to implement; but much better matchings, matchings that support estimation of treatment effects that's both more efficient and less biased, can be had using a flexible, optimal matching routine (Hansen, 2004). Coding such a routine is somewhat subtle, involving discrete optimization among other steps (Hansen and Klopfer, 2006), so it is perhaps unsurprising that until recently none was publicly available. The R package "optmatch" was developed to fill this gap.

The presentation will briefly discuss the internals of the package before presenting an overview of its architecture from the perspective of the user, and discussing an application. Among the issues covered will be how it combines with propensity scoring and other multivariate distances; its adaptations for large, memory-intensive problems; its use to produce optimal pair matches, matches with multiple controls, full matches, and full matches with restrictions; analogues of goodness-of-fit tests for matchings, including some from a related package, "ritools", that J. Bowers and I are developing; and matched analysis for treatment effects.

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

Hansen, B.B. (2004), Full matching in an observational study of coaching for the SAT, Journal of the American Statistical Association, 99, 609–618.

Hansen, B.B. and Klopfer, S.O. (2006), Optimal full matching and related designs via network flows, Journal of Computational and Graphical Statistics, to appear.

Statistical Principles to Live By

Frank E. Harrell Jr.
Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, Tennessee, USA

This talk deals with principles derived from over 30 years of applying statistics to biomedical research, collaborating with clinical and basic biological researchers and epidemiologists. The principles relate to statistical efficiency, bias, validity, robustness, interpretation of statistical results, multivariable predictive modeling, statistical computing, and graphical presentation of information. Topics to be discussed include respecting continuous variables, avoiding non-descriptive statistics, problems associated with filtering out negative results, overfitting, shrinkage, adjusting P-values for multiple comparisons without adjusting point estimates for same, and the false promise of multi-stage estimation and testing procedures, related to the use of bogus conditional techniques for computing what is advertised as unconditional variances or type I errors.

Adventures in High Performance Computing and R: Going Parallel

Justin Harrington & Matias Salibian-Barrera

Parallel computing refers to the ability to use multiple CPUs concurrently to perform calculations that would otherwise be carried out sequentially in a single CPU. While not all statistical applications can benefit from using parallel computation, those that can are able to scale near-linearly their processing times for increasing numbers of CPUs, all other things remaining constant.

Generally, researchers whose area of interest is not computer science tend to look at parallel computing as something that would require: (a) access to an expensive multiple CPU machine or cluster, and (b) that they completely re-write their code to accommodate this sophisticated architecture. In this talk I will argue that nowadays these concerns should not prevent R users from exploring (and using) parallel computing.

Although it is true that massive computer systems with multiple CPUs are still beyond the means of most individual researchers, technology now exists for establishing virtual "parallel computing" clusters using groups of non-homogeneous computers that sit idle for significant lengths of time (such as those in a teaching lab, for example). This "virtual cluster" can also be constructed using the CPUs inside a multi-CPU machine running a standard OS.

The other important consideration that keeps many researchers away from considering parallel computing is the perceived large overhead cost involved in re-writing their computer code to take advantage of multiple CPUs. Fortunately, this is not necessarily the case, as long as the original R code was written in an efficient way (from a single CPU point of view) because R has several readily available libraries that allow us to use multiple CPUs with minimal changes to a "standard" R program / function.

In this talk I will be discussing how R can take advantage of parallel processing. In particular, I will discuss the libraries `rpvm`, `Rmpi` and `snow` that facilitate migrating "standard" R code to take advantage of multiple CPUs (either in the same motherboard or across the internet). I will also briefly discuss MPI (Message Passing Interface) and PVM (Parallel Virtual Machine), the protocols used to control the architecture.

This talk will be presented from the perspective of a statistician rather than a computer scientist, and the focus will be on helping users get started. I will discuss how to use these libraries and demonstrate their application on one clustering method, Linear Grouping Analysis (Van Aeslt, Wang, Zamar & Zhu (2006)), where this strategy of parallel computing has been successfully applied.

Data Mining in R: Regularization Paths

Trevor Hastie

Regularization is a popular approach to model selection, with L2 and L1 taking center stage. Recently there has been a spate of research on efficient algorithms for computing regularization paths. In this talk I will discuss and demonstrate three R packages that implement path algorithms:

- LARS: Least angle regression and extensions, for lasso and related paths for regression.
- GLMPATH: An extension of LARS for L1 regularized generalized linear models, including coxpath().
- SVMPATH: a regularization path algorithm for fitting support-vector machine classifiers for all possible values of the cost parameter.

eRm - extended Rasch modelling

An R Package for the Application of Item Response Theory Models

Reinhold Hatzinger, Patrick Mair

Department of Statistics and Mathematics
Vienna University of Economics and BA
Augasse 2-6, 1090 Vienna, Austria
[reinhold.hatzinger, patrick.mair]@wu-wien.ac.at

Item response theory models (IRT) are growingly established in social science research, particularly in the analysis of performance or attitudinal data in psychology, education, medicine, marketing and other fields where testing is relevant. However, there is still a lack of computational implementations apart from commercially available special-purpose software (a comprehensive list is given at http://www.winsteps.com/rasch.htm). Several solitary routines have been published but there is no tool that allows for computing the various models in an integrative manner. The R package ltm (Rizopoulos, 2005) allows to fit some IRT models using a marginal ML approach but has a different focus and is restricted to binary data. We propose the R package eRm (extended Rasch modelling) for computing Rasch models and several extensions.

The main characteristic of IRT models, the Rasch model being the most prominent, concerns the separation of two kinds of parameters, one that describes qualities of the subject under investigation, the other relates to qualities of the situation under which the response of a subject is observed. Using conditional maximum likelihood (CML) estimation both types of parameters may be estimated independently from each other. IRT models are well suited to cope with dichotomous and polytomous responses, where the response categories may be unordered as well as ordered. The incorporation of linear structures allows for modelling the effects of covariates and enables the analysis of repeated categorical measurements.

Another aspect of Rasch type models is the concept of subject specific effects. If there is heterogeneity among subjects one might think of variance components and random effects parameters. In fact, this type of models using conditional ML estimation accounts for a different amount of individual propensity to certain reactions and can be seen as mixture models, where no specification of the distribution of the random effects parameters is made. These properties allow for the formulation of simple but very flexible models for longitudinal categorical data.

In a first version the eRm package fits Rasch models for binary and ordinal data, the rating scale model and the partial credit model. Linear reparameterisations through covariate structures allow for modelling diverse data sets including repeated measurements. We use an unitary, efficient CML approach with Newton-Raphsen and quasi-Newton algorithms to estimate the parameters. Graphical and numeric tools for assessing goodness-of-fit are provided.

Keywords:
Rasch Models, rating scale model, partial credit model, conditional ML estimation.

References:

G.H. Fischer and I. Molenaar, *Rasch Models - Foundations, Recent Developments, and Applications*, New York: Springer, 1995

systemfit: A Package to Estimate Simultaneous Equation Systems in R

Arne Henningsen and Jeff D. Hamann

Many theoretical models that are econometrically estimated consist of more than one equation. In this case, the disturbance terms of these equations are likely to be contemporaneously correlated, because some unconsidered factors that influence the disturbance term in one equation probably influence the disturbance terms in other equations of this model, too. Ignoring this contemporaneous correlation and estimating these equations separately leads to inefficient parameter estimates. However, estimating all equations simultaneously, taking the covariance structure of the residuals into account, leads to efficient estimates. This estimation procedure is generally called "Seemingly Unrelated Regression" (SUR) (Zellner, 1962). Another reason to estimate an equation system simultaneously are cross-equation parameter restrictions.[1] These restrictions can be tested and/or imposed only in a simultaneous estimation approach.

Furthermore, these models can contain variables that appear on the left-hand side in one equation and on the right-hand side of another equation. Ignoring the endogeneity of these variables can lead to inconsistent parameter estimates. This simultaneity bias can be circumvented by applying a "Two-Stage Least Squares" (2SLS) or "Three-Stage Least Squares" (3SLS) estimation of the equation system.

The **systemfit** package provides the capability to estimate linear equation systems in R (R Development Core Team, 2005). Although linear equation systems can be estimated with several other statistical and econometric software packages, **systemfit** has several advantages. First, all estimation procedures are publicly available in the source code. Second, the estimation algorithms can be easily modified to meet specific requirements. Third, the (advanced) user can control estimation details generally not available in other software packages by overriding reasonable defaults.

The **systemfit** package has been tested on a variety of datasets and has produced satisfactory for a few years. On the useR! conference, we would like to present some of the basic features of the **systemfit** package, some of the many details that can be controlled by the user, and the statistical tests for parameter restrictions and consistency of 3SLS estimation that are included in the package. While the **systemfit** package performs the basic fitting methods, more sophisticated methods are still missing. We hope to implement missing functionalities in the near future — maybe with the help of other useRs.

[1] Especially the economic theory suggests many cross-equation parameter restrictions (e.g. the symmetry restriction in demand models).

References

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Zellner A (1962). "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias." *Journal of the American Statistical Association*, **57**, 348–368.

Microeconomic Analysis with R

Arne Henningsen and Ott Toomet

Since its first public release in 1993, the free open source statistical language and development environment "R" (R Development Core Team, 2005) has been increasingly used for statistical analysis. While it has been prelevant in many scientific disciplines for a long time, it was not very widespread in economics in the first years. However, this situation has changed in recent years. Consequently, the number of extension packages for R that are suitable for economists has strongly increased in the last few years. One of these packages is called "**micEcon**" (Henningsen and Toomet, 2005) and provides tools for microeconomic analysis.

Initially, the **micEcon** package included only tools for microeconomic modeling. For example, it provides functions for demand analysis with the "Almost Ideal Demand System" (AIDS) (Deaton and Muellbauer, 1980). These functions enable the econometric estimation, calculation of demand (price and income/expenditure) elasticities and checks for theoretical consistency by one single R command. Second, **micEcon** contains tools for production analysis with the "Symmetric Normalized Quadratic" (SNQ) profit function (Diewert and Wales, 1987, 1992; Kohli, 1993). Additionally to the econometric estimation and calculation of price elasticities, it includes a function that imposes convexity on the estimated profit function using a new sophisticated method proposed by Koebel *et al.* (2003). Third, this package provides a convenient interface to "FRONTIER 4.1", Tim Coelli's software for stochastic frontier analysis (Coelli, 1996). Furthermore, **micEcon** includes tools for other functional forms, namely translog and quadratic functions.

About a year ago, we have added tools for sample selection models that are also often applied in microeconomic analyses. The **micEcon** package now includes functions to estimate these models using the two-step Heckman or an efficient maximum likelihood procedure. Furthermore, tools to calculate selectivity terms ("inverse Mill's ratios") even from bivariate probit models have been added.

On the useR! conference, we would like to present the capabilities of the **micEcon** package. Applied economists interested in microeconomic modeling will be invited to join our team and contribute to this package by providing tools for other types of microeconomic analyses.

References

Coelli TJ (1996). "A Guide to FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimation." CEPA Working Paper 96/7, Department of Econometrics, University of New England, Armidale NSW Australia.

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

Diewert WE, Wales TJ (1987). "Flexible Functional Forms and Global Curvature Conditions." *Econometrica*, **55**(1), 43–68.

Diewert WE, Wales TJ (1992). "Quadratic Spline Models for Producer's Supply and Demand Functions." *International Economic Review*, **33**, 705–722.

Henningsen A, Toomet O (2005). *micEcon: Tools for Microeconomic Analysis and Microeconomic Modeling*. R package version 0.2-1, http://cran.r-project.org.

Koebel B, Falk M, Laisney F (2003). "Imposing and Testing Curvature Conditions on a Box-Cox Cost Function." *Journal of Business and Economic Statistics*, **21**(2), 319–335.

Kohli UR (1993). "A symmetric normalized quadratic GNP function and the US demand for imports and supply of exports." *International Economic Review*, **34**(1), 243–255.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Resampling Libraries in S-PLUS and R

Tim Hesterberg <timh@insightful.com>
Research Department
Insightful Corp.

I'll give an introduction to the S+Resample library, and compare it to the S-PLUS/R boot library by Canty/Davison/Hinkley. While the S+Resample library has some technical capabilities no available in boot, its primary distinction is ease-of-use, both in simplifying user interaction from the command line, and in the addition of a GUI, intended for use even by introductory statistics students using the free student version of S-PLUS. In fact, the newest version of Moore and McCabe, Introduction to the Practice of Statistics (the most popular Intro Stat text in the U.S.), now has an optional resampling chapter, because of this project.

Conversely, the boot library has technical capabilities not in S+Resample. Unfortunately, the libraries are not compatible, which reflects past relationships between the S-PLUS and R communities. I conclude with a call for those relationships to change.

Least Angle Regression

Tim Hesterberg <timh@insightful.com>
Chris Fraley <fraley@insightful.com>
Research Department
Insightful Corp.

Least Angle Regression is a promising new technique for variable selection applications, offering a nice alternative to stepwise regression. It provides an explanation for the similar behavior of Lasso (L1-penalized regression) and forward stagewise regression, and provides a fast implementation of both. I'll describe work at Insightful to create an S-PLUS/R package, extending the existing "lars" package by Efron and Hastie. Extensions include the use of computationally-accurate methods, factor variables, and logistic regression. This also provides a model for other packages to work in both S-PLUS and R.

Keywords: regression, regularization, L1 penalty

Letter-Value Box Plots: Adjusting Box Plots for Large Data Sets

Heike Hofmann, Karen Kafadar, Hadley Wickham

January 26, 2006

Abstract

Conventional boxplots (Tukey 1977) are useful displays for conveying rough information about the central 50% of the data and the extent of the data. For moderate-sized data sets ($n < 1000$), detailed estimates of tail behavior beyond the quartiles may not be trustworthy, so the information provided by boxplots is appropriately somewhat vague beyond the quartiles, and the expected number of "outliers" and "far-out" values for a Gaussian sample of size n is often less than 10 (Hoaglin, Iglewicz, and Tukey 1986). Large data sets ($n \approx 10,000 - 100,000$) afford more precise estimates of quantiles in the tails beyond the quartiles and also can be expected to present a large number of "outliers" (about $0.4 + 0.007n$). The letter-value box plot addresses both these shortcomings: it conveys more detailed information in the tails using letter values, only out to the depths where the letter values are reliable estimates of their corresponding quantiles (corresponding to tail areas of roughly 2^{-i}); "outliers" are defined as a function of the most extreme letter value shown. All aspects shown on the letter-value boxplot are actual observations, thus remaining faithful to the principles that governed Tukey's original boxplot. We illustrate the letter-value boxplot with some actual examples that demonstrate their usefulness, particularly for large data sets.

Key words: boxplots, quantiles, letter value display, fourth, order statistic, tail area

Using R/Apache as the Statistical Engine for Web Applications

Jeffrey Horner

February 28, 2006

The R/Apache Project provides a solid foundation for web applications needing a statistical engine. With the R runtime engine embedded into the Apache 2.0 web server, and with the Apache and CGI data exposed to R functions, users can write applications entirely in the R language. The benefits to using R/Apache are:

- Users as statistical programmers can code in a language with which they are proficient,
- There is no need for a web enabled interface language to inflate process size,
- the cost of starting multiple R engines is hidden behind Apache's process management.

This presentation will demonstrate the R/Apache development process and explore strategies for session management, loading and storing data, dynamic graphics, and building web services. Current R/Apache limitations will also be discussed, such as supporting MS Windows and data persistence. In addition, the author will discuss research in R object serialization to backing stores with memcached and MySQL, and possibly a FastCGI interface to R.

The second international conference useR! 2006, Vienna, Austria, 15-17 June 2006

ROBUSTNESS ASSESSMENT FOR COMPOSITE INDICATORS WITH R

Luis Huergo[1], Ralf Münnich[2], Michaela Saisana[3]

[1]*University of Tübingen, Germany*, luis.huergo@uni-tuebingen.de

[2]*University of Trier, Germany*, ralf.muennich@uni-tuebingen.de

[3]*Joint Research Centre of the European Commission, Italy*, *michaela.saisana@jrc.it*

February 27, 2006

Composite indicators of countries performance are regularly used in benchmarking exercises or as policy-relevant interdisciplinary information tools in various fields such as economy, health, education or environment. They are calculated as weighted combinations of selected indicators via underlying models of the relevant policy domains. Yet, composite indicators can equally often stir controversies about the unavoidable subjectivity that is inherent in their construction. To this end, it is important that their sensitivity to the methodological assumptions be adequately tested to ensure that their methodology is sound and not susceptible to bias or significant sensitivity arising from data treatment, data quality [1,2], aggregation, or weighting [3,4].

In this presentation we use a combination of uncertainty and sensitivity analysis, coded in R programming language, to study how variations in the country scores derive from different sources of variation in the assumptions entailed in a composite indicator measuring the Knowledge Economy in the European Union. We focus on four major sources of uncertainty: (i) variation in the indicators' values due to imputation of missing data [5, 6], (ii) selection of weights, (iii) aggregation method (linear or geometric), and (iv) exclusion of one indicator at-a-time. The "UASA package" for the R statistical environment implements global variance-based sensitivity analysis for non-correlated input.

The aim of the analysis is to help gauge the robustness of the composite indicator scores, to increase the transparency in the development of the composite indicator, to identify the countries that improve or decline under certain methodological assumptions, and to help frame the debate around the use of such Index.

The discussion of the results is extended to re-assess the usefulness of combined sensitivity and uncertainty analysis to make quality judgments of composite indicators and how it can be formalized into a broader quality framework for knowledge performance measures and control policies.

The work is part of the project KEI (Knowledge Economy Indicators; cf. http://kei.publicstatistics.net) which is financially supported by the European Commission within the 6th Framework Programme under policy orientated research.

More information on composite indicators can be found at: http://farmweb.jrc.cec.eu.int/ci

The second international conference useR! 2006, Vienna, Austria, 15-17 June 2006

References

[1] Davison, A. C.,Münnich, R., Skinner, C. J., Knottnerus, P. und Ollila, P. (2004): The DACSEIS Recommended Practice Manual. DACSEIS deliverable D12.3, http://www.dacseis.de.

[2] Magg, K., Münnich, R., Wiegert, R., Åkerblom, M., Schmidt. K. (2006): Quality Concepts – State-of-the-Art. KEI deliverable 3.1. To appear.

[3] Saisana M., Saltelli A, Tarantola S. (2005): Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators, Journal of the Royal Statistical Society A, 168(2), 307-323.

[4] Saltelli A. Tarantola S., Campolongo, F. and Ratto, M.(2004): Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models, John Wiley & Sons publishers.

[5] Laaksonen, S., Rässler, S., Skinner, C., Oetliker, U. und Renfer, J.-P. (2004): Imputation and Non-response. DACSEIS deliverable D11.2, http://www.dacseis.de.

[6] Rubin, D.; Little, R. (2002): Statistical Analysis with Missing Data, Wiley & Sons.

SensoMineR: a package for sensory data analysis with R

F. Husson & S. Lê*
Agrocampus, 65 rue de St Brieuc, CS 84215, 35042 Rennes Cedex, France
sebastien.le@agrocampus-rennes.fr fax 02 23 48 58 93 phone 02 23 48 58 81

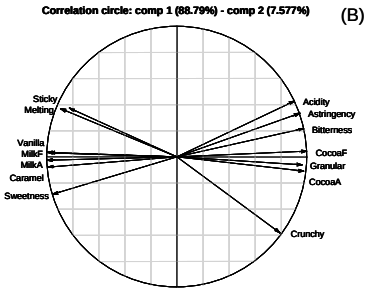
Sensory analysis is both a very lively and competitive field as can testify congresses such as Pangborn or Sensometrics. Consequently, users or practitioners can be disconcerted with the increasing number of methods at their disposal. We propose a new package, the SensoMineR package, conceived and programmed in R language; SensoMineR is completely free and can be downloaded at the following address: http://sensominer.free.fr. SensoMineR collects very classic methods used when analyzing sensory data as well as methods developed in our laboratory. SensoMineR provides numerous graphical outputs easy to interpret, as well as syntheses of results issuing from various analysis of variance models or from various factor analysis methods accompanied with confidence ellipses. SensoMineR tackles the following problems: characterizing products, relating sensory data and instrumental data, mapping consumers' preferences, assessing panel's performances, comparing panels' performances. During this presentation we will focus on the characterization of products and we will present two functions, the "decat" function or how to get unidimensional profiles of products synthesized in a single table, the "panellipse" function or how to get multidimensional profiles of products with confidence ellipses obtained by resampling techniques (bootstrap). The example that will be presented is provided with the package.

(A)

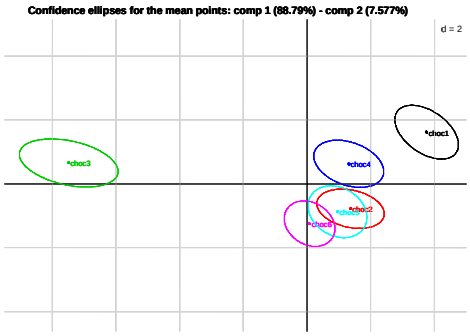
Average by chocolate

	CocoaF	Bitterness	CocoaA	Granular	Astringency	Acidity	Crunchy	Sticky	Melting	Sweetness	Vanilla	Caramel	MilkA	MilkF
choc1	8.069	7.069	7.086	3.448	4.759	4.655	5.966	3.759	4.741	3.138	1.103	1.672	3.586	1.569
choc2	6.914	4.948	6.552	3.155	3.155	3.138	7.707	3.828	4.328	4.621	1.81	2.776	4	2.379
choc4	6.69	5.19	6.259	3.552	3.69	3.931	6.103	4.103	4.379	4.293	2.121	2.672	4.103	2.586
choc5	6.793	4.879	6.793	3.069	3.103	3.086	6.638	3.224	4.741	5.224	1.793	3.414	4.172	3.121
choc6	6.224	4.19	6.362	3.172	2.759	2.672	7.328	3.931	4.207	5.621	1.914	3.259	4.569	3.362
choc3	3.379	1.397	4.672	1.603	1.207	1.569	2.983	5.034	7.31	7.603	3.672	6.328	6.052	7.707

(B)



(C)



(D)

P-values for the Hotelling's T2 tests

	choc1	choc2	choc3	choc4	choc5	choc6
choc1	1	4.199e-12	2.213e-24	5.377e-07	8.1e-13	2.319e-17
choc2	4.199e-12	1	7.884e-17	0.0008818	0.6708	0.00759
choc3	2.213e-24	7.884e-17	1	1.912e-17	1.599e-16	1.934e-15
choc4	5.377e-07	0.0008818	1.912e-17	1	0.0007292	3.787e-07
choc5	8.1e-13	0.6708	1.599e-16	0.0007292	1	0.06995
choc6	2.319e-17	0.00759	1.934e-15	3.787e-07	0.06995	1

Figure (A) shows the structure on the average data table: cells are colored if the mean of the product is significantly different from the overall mean (blue, if it is higher and pink if it is lower). Figure (C) represents the space products with confidence ellipses around the products. Figure (D) gives the p-values associated with the T square Hotelling test in order to test differences between sensory profiles of two products.

Automatic time series forecasting

Rob J Hyndman
Monash University, Australia

Automatic forecasts of large numbers of univariate time series are often needed in business. It is common to have over one thousand product lines that need forecasting at least monthly. In these circumstances, an automatic forecasting algorithm is an essential tool. Automatic forecasting algorithms must determine an appropriate time series model, estimate the parameters and compute the forecasts. The most popular automatic forecasting algorithms are based on either exponential smoothing or ARIMA models.

Exponential smoothing

Although exponential smoothing methods have been around since the 1950s, a modelling framework incorporating procedures for model selection was not developed until relatively recently with the work of Ord et al. (1997) and Hyndman et al. (2002). In these (and other) papers, a class of state space models which underly all of the exponential smoothing methods has been developed. Exponential smoothing methods were originally classified by Pegels' (1969) taxonomy. This was later extended by Gardner (1985), modified by Hyndman et al. (2002), and extended again by Taylor (2003), giving a total of fifteen methods seen in the following table.

	Trend Component	**Seasonal Component** N (None)	A (Additive)	M (Multiplicative)
N	(None)	N,N	N,A	N,M
A	(Additive)	A,N	A,A	A,M
A_d	(Additive damped)	A_d,N	A_d,A	A_d,M
M	(Multiplicative)	M,N	M,A	M,M
M_d	(Multiplicative damped)	M_d,N	M_d,A	M_d,M

Some of these methods are better known under other names. For example, cell N,N describes the simple exponential smoothing (or SES) method, cell A,N describes Holt's linear method, and cell A_d,N describes the damped trend method. The additive Holt-Winters' method is given by cell A,A and the multiplicative Holt-Winters' method is given by cell A,M. The other cells correspond to less commonly used but analogous methods.

Hyndman et al. (2002) describes how each exponential smoothing method corresponds to two state space models, giving 30 models in total. They also discuss an automatic algorithm for identifying an appropriate exponential smoothing model in a general class of state space models. I will review an implementation of the Hyndman et al. (2002) algorithm in the `forecast` package for R.

ARIMA models

Automatic ARIMA modelling has a longer pedigree, but is not widely used because of the computational time involved. To my knowledge, no automatic ARIMA algorithms are currently available in existing R packages. Furthermore, existing automatic ARIMA methods are based on information criteria which aim to identify the "correct" model rather than find a good forecasting model. I will review the main existing approaches and describe a new algorithm for automatic ARIMA forecasting where the aim is to "select the model that produces the best forecast" rather than "calculate forecasts from the best model". Finally, I will discuss the implementation of this new algorithm in R.

References

Gardner, Jr, E. S. (1985) Exponential smoothing: The state of the art, *Journal of Forecasting*, **4**, 1–28.

Hyndman, R. J., A. B. Koehler, R. D. Snyder and S. Grose (2002) A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, **18**(3), 439–454.

Ord, J. K., A. B. Koehler and R. D. Snyder (1997) Estimation and prediction for a class of dynamic nonlinear statistical models, *Journal of the American Statistical Association*, **92**, 1621–1629.

Pegels, C. C. (1969) Exponential forecasting: Some new variations, *Management Science*, **15**(5), 311–315.

Taylor, J. W. (2003) Exponential smoothing with a damped multiplicative trend, *International Journal of Forecasting*, **19**, 715–725.

Iterated function system and simulation of Brownian motion

Stefano Iacus and Davide La Torre

Several methods are currently available to simulate paths of the Brownian motion. In particular, paths of the BM can be simulated using the properties of the increments of the process like in the Euler scheme, or as the limit of a random walk or via L2 decomposition like the Kac-Siegert/Karnounen-Loeve series.

In Iacus and La Torre (2006, see http://arxiv.org/abs/math.PR/0601379) a IFSM (Iterated Function Systems with Maps) operator whose fixed point is the trajectory of the BM is proposed. In that paper we studied the application of this representation of stochastic processes to simulate their trajectories. The resulting simulated trajectories are self-affine, continuous and fractal by construction.

This fact produces more realistic trajectories than other schemes in the sense that their geometry is closer to the one of the true BM's trajectories. The IFSM trajectory of the BM can then be used to generate more realistic solutions of stochastic differential equations.

Pathwise approximations of stochastic processes remain a relevant topic, for example, in computational finance and numerical option pricing. In this work we will discuss some advances on this topic and present the new version of the IFS package.

R on Different Platforms: The useRs' Point of View

Stefano Iacus Uwe Ligges Simon Urbanek

R is available and widely used on various platforms, most notably under the operating systems Linux, Mac OS, and Windows. In our talk we will focus on (surprising?) analogousness but also differences between these operating systems. We try to answer some of the useR's questions on her/his way to become a developer:

- Where is the graphical user interface?
- What is the "right" editor to use?
- What is the "right" way to install and manage packages?
- How can I write platform independent code?
- How can I build my own package on my platorm / for other platforms?
- Which features are platform specific?

We will present some answers and examples for these questions and point to the appropriate documentation. UseRs will see that installing packages on Windows is not mysterious, Linux is great to run R jobs with low priority in the background, and R for Mac OS X already runs smoothly on Intel based Macs.

Matching and ATT Estimation via Random Recursive Partitioning

Stefano Iacus and Guiseppe Porro

Average Treatment Effect estimation is the problem of measuring the difference in outcome of two groups of treated and control units. In contexts where randomized experiment are difficult to be carried out (like in econometrics, sociology, political analysis, etc), the experimenter is only able to select observations to be to exposed a treatment and subsequently search for a control group from other data sources. In such observational studies, to reduce the bias due to different pre-treatment conditions between treated and control units, it is usually applied some matching method.

Due to the curse of dimensionality, direct matching on the space of covariates X is usually hard to carry out. Approaches based on different notions of similarity have been developed, like propensity score methods and distance based methods. In these cases, matching is based on a one-dimensional quantity instead of looking for similar individuals in the k-dimensional space X.

In this talk we present the Random Recursive Partitioning (RRP) method, which operates directly on the space of covariates X. The method randomly generates non-empty recursive partitions of the space X and evaluates how frequently two individuals lie in the same subset of X. RRP is indeed a Monte Carlo method on the set of possible partitions of X for estimating the likelihood of the "proximity" of two individuals. We also propose some tools to measure the extent of the common support between treated and control units. No average treatment effect estimation can in fact carried out without a sufficient overlap between the two groups.

Applications to real and simulated data and corresponding R code is presented.

A Unified User-Interface for Single and Multi-Equation Models (Also Known As, "Zelig: Everyone's Statistical Software")[1]

Kosuke Imai[2] *Gary King*[3] *Olivia Lau*[4]

Abstract

We propose a common framework for statistical analysis and software development built on and within the R language. Researchers in different academic disciplines have developed different statistical models, different mathematical notation, different parameterizations, different quantities of interest, and hence different computational implementations. The users of statistical software have much to gain by navigating the babel of R's many packages, but is is often far more difficult than it should be, given that packages have so much underlying statistical theory and computational structure in common. To address this problem, we have developed a conceptual framework and software package that offers:

- A common syntax for specifying univariate response, multivariate response, multilevel, and hierarchical models. In particular, we offer a user-specified, intuitive interface for identifying constraints across equations.
- A method to use this syntax with existing R packages, without modifying those packages (by re-defining function calls on the fly).
- A method to translate this syntax into matrices and arrays useful for programmers. For multiple equation models, we offer three implementation options (all of which work with user-specified constraints): an intuitive option that stacks matrices visually, a computationally efficient option that creates arrays of explanatory variables, and a memory-efficient option that coerces parameters to matrices.
- A framework for calculating quantities of interest with or without conditioning on the observed values of a particular unit, to provide users with substantive interpretation of model output.
- Methods to process lists of multiply imputed data, or stratified data, and combine model estimates when generating quantities of interest.
- An application programmer interface that makes it possible to dynamically generate a graphical user interface (GUI) for the models included in Zelig (see, e.g., the Virtual Data Center, for one example of a GUI which has already been implemented).

Our approach is intended to bring the computational flexibility and power of R to users who need to understand the data, the modeling assumptions, and the quantities of interest, but are not as focused on computational algorithms as applied statisticians. By creating a common interface between users and developers, and a set of tools to operate on that interface, we hope to extend the range of models that one can write and use in R, and the range of people to whom R is accessible.

[1]Our thanks to the National Institutes of Aging (P01 AG17625-01), the National Science Foundation (SES-0318275, IIS-9874747), and the Mexican Ministry of Health for research support. Current software is available from CRAN or `http://gking.harvard.edu/zelig/`.

[2]Assistant Professor, Department of Politics, Princeton University, `kimai@princeton.edu`

[3]David Florence Professor of Government, Department of Government, Harvard University, `king@harvard.edu`

[4]Ph.D. Candidate, Department of Government, and M.A. Student, Department of Statistics, Harvard University, `olau@fas.harvard.edu`

SEQUENTIAL MONTE CARLO METHODS IN R

THOMAS JAKOBSEN
JEFFREY TODD LINS

SAXO BANK A/S

Sequential Monte Carlo (SMC) methods, also known as particle filters, are an efficient means for tracking and forecasting dynamical systems subject to both process and observation noise. Applications include robot tracking, video or audio analysis, and general time series analysis.

Whereas the traditional Kalman filter is the optimal way of solving the tracking problem for linear, Gaussian models, other techniques such as SMC methods are needed in the nonlinear or non-Gaussian case. SMC methods maintain a set of particles to represent the posterior at time $t-1$ and then updates the weights of these particles to time step t by taking into account the observation y_t. To avoid an adverse increase in the variance of the importance weights, resampling steps are then usually carried out at regular intervals.

R's functionality provides an excellent basis for the development of flexible and efficient particle filter algorithms, and we show that it is possible to implement a number of interesting time series models, e.g., stochastic volatility models and adaptive factor models. We then demonstrate their use through examples involving real-world, financial, multivariate time series, where particle filter algorithms are used to track underlying factors. We also briefly describe how to use R's built-in optimization procedures to estimate fixed parameters.

We hope to include the developed algorithms and models in a forthcoming R package which will allow the user, through a simple interface, to choose between different approaches, e.g. sampling importance resampling (SIR), auxiliary particle filters (APF) and various resampling strategies.

References

[Arulampalam et al., 2002] Arulampalam, S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188.

[Doucet et al., 2001] Doucet, A., de Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Quantitative Analysis, Saxo Bank A/S, Smakkedalen 2, 2820 Gentofte, Denmark

E-mail address: tja@saxobank.com,jtl@saxobank.com

Date: February 27, 2006.

Computation and Aggregation of Quantiles from Data Streams

John M. Chambers David A. James Diane Lambert
Scott Vander Wiel

We describe the Incremental Quantile (IQ) method and its R implementation. IQ is a quick-and-dirty quantile tracker that we developed for monitoring network applications, and in particular it addresses two problems: (1) how to estimate, under strict computational requirements, multiple quantiles on a very large number of data streams or "agents" (e.g., groups of IP/port/time combinations), and (2) how to combine these agent quantiles to estimate quantiles of arbitrary data aggregates without re-processing the raw data stream.

The IQ method is attractive for its simplicity: As new data become available, IQ updates the current quantile estimates from the empirical CDF (eCDF) obtained by (weighted) averaging the current quantile estimates' eCDF and the new data eCDF. On the other hand, careful selection of IQ's probabilities and buffer sizes are needed for adequate performance.

We close by showing that IQ outperforms the previously implemented algorithm, but we also show through simulations the high price we pay in terms of root mean squared error for the scant use of computing resources.

Using the R language for graphical presentation and interpretation of compositional data in mineralogy – introducing the package *GCDkit-Mineral*

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

[1] *Institute of Petrology and Structural Geology, Charles University, Albertov 6, 128 43 Prague 2, Czech Republic; janousek@cgu.cz*

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

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

One of the problems we are facing in mineralogy is the dearth of universal, flexible and inexpensive software for recalculation of large compositional data sets acquired by microbeam techniques. The aim is to recast the chemical analyses from individual grains of various minerals into numbers of atoms per structural formula unit, and classify the individual data points according to rather complex rules based mostly on binary or ternary diagrams. The problem is that the recalculation and classification schemes differ strikingly for each of the main mineral species (IMA 2006).

We have decided to tackle the problem using the R language, as it provides powerful tools for data import/export, handling data matrices and production of publication quality graphics. An elegant solution to the computational problem uses S4 classes (Chambers 1998) to define algorithms as methods for each of the mineral species separately.

The raw data are imported into the system from the clipboard, text files or using the *RODBC* package (Ripley 2005). Individual analyses are split into classes according to the mineral species they belong to which enables the recalculation schemes to be defined, some of them rather complex, as independent S4 methods. There is a set of several generic functions that load the recalculation options for the given mineral class from two small and lucid external database files (ASCII and XML formats) that can be edited by users without prior R experience. Using these functions, the analyses are recast to structural formulae and the atoms are assigned to appropriate crystallographic sites. The minerals frequently form solid solutions and in these cases the data are transformed into a combination of two or more end-member compositions. Finally user-defined subsets of the numeric results can be copied to the clipboard, or exported via *RODBC* and *R2HTML* packages (Lecoutre 2003; Ripley 2005) to several formats (XLS, MDB, DBF and HTML). Special attention has been paid to provide routines for effortless data management, i.e. searching and generation of subsets, using regular expressions and Boolean logic.

In addition, our package for handling mineral compositions contains flexible high-level graphical functions. The diagrams are defined as internal templates that provide a means to create figure objects. The objects 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…). Most importantly, the templates are used as a basis for classification. Taking advantage of the algorithms originally developed for spatial data analysis (package sp of Pebsma & Bivand 2005), our general routine looks for the name of the polygon within the diagram (= graphical template), into which the analysis falls according to its x–y coordinates. The outcome can be either a name of the mineral or a link to another diagram, in the case of the more complex classification schemes. Following the compulsory rules of the International Mineralogical Association (IMA), in some cases the classification is not done graphically, but using external functions.

The package, named *GCDkit-Mineral (GCDkit* standing for 'Geochemical Data Toolkit'), is a part of a broader family of tools designed for mineralogists and igneous geochemists. The overall philosophy each of the packages is to a large extent similar. All their functions are accessible via graphical user interface, as well as in an interactive regime for R-literate colleagues. The core of the system, e.g. the data input/output, data management and graphical functions is identical in each case. The *GCDkit* family tools are distributed as freeware via the WWW; the current version can be downloaded from *http://www.gla.ac.uk/gcdkit.*

Chambers, J. M. (1998). *Programming with Data*. New York: Springer.
IMA (2006). Accessed on February 28, 2006 at http://www.minsocam.org/MSA/IMA/.
Lecoutre, E. (2003). The R2HTML package. *R-news* **3 ,** 33–36.
Pebsma, E. J. & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R-news* **5,** 9–13.
Originally Michael Lapsley and since Oct 2002 B. D. Ripley (2005). RODBC: ODBC database access. R package version 1.1-4.

pcalg: Estimating and visualizing high-dimensional dependence structures using the PC-algorithm

Markus Kalisch

February 22, 2006

We consider the PC-algorithm (Spirtes, Glymour and Scheines (2000)) for estimating the skeleton of a potentially very high-dimensional acyclic directed graph (DAG) with corresponding Gaussian distribution. The PC-algorithm is computationally feasible for sparse problems with many nodes, i.e. variables, and it has the attractive property to automatically achieve high computational efficiency as a function of sparseness of the true underlying DAG.

The restriction of underlying Gaussian distribution can be relaxed by using a robust scale estimator with high precision and high breakdown point (Q_n estimator).

We provide theoretical consistency results on this algorithm and analyze its properties in simulations. Furthermore, we introduce the new R-package `pcalg`, which performs both the standard and the robust version of the discussed algorithm. This package was used to obtain all simulation results.

The S Package System

Stephen Kaluzny
spk@insightful.com
Insiightful Corp.

The next release of S-PLUS will include a package system not unlike the package system in R. A goal of the system is the creation of source packages that can be installed both in R and S-PLUS. To achieve this goal many new low level functions often used by R packages have been added to S-PLUS and some existing functions now have additional arguments for R compatibility. An R.h include file allows most R package source code to be compiled under S-PLUS. Insightful developers have been working with several R package authors on converting their packages to run under S-PLUS. I will discuss the conversion process and describe areas that require further work. I will also show our package compatibility checker that examines a package source tree, identifying and possibly correcting non-portable areas.

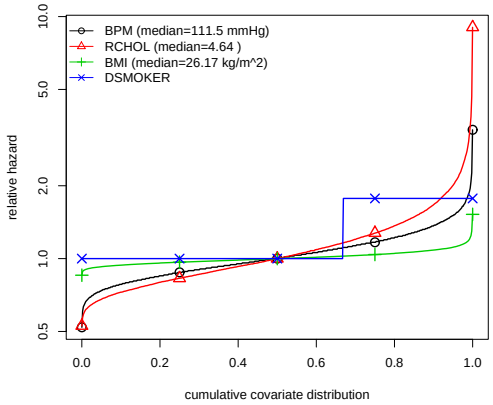
Visualizing covariates in proportional hazards model using R

Juha Karvanen

Consider an example where the hazard of coronary heart disease (CHD) is modeled by Cox's proportional hazards model with covariates cholesterol ratio, blood pressure (mmHg), body mass index (BMI, kg/m^2) and smoking (0/1). The estimated hazard ratios and their 95% confidence intervals are: cholesterol ratio 1.237 (1.200,1275), blood pressure 1.016 (1.012,1.021), BMI 1.015 (0.999,1.032), smoking 1.773 (1.572,1.999). We are interested in assessing the importance of these covariates in the population, e.g. answer questions such as "is smoking more serious risk factor of CHD than overweight in the population level?" The answer depends not only on the model coefficients and their statistical significance but also on the prevalence of smoking and overweight in the population.

We present a graphical method that visualizes the relative importance of covariates in a proportional hazards model. The key idea is to rank the covariates and plot the relative hazard as a function of ranks scaled to interval $[0, 1]$. This allows plotting of covariates measured in different units in the same graph. The reference hazard for relative hazards can be fixed to correspond to the cohort medians or some predefined reference values, e.g. BMI value 25. The method can be also utilized when comparing models. For instance, we may visualize the effect of taking logarithms of the covariates.

The visualization is implemented using R. An illustration is shown below.



1

Applied Econometrics with R

Christian Kleiber
Dept. of Statistics and Econometrics
Universität Basel, Switzerland
Christian.Kleiber@unibas.ch
and
Achim Zeileis
Dept. of Statistics and Mathematics
Wirtschaftsuniversität Wien, Austria
Achim.Zeileis@wu-wien.ac.at

Empirical research in economics commonly utilizes programming languages such as GAUSS or, more recently, Ox, as well as packages of canned routines such as EViews, RATS or TSP, to mention a few. The authors believe that R, with its flexibility, object orientation and superior graphics, has great potential in econometrics.
However, there appear to be at least two obstacles: First, R is being developed from the point of view of mainstream statistics, often using terminology unfamiliar to many econometricians (e.g., generalized linear models). This creates the impression, as may be witnessed on R-help, that some methods are unavailable in R, while in fact they just appear under different names. Second, classical statistics stems from the analysis of randomized experiments, while observational data are the rule in economic applications. This implies that modifications of classical procedures are required, some of which still need implementation in R.
The talk will report on the current status of Kleiber and Zeileis (2006), the first book on R exclusively devoted to econometric applications. Some gaps in R will be identified, notably models for panel data and some branches of microeconometrics.

Kleiber, C., and Zeileis, A. (2006). *Applied Econometrics with R*. New York: Springer, forthcoming.

Visualization of multivariate functions, sets, and data with package "denpro".

Jussi Klemelä
Department of Statistics, University of Mannheim

Package "denpro" implements several methods for visualizing multidimensional objects. The package is specialized for the visualization of density functions and density estimates, and multivariate continuous data. One of the basic ideas is to transform multivariate objects to low dimensional objects so that certain important shape characteristics are preserved.

Figure 1 visualizes a 2D Clayton copula density whose paramater is 4, and which has Student marginals with degrees of freedom 4. Frame a) shows a contour plot of the density, frame b) shows a tail probability plot of the 0.005% level set of the density (level 0.001742), and frames c) and d) show a location plot of this level set. The contour plot is defined only for 2D densities but the other visualizations may be used in higher dimensional cases.

Figure 2 visualizes a data of size $n = 1000$ generated from the density in Figure 1a. Frame a) shows a scatter plot of the data, frame b) shows a tail frequence plot of the data, and frames c) and d) show a tail tree plot of the data. Again, the scatter plot is defined only for 2D data but the other visualizations may be used also in higher dimensional cases.

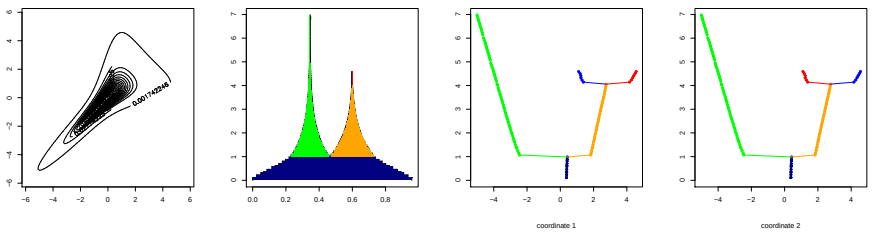


Figure 1: Visualization of a level set of a Clayton copula density which has Student marginals.

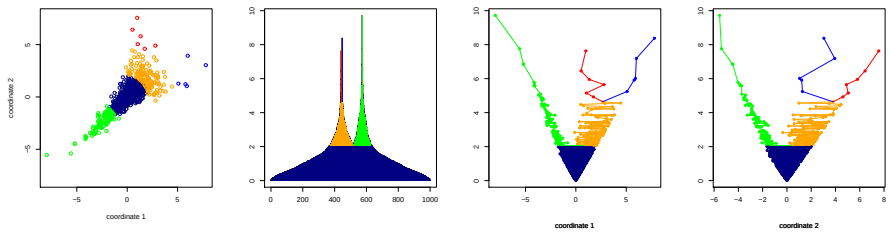


Figure 2: Visualization of data generated from the density in Figure 1a.

Integration of R into Wikis [1]

Sigbert Klinke[1], Sibylle Schmerbach[2], Olga Troitschanskaia[3]

Humboldt-Universität zu Berlin
[1]Wirtschaftswissenschaftliche Fakultät, Institut für Statistik und Ökonometrie
[2]Wirtschaftswissenschaftliche Fakultät, Institut für Öffentliche Finanzen, Wettbewerb und Institutionen
[3]Philosophische Fakultät IV, Institut für Erziehungswissenschaften, Abt. Wirtschaftspädagogik

Wikis for Teaching. The creation of the master and bachelor programs will increase the mobility of students within europe. Therefore we can not expect in the masters programs in economics or pedagogics that the statistics knowledge of the students is equivalent to the knowledge of the students which have visited the bachelor program at our faculty.

We utilize the Wiki technology to build up two wikis: (1) StatWiki, a wiki as an interactive statistical dictionary (in german) and (2) TeachWiki, a wiki for final theses in statistics of the students in our statistics lectures (in english). The StatWiki is currently based on Zope, Zwiki and LatexWiki and the TeachWiki is based on the MediaWiki which used for the Wikipedia.

Use of R. For statistics we need to integrate graphics and tables. We decided to utilize the programming language R and developed for Zope and the MediaWiki plug-ins which allow us to integrate R programs. The R plug-in allows to integrate raw R output (e.g. `<R output="text">1:10</R>`), HTML output generated by our `outHTML` command in R for tables and multi-page graphics into the wiki.

Graphical output is written as PostScript to a file and the ImageMagick package is used to convert the graphics to JPEG. Further work may include the utilization of forms to allow for further interactivity in the MediaWiki.

Weblinks
StatWiki - an interactive statistical dictionary for teaching:
`http://statwiki.wiwi.hu-berlin.de`
TeachWiki - a wiki for final theses in statistics lectures:
`http://stirner.wiwi.hu-berlin.de/mediawiki/teachwiki`
Zope - a open source content managment system: `http://www.zope.org`
Zwiki - a wiki plugin for Zope: `http://www.zwiki.org`
LatexWiki - a latex plugin for ZWiki: `http://mcelrath.org/Notes/LatexWiki`
MediaWiki - the wiki of the Wikipedia: `http://www.mediawiki.org`
Wikipedia - the free encyclopedia: `http://www.wikipedia.org`

[1]This work is supported by the Multimedia-Förderprogramm of the Humboldt-Universität zu Berlin and the CRC 649: Economic risk of the Deutsche Forschungsgemeinschaft.

Parametric link functions for binary response models: A Fisherian Holiday

Roger Koenker

The familiar logit and probit models provide convenient settings for many binary response applications, but a larger class of link functions may be occasionally desirable. Two parametric families of link functions are suggested: the Gossett link based on the Student t latent variable model with the degrees of freedom parameter controlling the tail behavior, and the Pregibon link based on the (generalized) Tukey λ family with two shape parameters controlling skewness and tail behavior. Implementation of the methods is illustrated in the R environment.

LabNetAnalysis - An instrument for the analysis of data from laboratory networks based on RExcel

Andrea Konnert

Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany
Department of Biostatistics, Roche Diagnostics GmbH, 82377 Penzberg, Germany

Abstract. For the definition of reference standards of diagnostic assays, laboratory networks are founded to ensure reliable patient results over space and time. An example of such a network is the IFCC HbA1c standardization network, where 10 - 13 laboratories define the reference values for the HbA1c measurements worldwide. Up to 15 samples are measured within each of these laboratories, the values of each sample are combined to define the reference concentration of this sample.
After the laboratories reported their values to the network coordinator, the following points need to be analyzed:

(i) Analysis of the measurements of an individual laboratory in comparison to the other laboratories.

(ii) Detection of outliers within the data.

(iii) Calculation of the uncertainties of the reference concentrations.

(iv) Comparison of the reference concentrations of one network with other national networks.

To be able to perform all these analysis by the network coordinator in a routine-like fashion, an instrument was needed that enables easy data-handling and storing, user-friendly operation and all statistical power to perform the analysis, which comprises basic statistics, mixed models and MCMC simulations. The interface between R - for the statistical analysis and Excel - for the data handling, storing and user-friendly interaction, provided by RExcel was found to be ideal for these requirements.
The talk will shortly describe the statistical analysis, afterwards the functioning of LabNetAnalysis based on RExcel will be shown, and end with a short story of the development process focusing on highlights and pitfalls during the development.

References

Konnert A, Berding C, Arends S et al., Statistical Rules for Laboratory Networks, JTEV, in press 03/06

Konnert A, Arends S, Schubert S et al., Uncertainty Calculation for Calibrators of the IFCC HbA1c Standardization Network, ACQUAL, in press 03/06

Baier Th, Neuwirth E, RExcel - using R from within Excel V1.5, http://sunsite.univie.ac.at/rcom/download/

Keywords

RExel, LABORATORY NETWORKS, UNCERTAINTY CALCULATION

Katarzyna Kopczewska*, MA
Faculty of Economic Sciences
Warsaw University

GEOGRAPHICAL BENEFITS IN SOCIO-ECONOMICS DEVELOPMENT IN POST-SOCIALIST COUNTRIES

ABSTRACT

Following the fall of the iron curtain in 1989 the post-socialist countries began to bridge the socio-economic gap between them and the Western European countries. The regions situated on the borderlines of the real socialism and capitalist economies became a bridge for integration of this part of Europe. In this research I will try to answer the question whether the countries situated along the iron curtain have benefited from their location and developed faster after 1989 than other regions. Purely economic factors, economic growth rate of Poland, the Czech Republic, Slovakia, and Hungary compared to their western neighbours Germany and Austria were subject to analysis as well as the changes in economic consciousness, cultural and social factors of these regions. This allowed for determining of beneficiaries and losers of the fall of the iron curtain, and also for examining the social aspects of development convergence and diffusion. Spatial error and spatial lag models with appropriate tests, correlogram, variogram and spatial statistics have been used for the purpose of the econometric analysis. The research was conducted on EUROSTAT and EVS data. Calculations were made in R with *spdep* and *geoR* packages.

The results indicate that, contrary to all expectations, the border regions of the post-socialist countries were developing slower whereas Austria's and Germany's border regions became the beneficiaries of the changes. This means that the post-socialist countries did not put emphasis on incorporating the border regions and did not redirect their international activities westwards. Germany and Austria on the contrary have grasped their chance on activating their eastern regions. During the economic growth of 1995-2000, geographical benefits measured with the GDP growth rate amounted to -10% for the eastern countries' border regions and +8% for the western countries' border regions.

Key words: spatial statistics, geographical benefits, post-socialist countries, socio-economic development, spatial models, correlogram, variogram

JEL classification: C31, F15, O18, O52, R12, Z13

* kkopczewska@wne.uw.edu.pl, WNE UW, ul.Dluga 44/50, 00-241 Warsaw, Poland

Cancer research – R package to analyze genomic regulation and tumor pathways based on array data from single nucleotide polymorphism (SNP) and comparative genomic hybridization (CGH) experiments

E. KORSCHING, W. NADLER[1], H. BÜRGER
Institute of Pathology, University of Münster (korschi@uni-muenster.de)
[1]John von Neumann Institute for Computing, Research Centre Jülich

Aims: Cancer research is focused on a better understanding of tumor progression pathways. Progression means that there are successive development stages of a certain tumor type which often can be discriminated by means of molecular biology methods or classically by morphologic classification of an ultra thin tissue section. This knowledge on types and grades in turn helps to understand and identify early events which makes it easier to intercept the progression on a level where the therapeutic action is feasible. In this context it is important to decipher marker molecules for diagnostic and also therapeutic purpose. Marker means that some molecular factors play a more specific role in a certain tumor event than other involved molecules.
One of the approaches to get a better insight in tumor development is to analyze genomic events by means of the CGH / matrix CGH and more recent by SNP analysis.
The SNP analysis on the whole human genome with a very dense coverage gives the possibility to go beyond the limits of resolution of the classical CGH and also the matrix CGH. We are trying to explore the new details arising with this high resolution and established a set of algorithms supporting this analysis approach.

Implementation: The set of algorithms was originally developed in S because the platform SPlus has an excellent visual data browser very useful in big and complex projects and also in routine work. But a lot of problems concerning the closed source platform their evolvement away from science and also some technical limitations convinced us to switch to R with our activities.
The functionality of the R code cover the following tasks: a) show the relation between classical CGH and SNP experiments b) associate SNP and gene expression data to analyze regulatory motives and c) visualize results in an appropriate full genome or chromosome mode. The implementation is actually based on the CSV format of the exported Affymetrix data, but can be easily adapted to other data structures.
The introduction of the 10K SNP microarray by Affymetrix has brought an average resolution of 210 kb up to 6 kb with the 500K SNP chip set. The developed software tools allow on one hand a rapid, global overview of all unbalanced chromosomal alterations within a tumor by the use of a smoothing procedure. On the other hand our tools allow a much more detailed view into the fine structure of chromosomal alterations giving a much better insight into the complex picture of chromosomal alterations.

Conclusion: The developed analysis tools are a stable platform to promote new models in genomic deregulation occurring in cancer events.

Reference: *Improvements in the analysis strategy make single nucleotide polymorphism (SNP) analysis a powerful tool in the detection and characterization of amplified chromosomal regions in human tumors. Korsching E. et al., Pathobiology 2006, in press.*

Calibrating the evidence in experiments with applications to meta-analysis

Elena Kulinskaya,*Stephan Morgenthaler†and Robert G. Staudte‡

June 4, 2006

1 Introduction

How much more evidence is there in a 'highly significant' p-value of 0.01 relative to one 'just significant' at 0.05? Why does the replication of an experiment lead, on average, to a higher p-value than the one just obtained? To answer such questions one must go beyond the traditional p-value which is conditional on the data and thus interpretable only in the context of the experiment just performed. One can achieve this by considering the *random* p-value which has a uniform distribution under the null hypothesis but a highly skewed distribution under alternative hypotheses.

When considered as a random variable, the p-value becomes another test statistic, and thus one can ask, which test statistic, if not the p-value, best captures the notion of 'evidence' ? Morgethaler and Staudte (2005) suggest that the answer is a transformation which takes the test statistic into *evidence* T which has a normal distribution with mean τ and variance 1 for all values of the distributional parameters of the test statistic. The mean evidence τ should grow from 0 as the alternative moves away from the null; and, further, for a fixed alternative should grow at the same rate as the alternative can

*Imperial College, London,e.kulinskaya@imperial.ac.uk
†École Polytechnique Fédérale de Lausanne, stephan.morgenthaler@epfl.ch
‡La Trobe University, Melbourne, r.staudte@latrobe.edu.au

be estimated, typically the square root of the sample size. The advantages of variance stabilization have long been appreciated by statisticians, (see Anscombe (1948) and Efron (1982), for example. They include small sample normal approximations, and confidence intervals for τ which can easily be transformed into confidence intervals for a model parameter.

2 Calibrating the evidence in p-values

For definiteness consider the simple model in which the test statistic is the sample mean $\bar{X}_n$ having the normal distribution with unknown mean μ and standard deviation $1/\sqrt{n}$. For testing $\mu = 0$ against $\mu > 0$ the random p-value is $PV = \Phi(-\sqrt{n}\bar{X}_n)$. The probit transformation $p \to \Phi^{-1}(1-p)$ clearly transforms PV to $T = \sqrt{n}\bar{X}_n$ which has the normal distribution with mean $\tau = \sqrt{n}\,\mu$ and variance 1, thus satisfying for any $\mu > 0$ the desirable properties of evidence (see $E_1 - E_4$ below).

Instead of reporting a p-value of 0.05, we advocate reporting evidence $T = 1.645$, plus or minus standard error 1. Further, on this scale a p-value of 0.01 is reported as 2.33, plus or minus standard error 1. So 0.01 reflects only about 41% more evidence than 0.01 in this example, subject to equal standard errors of 1. To obtain twice the expected evidence, a p-value of 0.0005 is required. This is more in keeping with what Bayesian statisticians have been arguing for years, although a recent (Selke, Bayarri and Berger, 2001) Bayesian calibration scale for the p-value, when examined from the point of view espoused here, shows that posterior probabilities of the null underestimate the expected evidence in the p-value by at least one standard deviation over the range of interest.

In the context of Neyman Pearson hypothesis testing, the expected evidence is simply the sum of the probits of the false positive and false negative rates, so once the expected evidence is found, a formula for the power function of a test can be deduced. In addition, bits of evidence on the probit scale are easily combined, facilitating the computation and interpretation of evidence for joint alternatives in multiple related experiments. Standard meta-analytic theory applies, but with *known* weights, which circumvents a major problem in meta-analysis.

Evidence for the two-sided alternative $\mu \neq 0$ is *not* simply $\Phi^{-1}(1-p^{\pm})$, where $p^{\pm} = 2\Phi(-\sqrt{n}|\bar{X}_n|)$ is the two-sided p-value, for this transformation is

not variance stabilized, having a singularity at $\bar{X}_n = 0$. However, the transformation $p^{\pm} \to T^{\pm} = \max\{0, \Phi^{-1}(1-p^{\pm})\}$ is 'nearly' variance stabilized. This example raises the question of how general is the calibration scale.

3 Calibration of evidence

Let θ be an unknown effect for which it is desired to test $\theta = 0$ against $\theta > 0$, and let S be a test statistic which rejects H_0 for large values of S. We want a measure of one-sided evidence T to satisfy:

E_1. The one-sided evidence T is a monotone increasing function of S;

E_2. the distribution of T is normal for all values of the unknown parameters;

E_3. the variance $\mathrm{Var}[T] = 1$ for all values of the unknown parameters; and

E_4. the expected evidence $\tau = \tau(\theta) = \mathrm{E}_\theta[T]$ is monotone increasing in θ from $\tau(0) = 0$.

In the simple example of a normal model with known variance all of the above properties hold exactly for evidence defined by the Z-test statistic; that is, the standardized effect. In general, properties $E_2 - E_4$ will hold only approximately, but to a surprising degree, even for small sample sizes.

We somewhat arbitrarily describe values of T near 1.645 as *weak* evidence against the null. Values of T which are twice as large we call *moderate* evidence, and values which are 3 times as large as *strong* evidence. Thus our definition of weak evidence follows Fisher's low standard when the null is true, but we are otherwise measuring evidence against the null on a different calibration scale, one which allows for interpretation whether or not the null hypothesis holds.

4 Example

Let X have the Binomial(n, p) distribution, with $0 < p < 1$. For testing $p = p_0$ against $p > p_0$, The classical transformation $a_n(p) = 2\sqrt{n}\ \arcsin(\sqrt{p})$ with $\tilde{p} = (X + 3/8)/(n + 3/4)$ does have an approximate normal distribution,

with unit variance (see p. 123, Johnson, Kotz and Kemp, 1995). Therefore we define the evidence against the null hypothesis for this one-sided alternative by $T = a_n(\tilde{p}) - a_n(p_0)$. Then, at least approximately, T is unit normal with expected value

$$\tau(p) = \mathrm{E}_{n,p}[T] \approx \{a_n(p) - a_n(p_0)\} - \frac{p - 0.5}{2\sqrt{np(1-p)}} \quad . \tag{1}$$

This T roughly satisfies properties $E_1 - E_4$. As an example, when $n = 9$, this two-term approximation to $\tau = \mathrm{E}_{9,p}[T]$ shown above is accurate to 0.05 for all $p \geq 0.5$ and the standard deviation $\mathrm{SD}_{9,p}[T]$ is within 0.05 of the target 1 for all $0.5 \leq p \leq 0.8$.

The maximum amount of evidence in a Binomial(n, p) experiment against $p = 0.5$ in favor of $p > 0.5$ occurs when $X = n$ and is $T_{max}(n) \approx \sqrt{n}\ \pi/2$. Thus to obtain 'strong' evidence against the null, the minimum sample size one needs must satisfy $5 \approx \sqrt{n}\ \pi/2$, or $n = 10$, and then one must observe $X = 10$. In the orthodox view, this sounds fairly difficult, for the p-value of this event would be $2^{-10} \approx 0.001$. But the p-value is computed under the null, and the null may well be false.

Many other examples of variance stabilizing transformations for test statistics are available in the references given below, but the requirements for a measure of evidence $E_1 - E_4$ are somewhat stronger. They make it easier to interpret evidence, to compare evidence obtained from different experiments, and to obtain simple confidence intervals for τ which can be converted into intervals for θ.

5 Evidence for heterogeneity

Given K studies measuring potentially different effects θ_k, for $k = 1, \ldots, K$ one often tests the null hypothesis of equal effects, or *homogeneity*, with an asymptotic Chi-squared test based on $Q = \sum_k w_k(\hat{\theta}_k - \hat{\theta}_w)^2$; Cochran (1954).

Unfortunately, when the weights in Q need to be estimated, the distribution of Q converges extremely slowly to its limit. But suppose it is possible to find evidence in the kth study $T_k \sim N(\tau_k, 1)$, where $\tau_k = \sqrt{n_k}\, m_k$, $m_k = m(\theta_k)$ and m is a monotone function free of k. Also let $\bar{m} = \sum n_k m_k / N$

be the weighted transformed effect, where $N = \sum_k n_k$. Then m_k can be estimated by $\hat{m}_k = T_k/\sqrt{n_k}$ and Cochran's $Q = \sum_k n_k(\hat{m}_k - \hat{\bar{m}})^2$; it measures heterogeneity of the m_k's directly, and of the θ_k's indirectly. Moreover this $Q \sim \chi^2_{K-1}(\lambda_Q)$, with $\lambda_Q = \sum_k n_k(m_k - \bar{m})^2$.

A variance stabilizing transformation of this Q to evidence is given by $T_Q = \{Q - \nu/2\}^{1/2} - \{\nu/2\}^{1/2}$, which satisfies $T_Q \sim N(\mathrm{E}[T_Q], 1)$ with $\mathrm{E}[T_Q] = \{\lambda_Q + \nu/2\}^{1/2} - \{\nu/2\}^{1/2}$. This T_Q satisfies $E_1 - E_4$ approximately, and is therefore a measure of *evidence for heterogeneity.*

6 Combinations of evidence on the probit scale

How one combines evidence in $\mathbf{T} = (T_1, \ldots, T_K)$ obtained in K studies depends on how much evidence T_Q one finds for heterogeneity of the θ_k's and on what specific alternative to the joint null $\theta_1 = \theta_2 = \ldots = \theta_K$ one wants evidence for. If there is only weak evidence for heterogeneity, one can assume the standard fixed effects model (all $\theta_k = \theta$) and find the evidence for $\theta > 0$ using $T_w = \sum_k \sqrt{w_k}\, T_k$, where $\sum_k w_k = 1$. Then, because $\tau_k = \sqrt{n_k}\, m(\theta)$, $T_w \sim N(\tau_w, 1)$ with $\tau_w = \sum_k \sqrt{w_k}\, \tau_k = m(\theta) \sum_k \sqrt{w_k n_k}$. A possible choice for $w_k = n_k/N$. Obvious confidence intervals for τ_w are easily transformed into intervals for θ, if desired.

If one chooses a fixed, but *unequal* effects model then there are several possible alternative hypotheses. For example, one can define an overall effect as the θ which transforms into a weighted average of the transformed effects $m_1, \ldots, m_K$ and find evidence for $\theta > 0$. This methodology is illustrated for one- and two-sample t-tests in Kulinskaya and Staudte (2006). Finally, one can assume a random transformed effects model which introduces a variance component on the range of the map m. Then inference on $\mu = m(\theta)$ can be carried out and transformed back into inference for $\theta = m^{-1}(\mu)$.

7 Summary

By means of variance stabilization, many routine test statistics can be transformed onto a calibration scale that allows for easy interpretation of results, and comparison and combination of evidence obtained in similar independent

experiments. While the proposed framework only leads to measures of evidence which are approximately normal, this has not been a hindrance to the greater goals of interpretation, combination and repeatability of evidence. It is basically a meta-analytic framework with *known* weights.

References

[1] F.J. Anscombe. The transformation of Poisson, binomial and negative binomial data. *Biometrika*, 35:266–254, 1948.

[2] P.F. Azorin. Sobre la distribución t no central I,II. *Trabajos de Estadistica*, 4:173–198 and 307–337, 1953.

[3] B. Efron. Transformation theory: How normal is a family of distributions? *The Annals of Statistics*, 10(2):323–339, 1982.

[4] N.L. Johnson, S. Kotz, and N. Balakrishnah. *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, New York, 1994.

[5] N.L. Johnson, S. Kotz, and N. Balakrishnah. *Continuous Univariate Distributions*, volume 2. John Wiley & Sons, New York, 1995.

[6] N.L. Johnson, S. Kotz, and A.W. Kemp. *Univariate Discrete Distributions*. Wiley, New York, second edition, 1993.

[7] E. Kulinskaya and R. G. Staudte. Confidence intervals for the standardized effect arising in comparisons of two normal populations. 2006. La Trobe University Technical Report No. 2006-4.

[8] S. Morgenthaler and R.G. Staudte. Calibrating significant p-values. 2005. Submitted for publication.

[9] C.D. Mulrow, E. Chiquette, L. Angel, J. Cornell, C. Summerbell, B. Anagnosetelis, M. Brand, and R.Jr. Grimm. Dieting to reduce body weight for controlling hypertension in adults (Cochran Review). In *The Cochran Library,* Issue 3. John Wiley & Sons, Chichester, UK, 2004.

[10] T. Selke, M.J. Bayarri, and J.O. Berger. Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55:62–71, 2001.

The `giRaph` package for graph representation in `R`

Luca La Rocca, Jens Henrik Badsberg and Claus Dethlefsen

February 28, 2006

Abstract

The `giRaph` package provides formal classes and methods to represent and manipulate graphs in the `R` language and environment for statistical computing and graphics. It is intended as a contribution to the `gR` project described by Lauritzen (2002). We consider a broad notion of graph, including graphs with loops, multiple edges and hyper-edges (i.e. edges involving more than two vertifces) both directed and undirected. In particular, *hyper-graphs* and *simple graphs* (as defined by Lauritzen, 1996) fall within the scope of our definition. Since there is no unique way to represent a graph which is optimal for all computations, we consider four different representations: incidence list, incidence matrix, adjacency list and adjacency matrix; see for example Ahuja *et al.* (1993). We provide classes for these representations, suitable to store graphs of different families, and classes for such families, handling alternative representations transparently with respect to the user. For each class we furnish a robust `initialize` method, that takes care of producing valid output from varied input, a nice `show` method, adopting typical graph notation, and methods to set and retrieve information. In addition, we implement conversions between graph representations, and between graphs of different families, by means of `coerce` methods. We also provide classes for directed and undirected edges, as well as for vertex sets, so that simple graph operations such as adding/removing an edge, or extracting an induced subgraph, can be performed via overloaded `+`, `-` and `*` operators. Finally, we provide an interface to other graph packages such as the interactive graphical tool `dynamicGraph` by Badsberg (2005) which is also part of the `gR` project.

Ahuja, R.K., Magnanti, T.L. and Orlin J.B. (1993) "Network Flows", Prentice-Hall, NJ.

Badsberg, J.H. (2005) "dynamicGraph", R package, version 0.2.0.1, available on CRAN.

Lauritzen, S.L. (1996) "Graphical Models", Clarendon Press, Oxford, UK.

Lauritzen, S.L. (2002) "gRaphical Models in R", R News, 2, 39.

Ecological Inference and Higher-Dimension Data Management

Olivia Lau[1] *Ryan T. Moore*[2] *Mike Kellermann*[3]

Abstract

Ecological inference (EI) takes contingency tables as the unit of analysis. These tables are described by marginal row and column totals (or proportions); the goal of inference is to determine the joint intra-table relationship between the rows and columns. Using a common example from political science, let the unit of analysis be a voting precinct in a state-wide election:

	Democrat	Republican	No Vote	Total
Black	?	?	?	423
White	?	?	?	219
Hispanic	?	?	?	14
Total	317	156	183	656

For a given election, we estimate each cell (e.g., the number of Blacks who voted for the Democratic candidate) for each precinct $i = 1, \ldots, I$, then aggregate across precincts to obtain election-wide results.

While some existing R packages (`MCMCpack` by Andrew Martin and Kevin Quinn and `eco` by Kosuke Imai and Ying Lu, for example) offer functions that analyze 2×2 models, we implement more general methods that can take more than two rows or columns. Our package will include:

- Extreme case analysis, or the method of bounds, suggested by Duncan and Davis (1953).
- Ecological regression described in Goodman (1953) using both frequentist point estimates and a Bayesian estimator that produces correct standard errors.
- R $\times$ C model described in Rosen et al. (2001) using three estimators: a Bayesian Markov-chain Monte Carlo algorithm, maximum likelihood, and penalized least squares.

Since the unit of analysis (each ecological table) is a matrix rather than a vector, studying the statistical problem of ecological inference requires computational innovation in higher-dimension data management. For example, in the case of the Bayesian R $\times$ C estimator, we need to keep track of an array of dimension:

$$\text{rows} \times \text{columns} \times \text{precincts} \times \text{simulations}$$

In typical electoral data, there may be four rows (Black, White, Hispanic, Other), three columns (Democrat, Republican, No Vote), and 11,366 precincts (in the case of Ohio in 2004), for a total of 136,392 cell parameters (about 1 GB of memory) *per simulation*. It is thus impossible to store every simulation, or even a substantially thinned number of simulations, without several terabytes of memory (supposing that R could handle that much). We propose to deal with this memory management issue for higher-dimension data in several ways:

- For each iteration (or every iteration saved), the user may specify a quantity of interest to be calculated and stored (rather than the parameter draws themselves).
- Rather than storing draws in the workspace, the Bayesian methods will have the option to `sink` draws to a file. Since these multi-dimensional data need to be formatted in two dimensions for disk storage, we will provide functions to reconstruct the higher dimensions upon reading the sunk file.

In addition, this package will provide wrapper functions to operate on the margins of higher-dimension arrays, providing useful summary and print functions.

[1]Ph.D. Candidate, Department of Government, and M.A. student, Department of Statistics, Harvard University, `olau@fas.harvard.edu`. An alpha version of this software may be found at `http://www.fas.harvard.edu/~olau/software/`.

[2]Ph.D. Candidate, Department of Government, and M.A. student, Department of Statistics, Harvard University, `rtmoore@fas.harvard.edu`

[3]Ph.D. Candidate, Department of Government, Harvard University, `kellerm@fas.harvard.edu`

References

Duncan, O. D. and Davis, B. (1953), "An Alternative to Ecological Correlation," *American Sociological Review*, 18, 665–666.

Goodman, L. (1953), "Ecological Regressions and the Behavior of Individuals," *American Sociological Review*, 18, 663–666.

Imai, K. and Lu, Y. (2005), *eco: R Package for Fitting Bayesian Models of Ecological Inference in 2x2 Tables*, R package version 2.2-1.

Martin, A. D., , and Quinn, K. M. (2005), *MCMCpack: Markov chain Monte Carlo (MCMC) Package*, R package version 0.7-1.

Rosen, O., Jiang, W., King, G., and Tanner, M. A. (2001), "Bayesian and Frequentist Inference for Ecological Inference: The $R \times C$ Case," *Statistica Neerlandica*, 55, 134–156, http://gking.harvard.edu/files/abs/rosen-abs.shtml.

GEOχP:an R package for interactive exploratory spatial data analysis

T. Laurent, A. Ruiz-Gazen [*], and C.Thomas-Agnan[†]
Université des Sciences Sociales, Toulouse 1
IMT-UT1 and GREMAQ, 21 allée de Brienne 31042 Toulouse, FRANCE

March 8, 2006

Abstract. Exploratory analysis of georeferenced data must take into account their spatial nature. GEOχP is a tool for researchers in spatial statistics, spatial econometrics, geography, ecology etc allowing to link dynamically statistical plots with elementary maps. This coupling consists in the fact that the selection of a zone on the map results in the automatic highlighting of the corresponding points on the statistical graph or reversely the selection of a portion of the graph results in the automatic highlighting of the corresponding points on the map. GEOχP includes tools from different areas of spatial statistics including geostatistics as well as spatial econometrics and point processes. Besides elementary plots like boxplots, histograms or simple scatterplots, GEOχP also couples with maps Moran scatterplots, variogram clouds, Lorentz curves, etc. In order to make the most of the multidimensionality of the data, GEOχP includes some dimension reduction techniques such as principal components analysis, sliced inverse regression and projection pursuit whose results are also linked to the map. It is flexible and easily adaptable. We illustrate the use of GEOχP with a data basis from education in the French Midi-Pyrénées region.

Key Words. Exploratory analysis, spatial econometrics, spatial statistics, interactive graphics, dimension reduction.

[*]e-mail : ruiz@cict.fr
[†]e-mail : cthomas@cict.fr

Rggobi2

Presented by Michael Lawrence;
work done in collaboration with Hadley Wickham.

"Serious" abstract: Rggobi2 — Bringing R and GGobi Closer

Rggobi2 aims to integrate R with GGobi, a software tool for exploratory data analysis that features multivariate visualization and interactive graphics, including linked plots and the grand tour. Rggobi2 is a rewrite of the original Rggobi, which was a valuable tool for loading quantitative data from R into GGobi and manipulating GGobi plots. Unfortunately, the lack of precisely defined goals led to inconsistent syntax in the R API and general instability throughout the interface. Realizing the value of synthesizing R's flexible data analysis with GGobi's interactive graphics, the project was restarted with a more systematic development approach and a greater level of commitment from the developers. The most significant improvement is the refactoring of the R API so that it follows a more natural R syntax. For example, GGobi datasets now masquerade as R data frames. The loading of R data into GGobi is now more robust and fully supports categorical variables. Besides these fundamental improvements, Rggobi2 aims to cover every functional aspect of GGobi, including full control over interaction and projection modes, listening to tour projections, translation of GGobi displays to R graphics, and embedding GGobi plots in RGtk2 interfaces. There is also a long term goal of enabling R to record an entire GGobi analysis session and later reproduce it. Rggobi2 is committed to fully realizing the marriage between numerical analysis and interactive graphics.

"Somewhat less serious" abstract: Rggobi2 — The Second Date Between R and GGobi

GGobi is a mature software tool with a career in exploratory data analysis. Its hobbies include multivariate visualization and interactive graphics, including linked plots and the grand tour. Several years ago, GGobi met R, a platform for statistical computing with whom we are all familiar. Their first date, Rggobi, was pleasant overall and even involved some serious conversation, including some data transfer and plot manipulation. However, it was a bit hard to break the ice and they were not quite sure how to approach one another. Admittedly, there were some awkward moments, like when they tried discussing categorical variables. Both GGobi and R agreed that there was a lot of chemistry between them, given GGobi's interactive graphics and R's flexible data analysis. Thus, they have decided to go on a second date, Rggobi2. They are resolving some of their minor differences and focusing on fundamentals like robust loading of R data into GGobi and a more consistent and natural R API. R is expanding their conversation to include every aspect of GGobi, including full control over interaction and projection modes, listening to tour projections, translation of GGobi displays to R graphics, and embedding GGobi plots in RGtk2 interfaces. Eventually they hope to become close enough that R can record an entire GGobi analysis session and later reproduce it. All signs point to a fruitful relationship.

The uroot and partsm R-Packages: Some Functionalities for Time Series Analysis

Javier López-de-Lacalle
School of Management and Economics
Universidad del País Vasco - Euskal Herriko Unibertsitatea
javlacalle@yahoo.es

23rd January 2006

This document reviews the tools maintained by the author for time series analysis. These tools are available within the uroot and partsm packages. Although the major concern of the author is macroeconomic time series, the functions in these packages can be of interest for other areas of research as well.

uroot performs unit root tests and graphics for seasonal time series. The statistical analysis provided by this package allows the user to determine the order of differencing in seasonal ARMA processes. This package also includes a graphical user interface [GUI] built by means of the tcltk package. The main feature of the GUI is the way in which different time series are organized through the tree widget. A root node is created when a time series is loaded, then transformations of those data (logarithms, first differences, subsamples,...) can be added to the corresponding node in the tree as a child node.

partsm fits periodic autoregressive time series models. These models can be regarded as time varying parameter models where the autoregressive parameters take different values for each season. Tests for periodicity in the autoregressive parameters, periodic integration, as well as PAR order selection criteria are also included.

This presentation is based on the documentation attached to the packages. It provides a guidance in the use of the functions implemented in the packages as well as recommendations for the practical analysis.

By presenting this document, the author expects to get some feedback from maintainers of other packages related to time series analysis and users alike. Suggestions for improvement of future versions of the packages are welcome.

Keywords: R, time series, seasonality, unit roots, PAR models.

Using R as a Wrapper in Simulation Studies

Tim F. Liao
University of Illinois

R is the computing software of choice today for many statisticians across a variety of disciplines, and is becoming the lingua franca of statistical computing because of its flexibility and availability. However, there are many statistical software packages that have been in use for the last two to three decades, and for many data analysts it will be useful if features of an existing program can be employed while the flexibility of R is taken advantage of. The inflexibility of many existing statistical packages makes a simulation study rather difficult, notably when a program, command, or subroutine must be called repeatedly a large number of times with a set of changing input values and with the output collected in an easily executable file.

In this paper I present a simulation application where the software LEM, a popular software for categorical data in the social and behavioral sciences, is called repeatedly from an R wrapper. The application is a simulation within a simulation. That is, R handles the external simulation of a large number times, allowing for varying input parameters. The specialty software handles the internal simulation of estimating a statistical model on randomly generated data. The wrapper allows the data analyst to change the input, store the output, and produce graphic interpretation of the simulation output. Although the example involves LEM in DOS mode, the same idea (not the code) can be generalized to other statistical programs in other platforms (with necessary modifications).

MARKOV DECISION PROCESSES, DYNAMIC PROGRAMMING, AND REINFORCEMENT LEARNING IN R

JEFFREY TODD LINS
THOMAS JAKOBSEN

SAXO BANK A/S

Markov decision processes (MDP), also known as discrete-time stochastic control processes, are a cornerstone in the study of sequential optimization problems that arise in a wide range of fields, from engineering to robotics to finance, where the results of actions taken under planning may be uncertain.

An MDP is characterized by mappings for a set of states, actions, Markovian transition probabilities, and real-valued rewards within the process. An optimal planning solution seeks to maximize the sum of rewards over states under some decision policy for state-action pairs given updated transition probabilities.

The concept of dynamic programming was introduced by Bellman and is a classical solution method for approaching MDPs, however, in practice, the applicability of dynamic programming may be prohibited by the sheer size of underlying state spaces for real world problems – Bellman's so-called "curse of dimensionality" – for whereas a linear program representing an MDP can be solved in polynomial time, the degree of the polynomial may be be large enough to render theoretical algorithms inefficient in practice.

In addition, many problems do not allow for direct observations of the state space or reward functions, but rather only of some noisy information about the current state. These so-called *partially observable* MDPs constitute a class for which exact solutions may only be found efficiently for the smallest of state spaces.

Reinforcement learning extends Bellman's equations and other approaches to methods which employ robust function approximations, in order to make solutions for MDPs and POMDPs computationally tractable, and many of the wide variety of these approaches leverage statistical methods, including least squares regression, Monte Carlo methods, simulated annealing, and Markov chain methods, available in many R packages.

We demonstrate dynamic programming algorithms and reinforcement learning employing function approximations which should become available in a forthcoming R package. We highlight particularly the use of statistical methods from standard functions and contributed packages available in R, and some applications of reinforcement learning to sequential stochastic processes.

References

[Bellman, 1961] Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press.

Date: February 27, 2006.

[Littman et al., 1995] Littman, M., Cassandra, A., and Kaelbling, L. (1995). Learning policies for partially observable environments: Scaling up. In Prieditis, A. and Russell, S., editors, *Machine Learning: Proceedings of the Twelfth International Conference*, pages 362–370. Morgan Kaufmann Publishers, San Francisco, CA.

[Sutton and Barto, 1998] Sutton, R. and Barto, A. (1998). *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA.

Quantitative Analysis, Saxo Bank A/S, Smakkedalen 2, 2820 Gentofte, Denmark

E-mail address: jtl@saxobank.com, tja@saxobank.com

Simple-R — A Windows-based interface to R for basic statistics

Gunther Maier
Institute for Regional Development and Environment
Vienna University of Economics and Business Administration

This paper will present Simple-R, a Windows-based interface to R intended for basic statistical analysis. The program is written in Visual Basic and intends to significantly lower the entry barrier to R for students and faculty members occasionally using statistical analysis methods. The program is not intended for experts and gives direct access only to a very limited set of R-functionality. The program offers, however, transition paths to the full set of R capabilities.

The presentation will give an overview of the philosophy behind Simple-R and will demonstrate the core features and capabilities of the program (in its current state of development). Strengths and weaknesses of the approach and possible further development steps will be discussed. Also, the results of an empirical investigation will be presented that tests, whether the Windows look and feel really can lower the entry barriers for novice users.

MASTINO: a suite of R functions to learn Bayesian Networks from data.

Massimiliano Mascherini, PhD
Department of Statistics
University of Florence
V.le Morgagni 59, 50134, Florence, Italy
mascher@ds.unifi.it

Abstract

Bayesian Networks (BNs), [2], are a widespread tool in many areas of artificial intelligence and automated reasoning because they perform probabilistic inference through very efficient algorithms. However, the problem of searching the BN that best depicts the dependence relations entailed in a database of cases it is hard to solve. Structural learning exploits algorithms which typically combine expert's knowledge with the information gathered in a database.

In this paper I present MASTINO: a suite of R functions to learn BNs from data. MASTINO is built on the top of the DEAL package, [1], and it provides several functions to learn Bayesian Networks from data in the score-and-search framework. In particular, the *P*-metric, [3], a new score to evaluate Bayesian Networks encoding prior information on structures, and the MGA algorithm, [4], an innovative genetic algorithm to search for the best Bayesian Networks, are implemented in MASTINO as well as many utility functions to work with BNs.

MASTINO has been successfully tested on several well-known Machine Learning benchmark datasets with excellent results. The package is freely available for use with R and it can be downloaded from the web site of the author: *http://www.ds.unifi.it/mascherini/*

References

[1] S. G. Bøttcher and C. Dethlefsen. DEAL: A package for learning bayesian networks. *Journal of Statistical Software*, 8(20):1–40, 2003.

[2] F. V. Jensen. *An introduction to Bayesian Networks*. Springer Verlag, New York, N.Y., 1996.

[3] M. Mascherini and F. M. Stefanini. Encoding structural prior information to learn large bayesian networks. *WP of the Department of Statistics - University of Florence*, 13, 2005.

[4] M. Mascherini and F. M. Stefanini. M-GA: A genetic algorithm to learn conditional gaussian bayesian networks. *Proceedings of the IEEE International Conference on Computational Intelligence for Modelling, Control and Automation*, 2005.

Four Dimensional Barycentric Plots in 3D

Geoffrey Matthews
Western Washington University
Bellingham, Washington, USA

February 17, 2006

An n-dimensional vector that describes a probability distribution over n possible outcomes, such as $\langle p_1, p_2, \ldots, p_n \rangle$, is overdetermined because of the requirement $\sum_i p_i = 1$. Hence points of a two-class distribution can be plotted on the line $[0, 1]$, with the endpoints representing the distributions $(0, 1)$ and $(1, 0)$, and an intermediate point, such as 0.3, representing $(0.3, 0.7)$. Points from a three-class distribution can be plotted over a triangle, with the corners of the triangle representing the distributions $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. Such plots (an example is seen in Figure 1, from the **triplot** help file in the **klaR** package [3]) are quite common in numerous disciplines.

Points over a four-class distribution can be plotted as points within a three-dimensional tetrahedron, with the four corners of the tetrahedron representing the distributions $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, *etc.* Such visualizations are useful in clustering and classification problems where points are given probabilistic membership in classes, and a 2D plotting method is given in the **klaR** package [3]. An example figure from the help page of the **klaR** package is provided as Figure 2.

I have developed a package, **quadplot3d**, using the **rgl** package [1] to create genuine 3D plots of four dimensional points in the probability distribution subspace. Figure 3 shows how the plots from Figure 2 look when replotted in 3D. The user can, of course, interact with the 3D figures, rotating, scaling, and changing the background, as with any other **rgl** plot.

In addition to plotting points, some of the features of the **misc3d** package [2] have also been ported to **quadplot3d**. Isosurfaces of a four dimensional function, for example, can be produced by the **quadcontour3d** routine, mimicking the **contour3d** function of the **misc3d** package. In Figure 4 I have plotted isosurfaces (at values 0.7, 1.5, and 1.8) of the *entropy* function, $-\sum_i p_i \log p_i$. The entropy function is widely used in data mining applications. An intuitive intuitive understanding of entropy, the placement of its troughs and ridges, for example, can be gained from the figure.

Visualizing points and functions in four dimensions has proved an important tool in my understanding of data as well as of the clustering, classification, and data mining algorithms that deal with probabilistic class membership. I believe this tool will provide a valuable adjunct to packages such as **klaR**, that attempt to bring together statistical and machine learning approaches. **R** provides a good platform for this kind of cross-disciplinary research.

References

[1] Daniel Adler. *rgl: 3D visualization device system (OpenGL)*, 2004. R package version 0.64-13.

[2] Dai Feng and Luke Tierney. *misc3d: Miscellaneous 3D Plots*. R package version 0.3-1.

[3] Claus Weihs, Uwe Ligges, Karsten Luebke, and Nils Raabe. klar analyzing german business cycles. In D. Baier, editor, *Data Analysis and Decision Support*, pages 335–343, Berlin, 2005. Springer-Verlag. (in print).

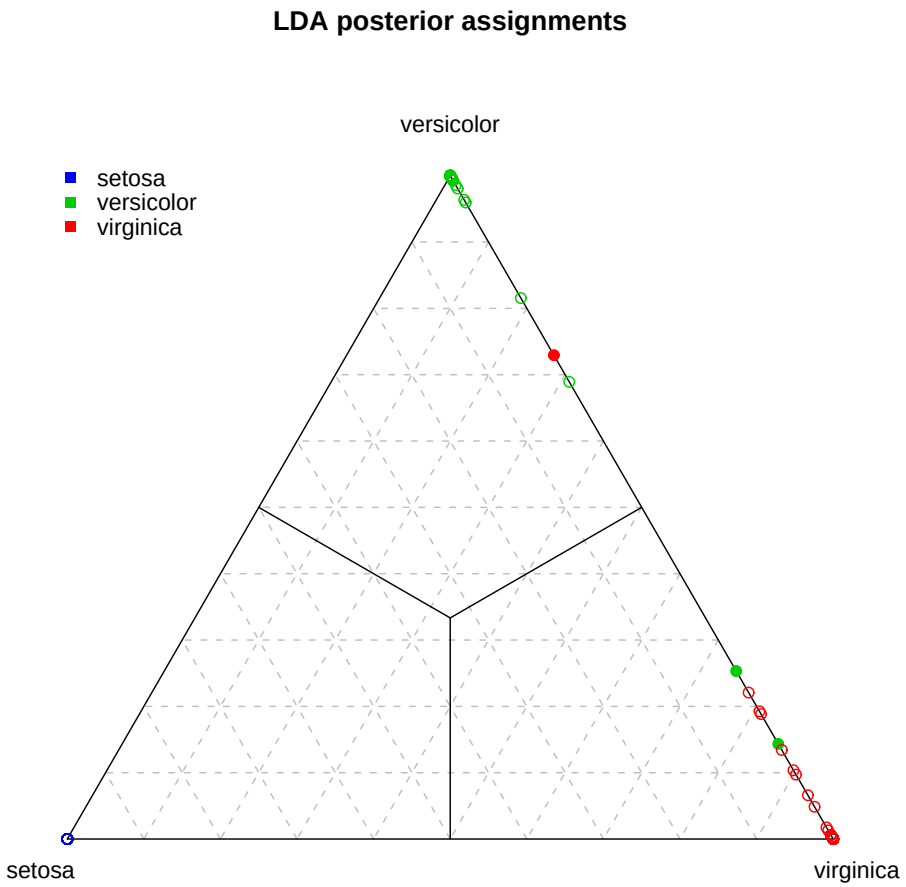


Figure 1: Barycentric plot of three dimensional points as provided by the **triplot** function of **klaR** package. Linear discriminants analysis (LDA) has provided a probability for each point's membership in the three classes of iris. Probability of membership in each of the three classes is given by distance from the "corner" positions, which represent unambiguous membership in the named class.

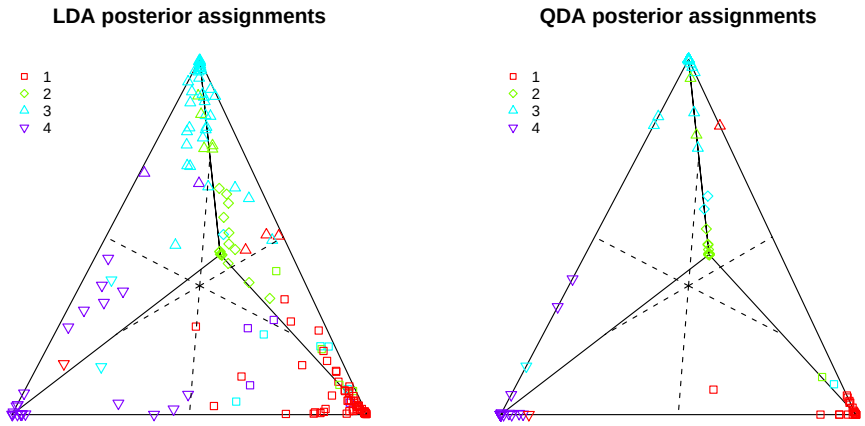


Figure 2: Two dimensional barycentric plots of four dimensional points as provided by the **quadplot** function of the **klaR** package. Here linear discriminants (LDA) and quadratic discriminants (QDA) have provided probabilistic membership for each of four classes. The corners in this case are the corners of a three dimensional regular tetrahedron, which has been projected down into two dimensions for the plot. The difficulties of visualizing four dimensions in a 2D projection are apparent.

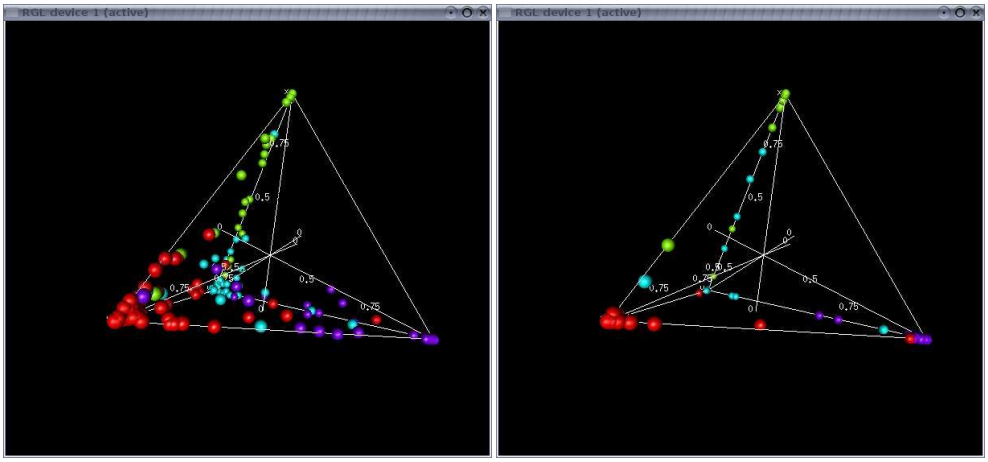


Figure 3: LDA and QDA posterior assignments in four dimensions (as in Figure 2) plotted in 3D using **quadplot3d**. 3D graphics provided in **R** by the **rgl** package are used to provide an interactive 3D view of the tetrahedron describing class membership. 3D hints such as size and perspective are apparent, and make the true relationships among the data points easier to see. Compare this figure with Figure 2.

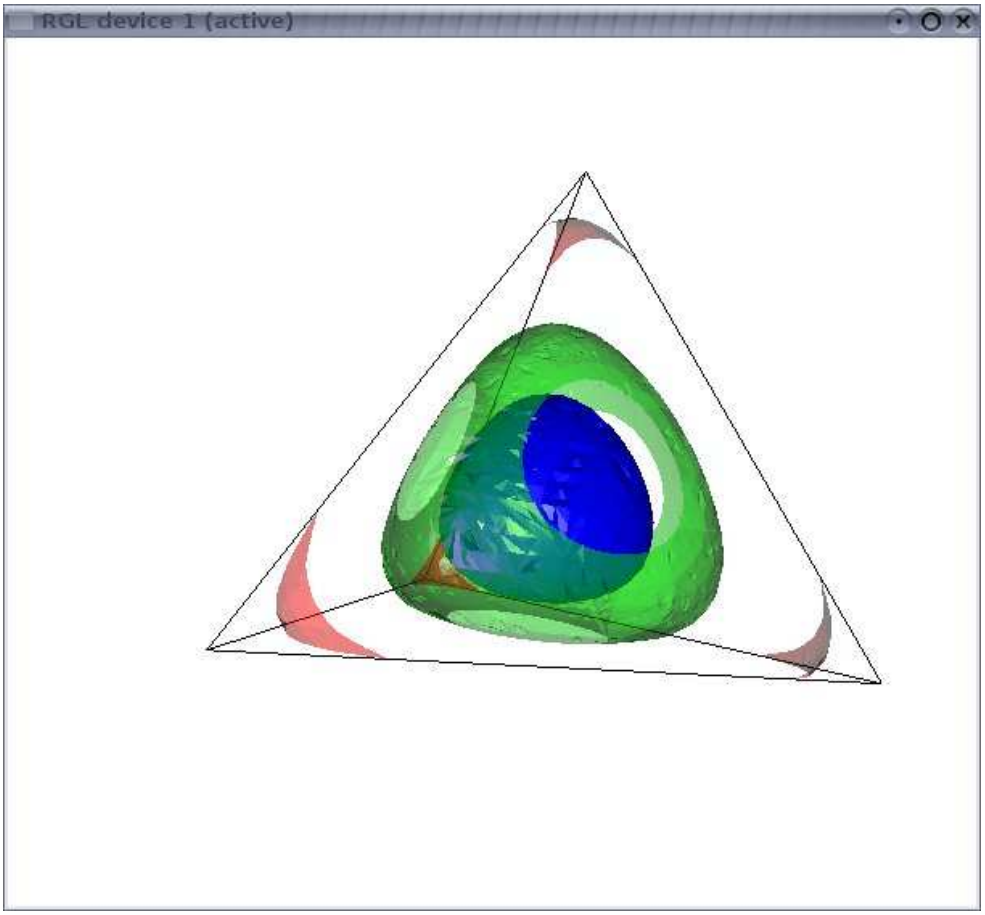


Figure 4: Isosurfaces of the entropy function in four dimensions plotted in 3D using **quadcontour3d**. A simple marching tetrahedra algorithm extracted surfaces of approximately equal value throughout the tetrahedra. Values of entropy used in the figure are 0.7, 1.5, and 1.8.

RIFFLE: an **R** package for Nonmetric Clustering

Geoffrey B. Matthews and Robin A. Matthews
Western Washington University
Bellingham, WA, USA

February 17, 2006

We present here an **R** package for RIFFLE, a nonmetric clustering technique [2]. This is a algorithm for clustering (unsupervised learning) that does not rely on a similarity measure for multivariate data, and uses only nonparametric (order) statistics. It is suitable for mixed nominal, ordinal, *etc.* attributes, as are often found in environmental data analysis.

The current implementation of RIFFLE in **R** has a number of improvements over the original implementation [2], utilizing a marginalized expectation maximization (EM) approach to speed up the search. This has advantages in avoiding local minima, as well. Also, a technique for creating useful seed clusterings has been developed (rather than completely random initial clusters, as in [2]), substantially speeding up the clustering and making the final cluster less susceptible to noise. Procedures are also provided to use the resulting clustering to find an optimal subset of the attributes, and to create a naive Bayes classifier.

RIFFLE has been used successfully in a variety of clustering tasks, and we have found it to be a useful, intuitive technique for graduate and undergraduate students in environmental sciences.

Although nonmetric clustering is over a decade old, it is not widely known. A recent authoritative survey [1],pp. 541-542, does not include a discussion of them, and laments, "How do we treat vectors whose components have a mixture of nominal, ordinal, interval and ratio scales? Ultimately, there are rarely clear methodological answers to these questions. ... We have given examples of some alternatives that have proved to be useful. Beyond that we can do little more than alert the unwary to these pitfalls of clustering." It is hoped that this **R** package will promote a more widespread use of nonmetric clustering in situations where it is appropriate.

References

[1] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification, Second Edition.* John Wiley & Sons, Inc., New York, NY, 2001.

[2] Geoffrey Matthews and James Hearne. Clustering without a metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):175–184, 1991.

Robust Statistics Collaborative Package Development: 'robustbase'

Martin Mächler and Andreas Ruckstuhl

R functions for robust statistics have been available for a while, in the standard 'stats' and the recommended 'MASS' package. However, researchers specialized in the field have for a while felt the lack of state-of-the-art methods on one hand, and of R tools that make robust data analysis as seamless as classical one, on the other hand.

Partly as result of a focused workshop last October, we've decided to start developing a package with at least two goals: - First, it should cover modern methods and algorithms for robust statistics, and thereby focus on methodology from the new monograph of Maronna, Martin and Yohai (2006). - Second, it should hopefully become the base of much further R developments within the field of robustness, and hence possibly define basic classes and methods.

The talk will give an overview on the new functionality in the package, notably robust GLM including model selection (Valentin Todorov will cover his own contributions in his own talk) but also comment on the experiences and tools used for collaborative package development (one maintainer, several package authors in diverse locations).

Statistical Approach to Operational Risk Management

Giulio Mignola and Roberto Ugoccioni,
Sanpaolo IMI Group, Torino, Italy

Banking supervising authorities require that, starting from 2007, banks calculate a regulatory capital for operational risks. Operational risks involve failures of normal business processes, e.g. mistakes, robberies, frauds, liabilities. In the most sophisticated approach, the capital requirement has to be computed taking into account both backward-looking historical loss data and forward-looking scenario analysis. No particular method is prescribed, but the risk measure at group level should provide a confidence level of 99.9% with a holding period of one year.

Because an industry best-practice in this field does not yet exists, the development of suitable methods required a tools which is both powerful and flexible, and an open source one was chosen to avoid an early lock-up in proprietary software.

Within Sanpaolo IMI, R is used at three levels:

- methodological research
- application prototyping
- production environment (in limited cases)

In historical loss data analysis, frequency and severity distributions of individual losses in a given risk class are studied, several distributions are fitted to the data with maximum likelihood estimation of parameters, establishing goodness-of-fit and choosing a best-fit distribution. The one-year period aggregate loss distribution is then computed from the characteristic functions via fast-Fourier-transform techniques and directly from the cdf via Monte Carlo techniques. The aggregated distributions for the different risk classes are then put together taking into account empirical correlations by making use of the copulas formalism. The regulatory capital is computed as the 99.9% quantile of the resulting total loss distribution.

In scenario analysis, in order to compute an equivalent figure, one has to obtain from "experts" within the bank (local managers) an estimate of the frequency and severity distributions: the method of moments and quantile matching has been chosen in order to gather from the interviewed the smallest and most precise information. To achieve this objective, one has to avoid open-answer questions and carefully choose answer ranges by calibrating a frequency dependent severity scale on the basis of the projected final risk measure. Again to focus the answers onto precise problems, the level of granularity in scenario analysis is greater than in loss data analysis, and it breaks the risk classes into event type sub-classes. This methodology requires preparing in advance, for each frequency and severity distribution type, curves of iso-UL (locus of points in parameter space with the same yearly aggregated unexpected loss), from which to compute answer ranges while the interview proceeds.

The risk measures from historical losses and scenario analysis are finally compound using Bayesian methods.

Using R via PHP: R-php

Angelo M. Mineo & Alfredo Pontillo
Università di Palermo

R-php is a project developed inside the Department of Statistical and Mathematical Sciences "Silvio Vianelli" of the University of Palermo (Italy); this project has as goal the realization of a web-oriented statistical software, i.e. a software that the final user reaches through Internet and that requires for its running only the installation of a browser, that is a program for the hypertext visualization. R-php is an open-source project with the code released by the authors with a General Public License (GPL) and can be freely installed.

A main feature of R-php is that all the statistical analyses exploit as "engine" R. Then, R-php is classifiable among the web projects dedicated to R. Beyond this project, there are others such as: R-web, R_PHP_ONLINE, and so on (see the web page `http://franklin.imgen.bcm.tmc.edu/R.web.servers`). Shortly, the main difference between R-php and the previously cited projects consists in the presence of a developed interactive module (R-php point-and-click) inside R-php, that allows to make some of the main statistical analyses without the user has to know the R statistical environment.

The R-php installation on a server is possible on numerous popular server platforms, such as Linux, Solaris, AIX or MacOS X and for its correct running is requested the installation of some additional software; the client side use is cross/platform and the only requirement is a browser supporting forms and the Javascript language.

The potential users of R-php are several: let's think about, for example, students, either inside didactic facilities, such as computer laboratories, or at home by means of a simple Internet connection, or a possible user that does not know and does not want to learn a programming environment as R, but wants to make some simple statistical analyses without buying the license of an expansive commercial statistical software with a developed graphical user interface.

In this talk, we describe the general design of R-php.

Graphical Exploratory Data Analysis Using Halfspace Depth

Ivan Mizera

The contribution centers around the implementation of two graphical exploratory tools based on different variants of halfspace depth. The first one is the `bagplot`, a bivariate generalization of the univariate boxplot proposed by Rousseeuw, Ruts, and Tukey (1999). The second one is the `lsdplot`, proposed by Mizera and Müller (2004). Some other, related procedures are explained, implemented and discussed as well.

While data depth in general, and halfspace depth in particular, offers a lot of data-analytic inspiration, it is the two-dimensional setting that is best shielded from various algorithmic curses of dimensionality haunting the general concept, and offers a visual perspective that might attract even conservative mainstream practicioners. Hence the choice of the subject, whose second part was driven by an evident desire for self-promotion, not that unknown from various R contributed packages; however, the first part is an attempted penance in the form of public service, not that unknown from various R contributions even more. In order to give the same care to both biological and adoptive child, a new contouring algorithm for location halfspace depth (rather than the existing S code cut and pasted) had to be developed in collaboration with David Eppstein. As a result, we aim at giving the useR not a quick brew, but a considerably matured product, in which not only `lsdplot`, but also `bagplot|` is capable of swallowing and digesting hundreds of thousands datapoints in mere seconds.

References

Peter J. Rousseeuw, Ida Ruts, and John W. Tukey (1999). The bagplot: A bivariate boxplot. *The American Statistician*, **53**, 382–387.

Ivan Mizera and Christine H. Müller (2004). Location-scale depth (with discussion and reejoinder), *Journal of the American Statistical Association*, **99**, 949–989.

KernGPLM – A Package for Kernel-Based Fitting of Generalized Partial Linear and Additive Models

Marlene Müller

Fraunhofer ITWM, Fraunhofer-Platz 1, D–67663 Kaiserslautern (Germany)
marlene.mueller@itwm.fraunhofer.de

Abstract

In many cases statisticians are not only required to provide optimal fits or classification results but also to interpret and visualize the fitted curves or discriminant rules. A main issue here is to explain in what way the explanatory variables impact the resulting fit.

The R package KernGPLM (currently under development) implements semiparametric extensions to the generalized linear regression model (GLM), in particular generalized additive and generalized partial linear models. A focus is given to techniques which are applicable for highdimensional data.

This covers in particular backfitting and marginal integration (Hengartner et al., 1999) techniques, which are both approaches for fitting an additive model when the underlying structure is truly additive. If the underlying structure is non-additive, however, both techniques may produce results that can differently be interpreted. While backfitting searches for the best projection on the additive function space, marginal integration estimators attempt to find the marginal effects of the explanatory variables. The KernGPLM package aims to provide estimation routines for the comparison of these different approaches.

Keywords

additive model, generalized additive model, generalized partial linear model, kernel smoothing, kernel-based regression

References

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., 2004. Nonparametric and Semiparametric Modeling: An Introduction. Springer, New York.

Hastie, T. J., Tibshirani, R. J., 1990. Generalized Additive Models. Vol. 43 of Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Hengartner, N., Kim, W., Linton, O., 1999. A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. Journal of Computational and Graphical Statistics 8, 1–20.

Müller, M., 2001. Estimation and testing in generalized partial linear models — a comparative study. Statistics and Computing 11, 299–309.

Müller, M.; Schimek, M. G., 2006. Classification of Highdimensional Data by Semiparametric Generalized Regression Models, Interface 2006, to appear.

TIMP: A package for parametric modeling of multiway spectroscopic measurements

Katharine M. Mullen, Ivo H.M. van Stokkum

February 24, 2006

Multiway spectroscopic measurements Ψ are collections of spectra representing a photophysical system at many different times or other conditions. Simultaneous analysis of such measurements in terms of a parametric model (*global analysis* in the photophysical literature) provides insight into the dynamics of the underlying system.

Measurements Ψ represent a superposition of the contributions of n_{comp} spectrally distinct components. The concentration and spectral property of each component may be represented as column l of matrices C and E, respectively, where C is of dimension $nt \times n_{comp}$, E is of dimension $nl \times n_{comp}$, nt is the number of times or other conditions at which spectra were measured, and nl is the number of (wavelength or wavenumber) points by which spectra are represented. The basic superposition model for such measurements is $\Psi = CE^T$.

Given Ψ, the inverse problem of recovery of the entries of C or E in terms of physically significant parameters (descriptive of, e.g., the decay rate of a component, or the location of the maximum of a spectrum) is often of interest. Adequate parameterizations of either C or E are nonlinear, and are often hierarchical in nature. For example, column l of C may be described as $C_l = exp(-k_l t) \otimes gaus(\mu, \sigma)$, where k_l is a decay rate parameter, t is time, $\otimes$ is convolution and μ and σ are location and width parameters, respectively. Parameters μ and σ may in turn be described as nonlinear functions of e.g., wavelength λ, so that C_l is wavelength-dependent and the model description is hierarchical [1]. Once a model is parameterized for C or E, the entries of the remaining matrix may often be treated as conditionally linear, taking advantage of the separable model form [2]. Intrinsically nonlinear parameter estimates may be derived by nonlinear regression, which has the advantage of returning parameter confidence estimates valuable in model interpretation and validation [3].

R has been used to prototype a problem-solving environment for parametric modeling of multiway spectroscopic measurements. Hierarchical models for multiway spectroscopic measurements find natural description as a hierarchy of S4 classes. S4 methods provide a means to differentiate the treatment of many different model types while maintaining uncomplicated calling code. The *nls* and *numericDeriv* functions provide a fast basis for nonlinear regression. An extended version of the problem-solving environment for public release is being developed as the package "TIMP".

References

[1] Ivo H. M. van Stokkum, Delmar S. Larsen and Rienk van Grondelle, "Global and target analysis of time-resolved spectra", *Biochimica et Biophysica Acta*, vol. 1657, pp. 82–104, and erratum, 1658, 262, 2004.

[2] Katharine M. Mullen, Mikas Vengris and Ivo H.M. van Stokkum, "Separable nonlinear models for time-resolved spectra", in I. García, L. G. Casado, E.M.T. Hendrix and B. Tóth, editors, *Proceedings of the International Workshop on Global Optimization*, pp. 183–188, September 2005.

[3] Douglas M. Bates and Donald G. Watts, *Nonlinear regression analysis and its applications*, John Wiley & Sons, 1988.

Can R Draw Graphs?

Paul Murrell

This talk will describe and demonstrate several new graphics features that are available with the release of R 2.3.0.

X-splines are smooth curves that are drawn relative to a set of control points, with a parameter controlling whether the curve approximates or interpolates each control point. We will look a new X-spline drawing primitive and show some simple applications such as using an open X-spline to draw a curve connecting a label to a location of interest, and using a closed X-spline to produce a novel plotting symbol.

Connectors are straight, bent, or curvy lines that join nodes in a graph or flow diagram. We will look at new features in the grid graphics package that build on X-splines to allow quite general connectors to be drawn between graphical objects, with examples showing the construction of simple graphs and flow diagrams.

Vector images are images described in terms of geometric shapes (as opposed to individual pixels as in bitmap images). We will look at a new package, grImport, which provides support for importing vector images into an R plot. Demonstrations of this package will include adding a background "watermark" to a plot and importing novel plotting symbols.

Testing volatility interactions in a constant conditional correlation GARCH model

Tomoaki Nakatani* and Timo Teräsvirta†

Department of Economic Statistics, Stockholm School of Economics,

P.O. Box 6501, SE-113 83 Stockholm, Sweden

February 2006

Abstract

In this paper, we propose an Lagrange multiplier (LM) test for the presence of volatility interactions among markets/assets. The null hypothesis is the constant conditional correlation (CCC) GARCH model of Bollerslev (1990) in which volatility of an asset is described only through lagged squared residuals and volatility of its own. The alternative hypothesis is an extension of that model in the way of Jeantheau (1998), where volatility is modelled, while keeping the conditional correlation structure constant, as a linear combination not only of its own lagged squared residuals and volatility but also of those in the other equations. As an example, we derive expressions for the LM test in the bivariate case along with the necessary derivatives of the likelihood function, and conduct simulation experiments to investigate finite sample properties of the test. Empirical applications are carried out for pairs of foreign exchange rates and of stock indices. Results indicate that there indeed exist volatility interactions in some pairs that are detected by the proposed test.

Keywords: Multivariate GARCH; Lagrange multiplier test; Monte Carlo simulation; Constant conditional correlation.
JEL codes: C12; C32; C51; C52; G1.

* E-mail: tomoaki.nakatani@hhs.se
† E-mail: timo.terasvirta@hhs.se

Estimating Consumer Demand for Hedonic Portfolio Products: A Bayesian Analysis using Scanner-Panel Data of Music CD Stores

Yuji Nakayama, Tomonori Ishigaki and Nagateru Araki
College of Economics, Osaka Prefecture University, Japan

Most products have hedonic and utilitarian attributes. Products that mainly offer hedonic benefit in the form of an affective experience (e.g. movies) are called *hedonic products*, while products providing the utilitarian benefit of pragmatic functionality (e.g. personal computers) are entitled *utilitarian products*. If a person prefers a specific category of hedonic products, he or she repeatedly buys products in that category, although it is rare that he or she repeatedly purchases an individual product . That is, many hedonic products are purchased as one of a person's possessions . Such products are categorized as *hedonic portfolio products*(Moe and Fader, 2001): a typical example is music compact disks (CD).

In this study, we conduct an empirical analysis using sales data from music CD stores in Japan. The distinguishing feature of our data is that the study contains the ID number of customers who purchase a specific music CD title, although no demographic data is collected to ensure privacy. We use this scanner-panel data to estimate each customer's demand for his or her collection of music CDs (e.g. Pop, Rock, Jazz or Classic), following earlier work by Kim, Allenby and Rossi (2002). For estimation, we use an R package, `bayesm`, developed by Rossi, Allenby and McCulloch (2005). Finally, we discuss the retailers' assortment and pricing strategy to improve both profitability and maintain customer satisfaction.

References

[1] Kim, J, Allenby, G.M., Rossi P.E. (2002) "Modeling Consumer Demand for Variety" *Marketing Science*, Vol 21 (3) 229–50.

[2] Moe, W.W., Fader, P.S. (2001) "Modeling Hedonic Portfolio Products: A Joint Segmentation Analysis of Music CD Sales" *Journal of Marketing Research*, Vol 38 (August) 376–85.

[3] Rossi, P.E., Allenby, G.M., McCulloch, R. (2005) *Bayesian Statistics and Marketing*, John Wiley & Sons, Ltd.

Neural network algorithms and related models

Stefan Neubauer, Georg Dorffner

Inst. of Medical Cybernetics and Artificial Intelligence, Center for Brain Research,
Medical University of Vienna

The NETLAB toolbox for MATLAB™ (Nabney 2002) is well established for teaching and research in the fields of pattern recognition and data analysis. To provide students and practitioners those tools also outside the Matlab framework, we have implemented an R package covering NETLAB's complete functionality. Although some tools for neural networks are already available in existing R packages, this new implementation enables consistent usage and should make it easier for existing NETLAB users to switch to R. The code is written 100% in pure R, thus making it easy to adapt models to one's own needs and ensuring platform independence. The original toolbox does not use object oriented methods, while in the R implementation S3 mechanisms were facilitated.

All tools and methods provided are theoretically well founded, with in-depth discussion in Bishop (1995). Complementary information for the toolbox - and thus implicitly also for the R counterpart - is given in Nabney (2002), where also more recent developments that are part of the toolbox are described.

The toolbox covers general purpose optimization routines, for example scaled conjugate gradient and quasi Newton methods. For density estimation, Gaussian mixture models (GMMs), Probabilistic Principle Component Analysis (PPCA), Generative Topographic Mapping (GTM) are available. The most commonly used neural network models are implemented, i.e. the multi-layer perceptron (MLP) and the radial basis function (RBF) network. Simpler models, such as K-means clustering, K-nearest-neighbor classifiers, and single layer network, are included to act as benchmarks. The Mixture Density Network (MDN) provides a general purpose model for conditional density estimation and to model multi-branched functions. For visualization, besides PCA, PPCA, SOM and the more principled alternative GTM, Neuroscale is included, which is a non-linear topographic projection that uses an underlying RBF network. The Bayesian approach to neural networks is addressed in a twofold way: the evidence procedure allows adding error bars to predictions and automatic determination of variable importance; Metropolis-Hastings and hybrid Monte Carlo methods are provided for sampling.

References

Bishop, C. M. (1996). Neural Networks for Pattern Recognition, Oxford University Press.
Nabney, I. T. (2002). Netlab: Algorithms for Pattern Recognition. Berlin, Springer.

RXL – A Free Excel Add-in for Introductory Business Statistics

Pin Ng
College of Business Administration
Northern Arizona University

In the last decade, there has been a trend among business schools to shift the focus of an Introductory Business Statistics course from the traditional approach of teaching statistics via formulae to an interpretive approach which emphasizes interpretations of statistical output obtained with the help of some statistical software. A survey of existing textbooks in the market reveals that a majority of them incorporate detailed instructions on *Excel* and its add-ins. Only a small portion of the remaining textbooks utilize other statistical software such as *Minitab*, *SPSS* or *SAS* to perform the statistical computations.

However, in a position paper to the Mathematical Association of America, the American Statistical Association (ASA, 2000) commented that "Generic packages such as Excel are not sufficient even for the teaching of statistics, let alone for research and consulting." Numerous studies have highlighted the deficiencies and dangers of using *Excel* as a statistical package for teaching and research.

As a result, there have been quite a few *Excel* add-ins written to address and attempt to solve the problems of using *Excel* and its add-ins in the Microsoft Data Analysis Toolpak. Some examples are *Analyse-it*®, *Fast Statistics*©, *Lumenaut*©, *N-SEA*©, *PopTools*, *SigmaXL*®, *statistiXL*©, *UNISTAT*®, and *XLSTAT*©. With the exception of *PopTools*, which is written specifically to analyze ecological models, these add-ins are commercial products that have an annual single user license fee. In light of the ever rising textbook prices and the costs of attending colleges/universities, it will be valuable to the students, instructors and researchers to have the freedom of using an *Excel* add-in that utilizes the familiar interface of *Excel* and offers an extended range of statistical procedures that are already available in *R* without having to be burdened with the usage cost. We attempt to accomplish this with an *Excel* add-in, *RXL*, that utilizes the "macro mode" and "worksheet functions" capability of *R-Excel* developed by Thomas Baier and Erich Neuwirth.

RXL will include within its menu driven GUI many of the procedures covered in a typical introductory statistics course. It also replaces a list of *Excel* functions that are commonly used in an introductory business statistics course with the corresponding *R* functions while retaining the valuable automatic recalculation feature of *Excel*.

RXL will be distributed under the GNU GPL Agreement. The GPL puts students, instructors and researchers in control of their usage of the software by providing them with the freedom to run, copy, distribute, study, change and improve the software, thus, freeing them from the bondage of proprietary software.

The continuous evolution of *RXL* will not only have a significant impact on the teaching of an introductory statistics course by providing a free alternative to the commercial proprietary software but also provide researchers in all disciplines who require sophisticated and cutting edge statistical and graphical procedures with a user-friendly interactive data analysis tool.

Spatial characteristics of vegetation index map in urban area derived by variogram analysis

Keiji Osaki
International Christian University
Tokyo Japan

It is important to keep monitoring the characteristics of vegetation in urban areas by remotely sensed data from an environmental viewpoint. Recent sensors on earth observation satellites have excellent and very high spatial resolution of a few meters on the surface of the earth. However, the higher the resolution of sensor becomes, the more difficult the analysis of land surface seems to be due to too fine aspects in the satellite imagery. The primary objective of the research presented here is to clarify the spatial characteristics of vegetation distribution in urban areas by analyzing properties of variograms . To depict the spatial patterns of the observed scenes, quantity measured by second-order statistics has been used in applications such as mining exploration and other engineering fields. Related practitioners call the field of spatial statistics 'geostatistics'. An important concept of the quantity inherent to the scenes is spatial continuity and is measured as covariance and semivariance. The semivariance plays a very important role in the analysis of data's spatial statistics in the present research. Among many vegetation indices, NDVI(normalized difference vegetation index) has been most widely used for investigation of environmental assessment. The NDVI data used for the current analysis is derived from the multi-spectral data of 'QuickBird' (earth observation satellite) with the resolution of 2.8m. The NDVI lies in its characteristics that can reduce the multidimensional data yielded from multi-spectral sensor systems to a single index which is sensitive to various characteristics related to vegetation activities such as biomass, productivity, leaf area, amount of photo-synthetically active radiation, and percent vegetative ground-cover etc.

The semivariogram is defined as the following equation:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1,N} (z_i - z_{i+h})^2$$

A useful measure of spatial variation in the values of a variable z is the semivariance, which is half the average squared difference in z values between pairs of sample points. The key to investigation of the semivariance is the construction of a semivariogram, which is a plot of the semivariance, as a function of distance h. At a distance referred to as "range", the semivariance levels off to a relatively constant value, referred to as the "sill". This implies that beyond this range, z values are no longer spatially correlated.

The variogram can provide the spatial structure or patterns of observed objects on the earth quantifying dissimilarity as a function of separation and direction. Here we skip the effects of anisotropy for simplified analysis.

As is shown in the following equation, we introduce two sets of parameters which represents the characteristics of variograms, that is, two 'sills' and 'ranges'. By nonlinear least squares regression method 'nls' in 'R', we can derive two sets of characteristic parameters of variograms which can let variograms be fitted to 'nested spherical model' (Hans Wackernagel, 2003)[1] :

$$\gamma(h) = \begin{cases} \sum_{k=1}^{2} \mathrm{Sill}_k \left\{ \frac{3}{2}\frac{h}{\mathrm{Range}_k} - \frac{1}{2}\left(\frac{h}{\mathrm{Range}_k}\right)^3 \right\} & \text{for } 0<h<\mathrm{Range}_k \\ \sum_{k=1}^{2} \mathrm{Sill}_k & \text{for } h > \mathrm{Range}_k \end{cases}$$

where sill1, range1, sill2 and range2 are defined by nonlinear least squares regression fitting. We can identify the difference between two areas i.e. "range" is larger for area, which contains much natural objects such as vegetation than for area like urban area with short-range which contains many artifacts. Since it demands vast computation cost for calculation of semivariances for remotely sensed scenes, we use ordinary simple 'random sampling ' method to select sampling pixels from the target areas.

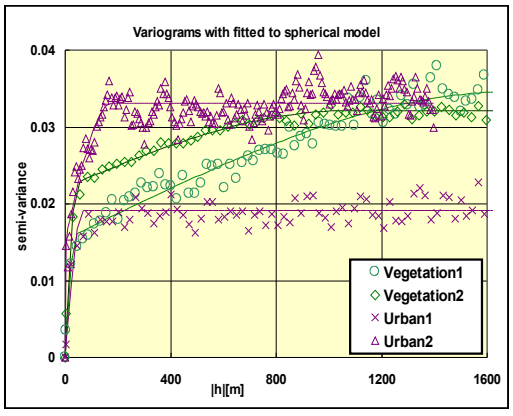


Figure Variograms for vegetation and urban areas with lines fitted to nested spherical model.

To extract interesting and important features of variograms in certain range of 'lag's where variograms would show rich vegetation characteristics, variograms and their fitted curves are calculated by nonlinear least squares regression fit to nested spherical model in the above equation. It is interesting shown in Figure that two types of rich and poor vegetation areas show their intrinsic spatial pattern in variograms. Variograms in urban areas show almost zero slope in the range of 'lag' (500m < |h| <1000m), while in richly vegetated areas have slower increase of variogram in that range.

References

[1] Hans Wackernagel, *Multivariate Geostatistics*, Springer, 2003.

Bayesian analysis of Dynamic Linear Models in R

By Giovanni Petris, University of Arkansas

In recent years state space models, and dynamic linear models (DLM) in particular, have become more and more important for the analysis of temporal data, particularly from a Bayesian perspective. The increase in computing power, coupled with the development of sophisticated MCMC algorithms has made possible the use of realistically complex models. We have developed a set of functions in R that provides a flexible environment for Bayesian time series analysis using DLM. We see R becoming the new standard for statistical computing, and we believe that having a package for DLM analysis available in R will greatly facilitate the use of Bayesian time series analysis both among statisticians and applied scientists in general. In the talk we will give an overview of the package, including the discussion of some computational issues and novel methods devised to deal with them.

Population ecology modelling with R: a comparison of object oriented approaches

Thomas Petzoldt[a] Karsten Rinke[b] Louis Kates[c]

February 28, 2006

The R system with the underlying S programming language is well suited for the development, implementation and analysis of ecological models and it is increasingly accepted among ecologists. Existing applications already cover a range from small conceptual process and teaching models up to large coupled models on the ecosystem scale. Small models can be implemented very easily in pure R. For larger ones, R is primarily used as an environment for data management, simulation control and data analysis, while the model cores are usually implemented in other languages like C++ or Fortran. This works perfectly at the extremes, but problems appeared with medium-sized models:

- non-trivial ecological models are based on more or less modular building blocks (submodels), which are either the underlying base equations or complex models themselves.
- both, data and procedural code (submodels) are highly variable. A typical example is the comparison of similar models with identical data, but with slightly different submodels.

The `simecol` package was developed in order to provide an open and minimally standardized structure to implement and run ecological models; however, the S3 list-based structure of the first version had to be extended to organize more complex applications. Different object oriented approaches (S3 lists, proto, R.oo, S4) were evaluated using the example of a typical and yet simple Lotka-Volterra type model. The implementations were compared with respect to usability, flexibility, conformity to common standards and performance. A larger model[1] for the genus *Daphnia* is presented to demonstrate some of the object oriented approaches. This model represents a more complex and computation intensive simulation, a bioenergetic model which accounts for demographic population structure.

With the examples used here we found it advantageous to use OOP and the difference between using OOP and not using OOP was more significant than which OOP framework was used. If one OOP implementation exists it is relatively easy to transform it to one of the other versions, but distinctive differences and specific features of the packages remain. The purpose of this presentation is to compare the technical frameworks available in the R environment according to their suitability to organize ecological models. Reference applications are provided that can help ecologists to structure their work. Moreover, the examples demonstrate that it would be feasible to use R and OOP as a medium for the distribution and share of ecological modelling code.

[a]Technische Universität Dresden, Institute of Hydrobiology, 01062 Dresden, Germany, thomas.petzoldt@tu-dresden.de, `http://tu-dresden.de/Members/thomas.petzoldt`

[b]Universität Konstanz, Limnological Institute, Mainaustrasse 252, 78464 Konstanz, Germany, karsten.rinke@uni-konstanz.de, `http://www.uni-konstanz.de/limnologie/ags/Karsten/`

[c]GKX Associates Inc., Waterloo, ON, Canada, lkates@alumni.princeton.edu

[1]Rinke, K. & Vijverberg, J. (2005) A model approach to evaluate the effect of temperature and food concentration on individual life-history and population dynamics of *Daphnia*. Ecological Modelling, 186, 326-344

Teaching the Theory of Information and Coding with R

Rafael Pino Mejías, Mª Dolores Cubiles de la Vega
Statistics Department, University of Seville

The R system is a very appropriate resource for teaching statistic topics in a computer science faculty. Its powerful programming language may be used by the students to develop algorithms and to explore theoretical questions for a better understanding. These reasons guided us to use the R system as the main tool for the computer classes accompanying the subject "Theory of Information and Coding" in the Computer Science Engineering career offered by the University of Seville. So, from the interaction between the teachers and the students many R programs covering different tasks have been designed. For example, we have used R for visualizing the entropy, for simulating communication channels including the estimation of the probability of bit and block errors, and several linear and circular codes have been programmed. Moreover, some elements of data and image compression codes have also been written in R.

Currently, we are supervising the building of a package including these resources and a web page. Our presentation will describe these teaching experiences and the main developed programs.

Bayesian Modeling in R with JAGS

Martyn Plummer
International Agency for Research on Cancer

JAGS is a C++ library for analyzing Bayesian hierarchical models using Markov Chain Monte Carlo. It is not wholly unlike WinBUGS (Bayesian inference Using Gibbs Sampling) [1] and has many of the same features, with the notable exception of a graphical user interface.

JAGS takes a description of a model in the highly flexible BUGS language and a data set defining the observed variables. It then provides a sequence of samples from the posterior distribution of the unobserved variables which can be used for approximate Bayesian inference. The capabilities of JAGS can be extended with dynamically loadable modules. A module can extend the modeling language by defining new functions and distributions, or it can improve performance by defining new samplers for specific modeling situations, such as mixture models or generalized linear models.

This talk will review the JAGS package for R which is currently under development. The package uses R's simple interface to external references [2]. A JAGS model object in R contains an opaque pointer to an external C++ object. JAGS model objects therefore have a mutable state, which makes them different from the more familiar "fitted model" objects, such as those created by the `glm()` function, and requires an object-oriented user interface.

References

[1] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual, Version 2.0, June 2004. `http://mathstat.helsinki.fi/openbugs/`

[2] Tierney L. Simple References with Finalization. `http://www.stat.uiowa.edu/~luke/R/simpleref.html`

Data Profiling with R

Submitted by Jim Porzak, VP of Analytics, Loyalty Matrix, Inc., San Francisco, California.
JPorzak@LoyaltyMatrix.com

Data profiling should be the first step in any data mining project where we are not 100% certain that the source data actually is what it purports to be -- in other words, always. Done correctly, data profiling will discover data quality issues at the beginning of an analysis project before they impact subsequent processing, or worse, conclusions.

Data profiling is generally considered as part of the data quality process. While data quality is best achieved when, and where, the data is sourced, analytic practitioners don't have the luxury of waiting until a client achieves a state of data quality nirvana. We need to understand the data's limits and deal with it. Data profiling should not be confused with exploratory data analysis (EDA). EDA presupposes valid data. Data profiling discovers invalid or suspect data that must be corrected, discarded, or dealt with in subsequent data hosting and analysis.

A good data profiling tool will

- Require minimum input from an analyst to run.
- Do column profiling with simple statistics, plots, patterns, exceptions and domain detection.
- Do dependency profiling to identify any intra-table dependencies impacting normalization.
- Do redundancy profiling to discover keys between tables and other overlapping variables.
- Produce easy to use output which will be useful to analysts and meaningful to clients.
- Save findings to an accessible data structure for subsequent use.

Unfortunately commercial data profiling tools are generally targeted at "enterprise" data integration projects. They tend to be highly flexible and thus complex to use. Of course, they come with an enterprise worthy price tag.

At Loyalty Matrix, we must deliver data-driven insights quickly and economically. One of our tricks is to profile all new data sets upon receipt. Our data profiling tool has evolved over the last year having been used on dozens of projects.

Since our clients send us gigabyte, or low terabyte, datasets, we routinely load all client data into a SQL relational database (RDBMS), typically MS-SQL or MySQL. The initial step in any project is loading the raw data without transformation into the RDBMS. Our profiling tool takes advantage of this by using a combination of R and SQL (via RODBC) to optimize processing time and flexibility. We use grid graphics to create data summary panels that can be displayed individually or combined into a comprehensive report. Summary statistics and data hypothesis are written back to the RDBMS for subsequent reporting and integration into metadata repositories.

This paper will

- Review our requirements for the data profiling tool.
- Describe the high level design and structure of the tool.
- Show examples of the integration of R and SQL to achieve optimum processing.
- Show examples of grid graphics design and code for the data summary panels.
- Conclude with real-world examples of quality issues discovered with help of the tool.

end

Spike Sorting with R and GGobi

Christophe Pouzat, Andrea Ridolfi and Pascal Viot

February 27, 2006

Neurobiologists are more and more interested in studying neuronal populations. To this aim, one of the most popular experimental techniques is Multi-Electrode Array (MEA) recording which collects the activity of several neurons from each of many electrodes. But, in order to be really informative this technique requires the difficult "spike sorting" problem to be solved. The raw recorded data are indeed a mixture of signals (action potentials or "spikes") originating from many neurons. Spike sorting techniques aim at un-mixing these signals to get the single neurons spike trains. They in fact require the application of three classical data processing stages: 1) Pre-Processing, the spikes must be detected and the dimension of the space in which they are represented must be made as small as possible. 2) Clustering, individual spikes must be grouped into clusters corresponding to the different neurons whose number is not known a priori. 3) Classification, some detected events correspond in fact to two spikes from two neurons occuring almost simultaneously, they must be identified and classified as such; recordings are moreover long (several hours) and stable, the time consuming clustering stage is therefore typically performed on the first minutes of data and subsequent parts are directly classified.

R and GGobi provide wonderful tools to address all these issues. After several years of software development in pure C interfaced with Scilab and then R, we are presently rewritting our software, SpikeOMatic[1] as an R package using default and contributed packages like mclust, e1071, XML. S4 type of classes and methods allow both easy and intuitive data manipulation, because our objects behave in most cases like matrices, and a full access to R functionalities. Many analysis steps in spike sorting, that is, in clustering and/or classification, can also be trivially parallelized thanks to Rmpi and/or rpvm together with snow. Finally the powerfull data visualization features of R and GGobi provide crucial help to the neurophysiologist trying to understand the method he is using and/or to evaluate the trustworthiness of his results.

[1] `http://www.biomedicale.univ-paris5.fr/physcerv/C_Pouzat/newSOM/newSOMtutorial/newSOMtutorial.html`

Applied Bayesian Inference in R using **MCMCpack**

Kevin M. Quinn Andrew D. Martin

February 24, 2006

MCMCpack is an R package designed to allow users to perform Bayesian inference via Markov chain Monte Carlo (MCMC) for models commonly used in the social sciences. Currently **MCMCpack** allows the user to perform Bayesian inference via simulation from the posterior distributions of the following models: linear regression (with Gaussian errors), a general linear panel model, Quinn's dynamic ecological inference model, Wakefield's hierarchical ecological inference model, a probit model, a logistic regression model, a one-dimensional item response theory model, a K-dimensional item response theory model, a robust k-dimensional item response theory model, a Normal theory factor analysis model, a mixed response factor analysis model, an ordinal item response theory model, a Poisson regression, a tobit regression, a multinomial logit model, an SVD regression model, and an ordered probit model.

The posterior samples returned by each function are returned as mcmc objects, which can easily be summarized and manipulated by the **coda** package. **MCMCpack** also contains densities and random number generators for commonly used distributions that are not part of the standard R distribution, a general purpose Metropolis sampling algorithm, functions to compute Bayes factors for some models, a handful of teaching models, and some data visualization tools for ecological inference.

MCMCpack is very much a work in progress. We are interested in demonstrating the user interface, taking a look "under the hood" at some of the code base, and demonstrating three recently added features that increase the power of **MCMCpack** as an estimation engine: the addition of a generic Metropolis sampler for quickly fitting arbitrary models (with the log-posterior density programmed in R), the ability to simultaneously fit models with dispersed starting values and suitable random number generators on clusters of machines, and for some models computation of log-marginal likelihoods to facilitate computation of Bayes factors.

np – A Package for Nonparametric Kernel Smoothing with Mixed Datatypes

Jeff Racine

This package provides a variety of nonparametric kernel methods that seamlessly handle a mix of continuous, unordered, and ordered factor datatypes.

All estimation methods are fully multivariate, i.e., there are no limitations on the number of variables one can model (or number of observations for that matter).

Nonparametric methods include unconditional density (distribution), conditional density (distribution), regression, mode, and quantile estimators along with gradients where appropriate, while semiparametric methods include single index and partially linear models.

A number of tests are included such as consistent specification tests for parametric regression and regression quantile models along with tests of significance for nonparametric regression.

A variety of bootstrap methods for computing standard errors, nonparametric confidence bounds, and bias-corrected bounds are implemented.

A variety of bandwidth methods are implemented including fixed, nearest-neighbor, and adaptive nearest-neighbor.

A variety of data-driven methods of bandwidth selection are implemented, while the user can specify their own bandwidths should they so choose (either a raw bandwidth or scaling factor).

A flexible plotting utility, `np.plot()`, facilitates graphing of multivariate objects. An example for creating postscript graphs using the `np.plot()` utility and pulling this into a LaTeX document is provided.

The function `np.kernelsum()` allows users to create or implement their own kernel estimators or tests should they so desire.

The underlying functions are written in C for computational efficiency. Despite this, due to their nature, data-driven bandwidth selection methods involving multivariate numerical search can be time-consuming, particularly for large datasets. A version of this package using the **Rmpi** wrapper is under development that allows one to deploy this software in a clustered computing environment to facilitate computation involving large datasets.

Using R to Reduce Pesticide Usage in the Horticultural Industry

Lisbeth Riis, Mikkel Grum
Scarab Consult, Nairobi, Kenya

Most growers of high value horticultural produce scout their fields regularly for pests and diseases, but have very limited capacity to make good use of the data. Scarab provides growers' scouts with PocketPCs and GPSs. Each pest and disease observation is geo-referenced and given a timestamp. Data is submitted to a server via GSM. Using R and Latex, we analyse the data to produce reports with maps of every pest and disease in each greenhouse, enabling timely and accurate intervention. Growers can limit their spraying to the affected spots only, or release just the right number of natural enemies in the right place. This enables a shift from interventions based on economic thresholds to immediate intervention whenever a problem is detected, keeping problems small and costs low.

R has a broad range of useful features that have made this possible. While the interactive interface is useful for development, all routine analyses are run as batch files. GPS is not accurate enough on its own, so we use robust linear models to adjust coordinates and remove outliers. Interpolation provides estimates of pest and disease levels between observation points. R's graphics capabilities encourage the use of highly informative graphics. R's database connectivity provides good options for fetching data directly from databases and storing results. Sweave provides fairly flexible automated reporting with Latex.

We are moving the system to the new R spatial foundation classes to make it easier to take advantage of R's spatial capabilities, particularly the geostatistics and graphical capabilities. Other plans include improved analysis of scout performance; analysis of the results of pest and disease control interventions; and use of the new sudoku package to plan scout rotations.

Can R speak your language?

Brian D. Ripley
University of Oxford

"Internationalization" was introduced into R 2.1.0 and has been enhanced in subsequent releases, to the point that the process is pretty much complete in R 2.3.0. The talk will discuss

- handling non-American languages;
- dealing with more than one language at a time;
- using non-English glyphs in graphics - how these changes impact package maintainers.

Bayesian Statistics with Marketing Data in R

Peter Rossi
The University of Chicago Graduate School of Business

By their very nature, marketing problems involve data on a large number of decentralized units. For example, marketing researchers often obtain survey data involving choices between alternative products for a large number of respondents. Demand data on sales of products is collected at the level of the individual product and store. Hierarchical or mixed or multi-level models are often very useful for analyzing these types of data. Analysis of three datasets is illustrated using bayesm, an R package for Bayesian analyses. The examples include a conjoint survey used for product design, a customer satisfaction survey and key account data on retail sales. In addition, some recent developments using Dirichlet Process Priors are briefly discussed.

A package on Robust Kalman filtering

Peter Ruckdeschel

Universität Bayreuth
Mathematisches Intstitut
D-95440 Bayreuth

eMail: Peter.Ruckdeschel@uni-bayreuth.de

WWW: www.uni-bayreuth.de/departments/math/org/mathe7/RUCKDESCHEL

Bernhard Spangl

Group Applied Statistics and Computing
Dept. of Spatial, Landscape, and Infrastructure Sciences
BOKU - Univ. of Natural Resources and Applied Life Sciences, Vienna
Gregor-Mendel-Strasse 33
A-1180 Vienna

eMail: bernhard.spangl@boku.ac.at

WWW: www.rali.boku.ac.at/statedv.html

We want to discuss a proposal on an implementation of Robust Kalman filtering based on `S4` classes. To do so, we are geared to the existing implementations of the Kalman filter from the basic R distribution (cf. [5] and [1]) as well as from the bundle `dse` (cf. [2]). By means of the `setOldClass` mechanism (cf. [5]), we register existing `S3` classes from these implementations as `S4` classes and extend them for our purposes. As generic functions we will present implementations of the classical Kalman filter, the ACM filter from [3], and the rLS-filter from [4].

References:

[1]. Durbin, J. and Koopman, S. J.(2001): *Time Series Analysis by State Space Methods.* Oxford University Press.

[2]. Gilbert, P. (2005): Brief User's Guide: Dynamic Systems Estimation (DSE). Available in the file `doc/dse-guide.pdf` distributed together with the R bundle `dse`, to be downloaded from `http://cran.r-project.org`

[3]. Martin, D. (1979): Approximate conditional-mean type smoothers and interpolators. In *Smoothing techniques for curve estimation, Proc. Workshop, Heidelberg 1979*, Lect. Notes Math. 757, p. 117-143

[4]. Ruckdeschel, P. (2001): Ansätze zur Robustifizierung des Kalman Filters. Bayreuther Mathematische Schriften, Vol. 64.

[5]. R Development Core Team (2005): R:*A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`

RJaCGH, a package for analyisis of CGH arrays with Reversible Jump MCMC

Oscar Rueda[1,2], Ramón Díaz-Uriarte[1]

[1]Bioinformatics Unit, Spanish National Cancer Center (CNIO), Melchor Fernández Almagro 3, Madrid, 28029, Spain

Abstract

Frequently, changes in the number of DNA copies are associated to cancer activity. Gains are related to oncogene activation, and losses to tumor suppressor inactivation. Comparative Genome Hybridization (CGH) is a technique that allows to detect such changes.
RJaCGH is a package that performs analysis of CGH arrays through a non-homogeneous hidden Markov model. It assumes that the true number of DNA copies follows a Markov chain whose transition matrix is dependent of the distance to the next gene. But, due to the fact that what is measured in the array are fluorescent intensities, the states can't be directly observed. Instead, the emissions are modelled through independent gaussian distribution conditional to the hidden states. One of the main problems with this kind of models is the selection of the number of hidden states. Here, we address that issue through a bayesian analysis and the use of Reversible Jump Markov Chain Monte Carlo (RJMCM) techniques. This method allows us to answer several important biological questions, such as estimating the probability of no changes in DNA copy number in a given chromosome, or the probability that a particular gene is gained or lost.

[2]Corresponding author. Email: omrueda@cnio.es

3D Semantic Knowledge Retrieval

Eduardo San Miguel Martín.
Complutense University of Madrid

A work on Latent Semantic Analysis is presented. Upon the basis of semantic similarities a 3d framework for representing knowledge is built. This framework allows representation of words, sentences and whole texts semantic relationship.

LSA is a computational model of human knowledge representation that approximates semantic relatedness judgements. LSA has proven useful in a variety of comprehension and text processing situations. Although, several things are unsolved yet: representation alternatives, similarity measures, semantic bases choices, etc…

This work is an attempt to represent semantic knowledge in a tri - dimensional fashion using R. Due to the multidimensional nature of the semantic space our work has tried two alternatives: first, representing same knowledge from several perspectives at the same time; second, scaling dimension when possible.

LSA algorithm was fully implemented in R, Porter´s stemming algorithm included. Semantic space was built from a collection of Spanish texts. Measures used were cosine, Euclidean distance inverse and vector length.

Results showed that 3d plots are a useful tool in representing Latent Semantic Similarities. This kind of visualization offers a wide range of exploration capabilities.

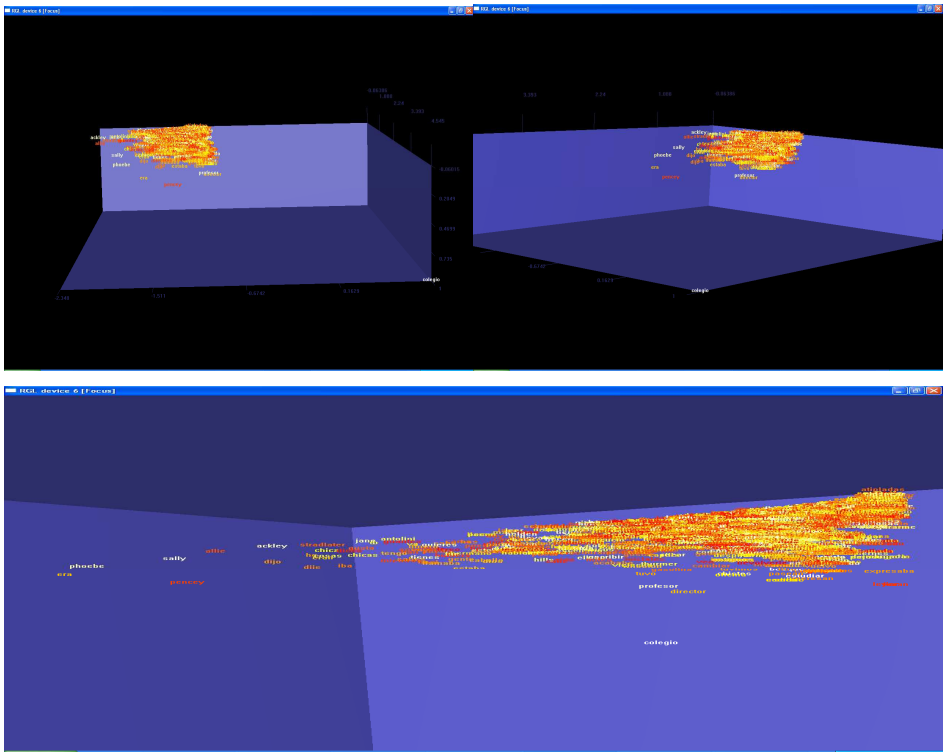


Figure 1.- 3d plot for word “Colegio” (“school”) from a semantic base built from “The catcher in the Rye”

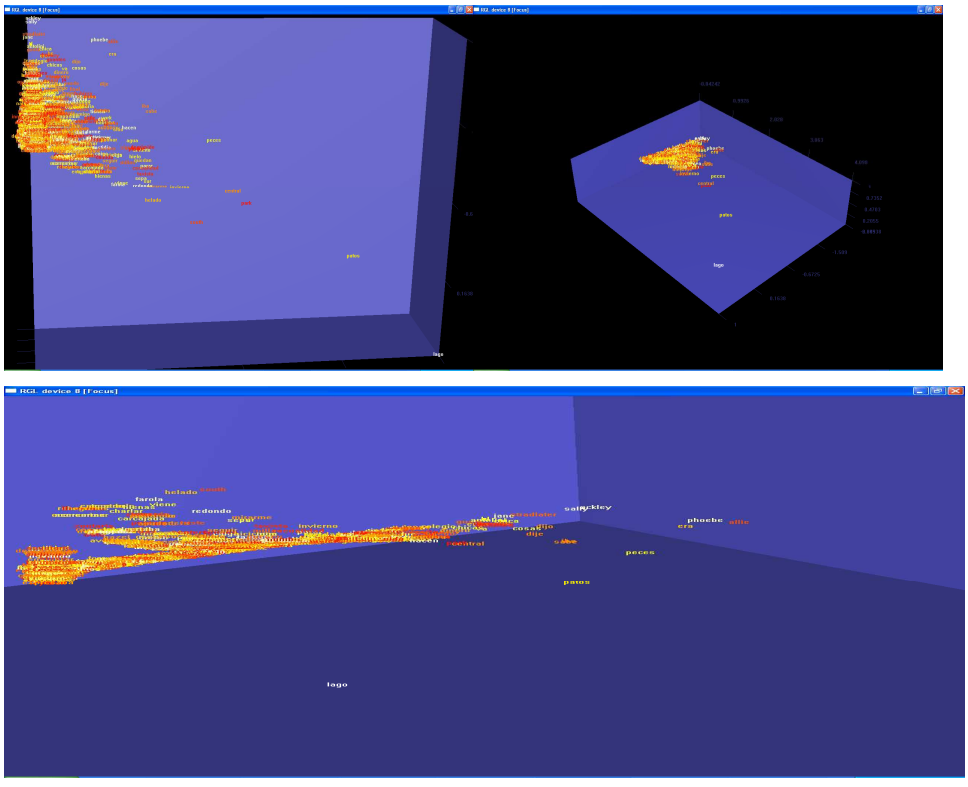
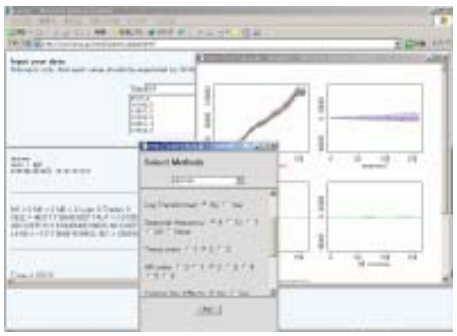


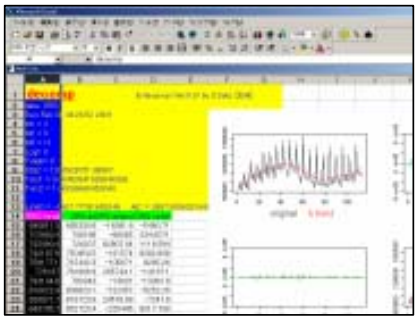
Figure 1.- 3d plot for word “Lago” (“lake”) from a semantic base built from “The catcher in the Rye”.

Web Decomp and E-Decomp – Time Series Analysis using R

Seisho Sato (The Institute of Statistical Mathematics, sato@ism.ac.jp)

In this paper, we introduce “Web Decomp” and “E-Decomp”. Both software are based on “R” or “S” and the core system is same. We developed first “Web decomp” which is web application using S or R. Then we developed E-Decomp which is Excel version of “Web Decomp”. It was easy to re-make the application because R has good abilities in their extension and portability. The both software are shown in my web site “http://www.ism.ac.jp/~sato/”. In Web Decomp and E-Decomp, users can easily apply various time series methods, for example, trend estimation, seasonal adjustment, AR fitting, ARMA fitting and so on.





Web Decomp and E-Decomp

The list of main methods is shown as followings. We note that all method are applied for univariate series.

a. Decomp --- The seasonal adjustment method by using a state space modeling which was developed by Kitagawa and Gersch(1984) , “A smoothness priors --- state space modeling of time series with trend and seasonality, *JASA*, **79**,386,378-389.

b. ARfit --- AR fitting by minimum AIC method

c. ARMAfit --- ARMA fitting. User needs to select an ARMA orders.

d. AutoCor --- Plot autocorrelations of the data.

e. Spectrum --- Plot power spectrum calculated from autocovariances of the data.

The program consists of a GUI part and a computational part. The GUI part is written by JAVA Script(for Web Decomp) or EXCEL-VBA(for E-Decomp). And the computational part is by R(or S) and the DLL which is called from R(or S). These two parts are linked by using CGI program(for Web Decomp) or “R-(D)COM Interface” (for E-Decomp) which was developed by Mr. Thomas Baier.

The rpm package: aligning LC/MS mass spectra with R

Saussen B.[(1)], Kirchner M.[(1,2,3)], Steen H.[(2,3)]
Steen J.A.J.[(2,3)], Hamprecht F.A.[(1)]
[(1)]Interdisciplinary Center for Scientific Computing,
University of Heidelberg, Germany
[(2)]Steen & Steen Lab, Children's Hospital Boston, Boston, USA
[(3)]Systems Biology, Harvard Medical School, Boston, USA

The application of mass spectrometry methods in Systems Biology and particularly in Proteomics is a rapidly evolving and promising field, that provides high-accuracy qualitative and quantitative measurements.

For the LC/MS analysis of complex (e.g. organic) samples, mass spectrometers (MS) are operated in line with liquid chromatography (LC) systems, which uses an organic solvent gradient to separate the sample based on the chemical properties of its constituents before subjecting it to ionization and mass analysis.

Despite the usefulness of the LC separation, this procedure also presents several analytical challenges while processing data. Due to physical properties of the LC process, multiple runs of the same sample as well as comparative runs of different samples suffer from non-linear shifts in the retention time domain, rendering direct comparisons difficult or impossible.

Proper sample registration is also required for samples which do not use the LC separation dimension as one often observes small shifts along the mass/charge domain amongst multiple samples in MS analysis.

We provide an R package called "rpm", which implements the Robust Point Matching Algorithm (RPM) of [Chui and Rangarajan, 2000] which we successfully applied to the registration of real-world MALDI and LC/MS data. With RPM being non-landmark-based, outlier-insensitive and capable of modeling non-linear transformations, we were able to overcome the drawbacks of state-of-the-art methods that are mainly based on piecewise linear alignment of hand-picked landmark peaks and to achieve proper registration of samples.

We consider this work a successful effort to integrate R further into the bioinformatics and proteomics field as a platform for efficient data analysis and prototyping.

mgarch: A Package for the Analysis of Multivariate Garch Models

Harald Schmidbauer and Vehbi Sinan Tunalioglu

In recent years, conditional heteroskedasticity (CH) models have become the workhorse in the study of returns on assets, their purpose being in particular to provide insight into the impact of news on the volatility of asset returns. However, little free software is available for the analysis, in terms of CH models, of multivariate time series. We present a package which tries to make a contribution in this direction. Our package provides elementary functionality to build a synchronized multivariate time series of daily or weekly returns, on the basis of separate univariate level series, which need not be in sync (i.e., different days may be missing). The main part of the package consists of functions which permit the estimation of MGARCH-BEKK and related models, among them a novel bivariate asymmetric model which is capable of distinguishing between positive and negative returns. Diagnostic tools are also included.

Parallel Computing in R using NetWorkSpaces

N. Carriero[(1)], J. Lai[(1)], M. Schultz[(1)], S. Weston[(1)], G. Warnes[(2)]
[(1)]Scientific Computing Associates Inc
[(2)]Pfizer Research Laboratory

Statisticians often encounter computationally intensive problems which can be efficiently processed by splitting the calculations across several computers. The primary barrier to doing this splitting is the complexity of the software tools and programming models traditionally used for handling and managing this splitting.

NetWorkSpaces, developed by Scientific Computing Associates, Inc. (SCAI), provides a simple, but powerful[1], "globally shared namespace" programming model which is very similar to the namespace concept available in R and other interactive programming environments. NetWorkSpaces is implemented via an open-source server plus "adapter" packages for R and other interactive programming environments.

The R 'nws' package (available on CRAN) provides a simple and clean interface for using NetWorkSpaces for data analysis and programming tasks. In particular, it provides a simple mechanism ('sleigh') for splitting a single computation across a set of collaborating machines for processing. This code takes care of all of the details of launching jobs, managing the interaction between jobs, and retrieving all of the results, making it straighforward for even novice R programmers to utilize.

In this presentation present an overview of NWS in R, the supporting system NWS, and several prototypical applications.

[1]The NetWorkSpaces model is exceptionally flexible and powerful despite its simplicity. Unlike many competing approaches, it naturally allows automatic load balancing, dynamic addition and removal of workers, arbitrary communication patterns, and other advanced features.

Managing Large Sets Of Models

Ralf Seger, Antony Unwin
Augsburg University

Abstract

In recent years computer hardware and software has improved to such an extent that it is common to fit large numbers of models, either to ensure the "best" is found or to combine the results from all of them. However, standard software is usually designed for fitting single models, perhaps including some residual analysis and model refinement, but not for analysing many models. Fitting large numbers of models and summarising the results effectively has to be done mainly by hand.
This paper describes the development of a software, Moret, which automatically records and stores the results of models produced in R, providing overviews of all models fitted and ready access to model criteria and estimates for metamodel analyses.

Turing Output of Item Response Theory Data Analysis into Graphs with R

Ching-Fan Sheu and Cheng-Te Chen
Department of Psychology, National Chung Cheng University, Chia-Yi, Taiwan
E-mail: psycfs@ccu.edu.tw

Item response theory (IRT) has become increasingly important for the analysis of measurement data in behavioral research. Many popular software packages for data analysis using IRT focus on parameter estimation and possess only crude graphical capability. However, graphical techniques have been widely recognized as essential tools for data analysis. Plotting enables researchers to gain insights into the relationship between variables, to assess outliers as well as to check model assumptions and so on. Within the framework of item response theory (IRT), person-item map, item characteristic curves (ICCs), and residual plots of fit statistics are three important graphical methods to examine how well the models fit the data and to aid in the interpretation of the features of these fitted models.

The purpose of this presentation is to illustrate the flexibility of using R to create high quality graphs in IRT data analysis. We will present two real data examples to illustrate the implementation of graphing person-item maps, item characteristic curves and residual plots in R.

REFERENCES

1. Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum Associates, Inc.
2. Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299-314.
3. Venables, W.N., Smith, D.M., & the R Development Core Team (2003). *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics* (Version 1.8.0).

Online Applications with Rpad

Tom Short*, Philippe Grosjean†

Rpad (http://www.rpad.org) is an open-source program that provides interactive workbook-type web pages using R as the calculation engine. The R scripts that perform the calculations are embedded in the web page. All required code is embedded in a R package distributed on CRAN. The following topics are discussed:

Server-side functionality. The server uses a modified version of an open-source Perl application called Statistics-R by Graciliano Monteiro Passos. This tool initiates and controls multiple instances of R. A small set of Perl scripts use Statistics-R to pass commands and data from the web server to R. Each webpage gets its own R process and its own temporary directory for reading and writing files. After a definable period of inactivity from the user, the R process shuts down.

Mini-webserver in R using Tcl socket communications. In addition to running Rpad on a dedicated web server, Rpad can also be run from within R using a built-in mini web server inside R. The mini web server is about 500 lines of Tcl code adapted from tclhttpd (http://www.tcl.tk/software/tclhttpd/) and it is used to quickly run a local Rpad server (i.e., no other requirements than to install R and Rpad from CRAN).

Browser-side functionality. Rpad is a type of application based on AJAX technologies. AJAX stands for Asynchronous JavaScript plus XML, but it is used widely to mean any web-based application that dynamically updates portions of the page from the server without refreshing the whole page. When Rpad calculates a page, the browser updates only the results portions of the web page. So, a user can change inputs, hit the calculate button, and the graphs or data or other results update within the existing page. All communication between the browser and the server is plain text. The results from the server are displayed as plain text or as HTML (possibly using functionalities of the R2HTML package). Graphs are normally created as PNG files on the server and passed as an anchor tag to the browser for display within the page. HTML form elements (inputs, checkboxes, and radio buttons) are sent to the server as R variables, which makes it easy to code simple form-type user interfaces.

Browser "editing". Rpad provides different options for the user to view and change the R scripts and other portions of the webpage. Early versions used HTMLArea to provide editing, and the most recent version of Rpad uses the Dojo toolkit to provide editing and user interactivity (http://www.dojotoolkit.org).

Integration of Rpad with a wiki. A Wiki dedicated to R is currently under construction. It will be available before the workshop (probably under an address like http://wiki.r-project.org or http://www.r-project.org/wiki). We plan to integrate Rpad with the Wiki pages to allow interactivity and experimentation on these pages. Two solutions are investigated: (1) direct integration of Rpad in the R Wiki server (or in a secured dedicated Rpad server), and exportation of the wiki pages in Rpad format. These pages can then be edited and viewed locally, using the local mini-webserver in R/Rpad.

Security issues. There is no built in security in Rpad. The user has complete access to any command in R and also to the system shell. For protection, the system needs to be locked down on the server. Write protect any files and databases that are a concern, and lock out access to the server user to other parts of the system. More advanced options to protect against malicious users are to put the server and Rpad components in a chroot jail, use a virtual server to supply Rpad (Xen or VServer), or run the server from a write-protected system disk (like a Quantian DVD).

*EPRI Solutions, Inc.,Schenectady, NY, USA, tshort@eprisolutions.com
†Numerical Ecology of Aquatic Systems, Mons-Hainaut University, Belgium, phgrosjean@sciviews.org

snapCGH (segmentation, normalisation and processing of arrayCGH data) and methods for combining with gene expression information

Mike Smith, John Marioni, Natalie P Thorne, Simon Tavaré
Hutchison/MRC Research Centre, Department of Oncology, University of Cambridge, England

Array comparative genomic hybridisation (arrayCGH) is a technique allowing detection of copy number changes of DNA segments. A microarray is assembled by spotting DNA onto a substrate. Next, test and reference DNA, each labelled with a different fluorochrome, are hybridized to the array. The ratio of the hybridized intensities along the chromosomes provides a measure of the relative copy number between the test and reference samples.

Several statistical methods for the detection of copy number changes (segmentation) have been proposed, but currently the relative merits of each have only been briefly investigated. This is partly due to the disparity between the required formats of input data for the separate methods, which results in, at most only a couple of methods being employed on a given set of data. We have designed a new package (*snapCGH*), which defines a common object type allowing straightforward implementation of several different methods on a single dataset. The *snapCGH* package is also designed to be compatible with the widely used microarray package *limma*, allowing smooth transitions between each stage in the analysis.

The package also implements a new segmentation algorithm, called BioHMM, a heterogeneous hidden Markov model that enables the incorporation of biological data (e.g. distance between probes and probe quality) currently neglected by other segmentation methods. Additionally we have provided the facility to simulate arrayCGH data from multiple experimental platforms to allow more thorough comparisons of segmentation methods i.e. assessment of their performance with different array platforms.

We are also developing ways to incorporate gene expression information with arrayCGH data. These include methods of displaying both types of data alongside each other in meaningful ways and the dynamic retrieval of web-based resources initiated through user interaction with plotting functions.

We will present the methods we have developed for visualising and comparing segmentation methods. Addtionally we will give an overview and example of how users can combine and analyse arrayCGH and expression data together using the *snapCGH* package.

The latest version of the *snapCGH* package is available from:
http://www.bioconductor.org/packages/bioc/1.8/html/snapCGH.html

RpostGIS

an R-library for using PostGIS spatial structures and functions

Norbert Solymosi[1], Andrea Harnos[1], Jenő Reiczigel[1], Ferenc Péter Speiser[2]

[1]Department of Biomathematics and Informatics, Faculty of Veterinary Science, Szent István University, Budapest
[2]Department of Automation, University of Veszprém

Recently, it is more and more widespread in geographical information systems to store the vector graphical maps in databases instead of the former file based solution. This makes it possible to relate the map tables without any interface to the descriptive tables and to access the maps by a large number of users, and even the security of access is getting better.

For the PostgreSQL[1] database server the PostGIS[2] extension based on the OpenGIS *"Simple Features Specification for SQL"*[3] has the above capabilities. Besides of storing the spatial structures, the PostGIS has several functions for handling spatial objects. Beyond that, depending on the installation, the GEOS[4] functions also allows several spatial transformations. So, we can use them through simple ODBC connection on the maps stored in the database.

In the R environment several libraries enabling spatial statistical processes are accessible and offer a great number of functions. But the spatial structures and vector graphical maps are readable only in ESRI shape file formats. Because we could not find such a library that is able to read PostGIS tables, we developed one ourselves.

The RpostGIS library makes it possible to read the maps through the ODBC connection and transformed or generated by the PostGIS and GEOS functions to the R system to apply further operations[5].

The presented package enables direct use of maps based on databases of spatial information systems in the R environment.

1 http://www.postgresql.org/
2 http://postgis.refractions.net/
3 http://www.opengis.org/docs/99-049.pdf
4 http://geos.refractions.net/
5 Norbert Solymosi, Jenő Reiczigel, Andrea Harnos, József Mészáros, László Molnár D., Franz Rubel. Finding spatial barriers by monmonier's algorithm. ISCB 2005, Szeged, 2005.

SHOGUN - A Large Scale Machine Learning Toolbox

Sören Sonnenburg[†], Fabio De Bona[♮],Gunnar Rätsch[♮]

[†] Fraunhofer Institut FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany

[♮] Friedrich Miescher Laboratory, Spemannstr. 35, 72076 Tübingen, Germany

Soeren.Sonnenburg@first.fraunhofer.de,

{Gunnar.Raetsch,Fabio.De.Bona}@tuebingen.mpg.de

Abstract

We have developed an R Interface for our Machine Learning Toolbox SHOGUN. It features algorithms to train hidden markov models and learn regression and 2-class classification problems. While the toolbox's focus is on kernel methods such as Support Vector Machines, it also implements a number of linear methods like Linear Discriminant Analysis, Linear Programming Machines and Perceptrons.

It provides a generic SVM object interfacing to *seven* different SVM implementations, among them the state of the art LibSVM[1] and SVM^{light}[2]. Each of these can be combined with a variety of kernels. The toolbox not only provides efficient implementations of the most common kernels, like the Linear, Polynomial, Gaussian and Sigmoid Kernel but also comes with a number of recent string kernels as e.g. the Spectrum or Weighted Degree Kernel (with shifts). For the latter the efficient linadd[4] optimizations are implemented. Also SHOGUN offers the freedom of working with custom pre-computed kernels.

One of its key features is the "combined kernel" which can be constructed by a weighted linear combination of a number of sub-kernels, each of which not necessarily working on the same domain. An optimal sub-kernel weighting can be learned using Multiple Kernel Learning.[3]

The input feature-objects can be dense, sparse or strings and of type int/short/double/char and can be converted into different feature types. Chains of "preprocessors" (e.g. substracting the mean) can be attached to each feature object allowing for on-the-fly pre-processing.
SHOGUN also supports Matlab™, Octave and Python-numarray. The Source Code is freely available for academic non commercial use under http://www.fml.mpg.de/raetsch/shogun.

References

[1] C.-C. Chang and C.-J. Lin. Libsvm: Introduction and benchmarks. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2000.

[2] T. Joachims. Making large–scale SVM learning practical. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 169–184, Cambridge, MA, 1999. MIT Press.

[3] S. Sonnenburg, G. Rätsch, S. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 2006. accepted.

[4] Sören Sonnenburg, Gunnar Rätsch, and Bernhard Schölkopf. Large scale genomic sequence SVM classifiers. In *Proceedings of the 22nd International Machine Learning Conference*. ACM Press, 2005.

R-ICE - A Modular R GUI

Hutcha Sriplung, Edward McNeil,
Apiradee Lim and Naratip Junsakul

R is an environment which is powerful in statistical computation and graphics. A number of R GUIs have been developed using different approaches and computer languages. Some are limited to a specific platform and a few are available in all main platforms, namely Windows, Linux, and Macintosh OSX.

R-ICE is another R GUI abbreviated from Integrated Computing Environment for R. The main concepts underlying R-ICE are its customizability, open environment, modularity, and platform independence. It is created with the tcltk package within R itself. Customizability means that R-ICE can be modified to speak any human language. Users can select some of its components to fit their works and also create new GUIs to do many things more. Open environment means users can use it, share it, and modify the source code under the basic concepts of the GPL license .It is open for developers to create their own modules and share with others. Modularity means it comprises a number of modules that may or may not be dependent on the others and it is a plug and play environment. Platform Independence means that R-ICE modules are, in fact, R packages that depend on the tcltk library in R.

R-ICE consists of four groups of modules; global, main, associated, and extended. At the moment there is only one global module called *ice*, one main module called *ice.main*, and 6 associated modules, *ice.dataman*, *ice.summary*, *ice.graph*, *ice.statis*, *ice.commands* and *ice.objects*, and two extended modules which are epid, *ice.epid*.

The global module, *ice*, collects additional functions, especially those for data management and summaries, thus, it does not have its own GUI interface. The main module, *ice.main*, is the GUI responsible for basic file and object management, and setting some preferences in the ICE environment. The associated modules deal with basic data frame management, object summary, graphics, basic statistics, and other basic functions. The extended modules are the plug for creativity. Basically, this plug is designed for developers to encapsulate an existing R package with a GUI with the same fashion of menus and dialog boxes. This is demonstrated with the epid and *ice.epid* modules, where epid contains some functions for epidemiology and *ice.epid* is the GUI which calls the epid functions.

R-ICE is open for developers to join. The R-ICE web site is http://www.r-ice.org.

The Generalized Additive Models for Location, Scale and Shape in R

Bob Rigby, Mikis Stasinopoulos [1] and Calliope Akantziliotou [2]

[1] STORM: Statistics, Operational Research and Mathematics research centre, London Metropolitan University, Holloway Road, London, N7 8DB, UK
[2] Bank of Greece, 21 E. Venizelos Ave., Athens, Greece

Abstract

Generalized Additive Models for Location, Scale and Shape (GAMLSS) were introduced by Rigby and Stasinopoulos (2005). GAMLSS is a general framework for univariate regression type statistical problems. In GAMLSS the exponential family distribution assumption used in Generalized Linear Model (GLM) and Generalized Additive Model (GAM), (see Nelder and Wedderburn, 1972 and Hastie and Tibshirani, 1990, respectively) is relaxed and replaced by a very general distribution family including highly skew and kurtotic discrete and continuous distributions. The systematic part of the model is expanded to allow modelling not only the mean (or location) but other parameters of the distribution of y as linear parametric, non-linear parametric or additive non-parametric functions of explanatory variables and/or random effects terms. Maximum (penalized) likelihood estimation is used to fit the models. The algorithms used to fit the model are described in detail in Rigby and Stasinopoulos (2005). For medium to large size data, GAMLSS allow flexibility in statistical modelling far beyond other currently available methods.

The most important application of GAMLSS up to now is its use by the Department of Nutrition for Health and Development of the World Health Organization to construct the worldwide standard growth curves. The range of possible applications for GAMLSS though is a lot more general and examples will be given of its usefulness in modelling medical and insurance data.

In the talk we will describe the GAMLSS model, the variety of different (two, three and four) distributions that are implemented within the GAMLSS package and the variety of different additive terms that can be used in the current implementation. New distributions and new additive terms can be easily added to the package. We shall also discuss the difference of GAMLSS with other available packages in R such as *gam* and *mgcv*. More recent work, for example the inclusion of non linear parameter components as additive terms and the inclusion of truncated distributions and censored data within the GAMLSS family, will be also discussed.

References

Hastie, T.J., and Tibshirani, R.J. (1990) *Generalized Additive Models*. London: Chapman & Hall.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *J. R. Statist. Soc. A*, **135**, 370-384.

Rigby, R.A. and Stasinopoulos, D.M. (2005) Generalized Additive Models for Location, Scale and Shape (with discussion). *Appl. Statist.*, **54**, 1-38.

Variable Selection Bias in Classification Trees and Ensemble Methods

Carolin Strobl, Achim Zeileis,

Anne-Laure Boulesteix and Torsten Hothorn

Carolin.Strobl@stat.uni-muenchen.de

Standard classification tree algorithms, such as CART (Breiman, Friedman, Olshen, and Stone, 1984) and C4.5 (Quinlan, 1993), are known to be biased in variable selection, e.g. when potential predictor variables vary in their number of categories. The variable selection bias evident for predictor variables with different numbers of categories in binary splitting algorithms is due to a multiple testing effect: When potential predictor variables vary in their number of categories, and thus in their number of potential cutpoints, those variables that provide more potential cutpoints are more likely to be selected by chance. This effect can be demonstrated for the rpart routine, which is an implementation of the CART algorithm in R.

Ensemble methods have been introduced to increase the prediction accuracy of weak base learners such as classification trees. However, when biased classification trees are employed as base learners in ensemble methods variable selection bias is carried forward. Simulation results are presented that show variable selection bias for the gbm routine for boosting and for the randomForests routine. Both ensemble methods provide variable importance measures for variable selection purposes that are biased when potential predictor variables vary in their number of categories:

Unsurprisingly, variable importance measures that are based on the individual trees' biased impurity measures are again biased. But also variable importance measures based on the decrease of prediction accuracy after permutation (Breiman, 1998) are biased. This bias can partially be explained by the fact that variables that are preferred in the biased individual trees acquire more influential positions close to the root node and have more effect on the prediction accuracy.

Variable selection bias in individual classification trees can be eliminated by using split selection criteria that account for multiple testing and sample size effects (Strobl, Boulesteix, and Augustin, 2005; Hothorn, Hornik, and Zeileis, 2006). However, empirical experiments suggest that certain resampling schemes, for example the bootstrap used in several random-forest-like ensemble methods, may itself induce preferences towards variables with many splits, even when unbiased classification trees are used as base learners.

We give an overview over sources of variable selection bias in individual classification trees and its effects on variable importance measures in ensemble methods based on classification trees. The underlying mechanisms are illustrated by means of simulation

studies.

References

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics 26(3)*, 801–849.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and Regression Trees*. New York: Chapman and Hall.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics (to appear)*.

Quinlan, R. (1993). *C4.5: Programms for Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc.

Strobl, C., A.-L. Boulesteix, and T. Augustin (2005). Unbiased split selection for classification trees based on the Gini Index. *SFB-Discussion Paper 464, Department of Statistics, University of Munich LMU*.

Remittances and Political Liberalization

Yu-Sung Su*

February 12, 2006

Brief Overview

The paper examines 72 countries from 1977 to 2000, using Bayesian multilevel modeling with R and WinBUGS to test the political effect of remittances on regime change. The paper shows that remittance inflows do enhance the chance of democratization by improving the governance of a regime.

Abstract

Money buys influence. Therefore, that in countries where remittances are important, the political effects shall not be inconsequential. Yet past scholastic works have devoted more effort in studying the economic effects than the political effects of remittances.

In addition, systematic and statistical analyses of the relationship between political liberalization and economic determinants, using pooled cross-national and time-series regression analysis, are rich, though; they are weak at validating their research question cross-nationally and over times.

Therefore to fill in the research gap of the political impact of remittances on the recipient country on one hand, and to improve the generalizability of cross-national analysis of democratization on the other hand, the paper will systematically examine 889 country-year-units, which includes 72 countries from 1977 to 2000, using Bayesian multilevel modeling with R and WinBUGS. The political phenomenon in examining is democratization.

The paper shows that remittance inflows do enhance the chance of an autocracy to democratize. The causal mechanism in behind is that remittance inflows can bring about better regime performance.

*Ph.D. Student of the Graduate Center of the City University of New York. E-mail: ys463@columbia.edu

Stability of Cluster Analysis

Matthias Templ[1,2] and Peter Filzmoser[2]

[1] Statistics Austria, Vienna, Austria
[2] Department of Statistics and Probability Theory, Vienna University of Technology

Abstract

Cluster analysis is a method for finding groups in multivariate data without providing any information about group membership (unsupervised classification). Many different clustering algorithms have been proposed in the literature, and many methods are implemented in R. Unfortunately, for real data sets without obvious grouping structure, different cluster algorithms will in general give slightly different results, sometimes even completely different results. For the user it would thus be important to know which clustering methods are "ideally suited" for analyzing the data at hand.

To start with, one first has to think about preparing the data for clustering. Is it necessary to transform or to scale the data?

Secondly, some cluster algorithms are based on distances. Which distance measure is appropriate? Will the results heavily depend on the distance measure used?

Thirdly, which clustering method should be chosen? Some cluster methods require knowledge on the number of clusters. Will the method and/or the selected number of clusters heavily influence the outcome? Which clustering methods give stable results?

A number of different validity measures have been proposed which help to determine the "correct" number of clusters. Which validity measures are really helpful with this decision?

We will try to provide answers to the above questions using a real data set from geochemistry. A tool in R has been developed which allows a flexible handling of various clustering methods based on different distance measures and evaluated on different validity measures.

Keywords: Cluster analysis, Multivariate methods, Stability, Robustness

Extending interactive statistical graphics

Martin Theus Simon Urbanek

Interactive graphics are often confined to fixed implementation and lack extensibility. On the other hand a programmable environment like R is the ideal platform to provide flexible extensions to existing graphics. The iPlots 2.0 package offers fully interactive, programmable objects such as rectangles, polygons, lines and points, which can be used to build custom interactive graphics with R commands alone. Each such object is linked to underlying data records, and responds in the usual way to interactions like selections, linked-highlighting or queries.

This talk gives an overview of the extensibility of iPlots 2.0 along with several examples, which illustrate the power of custom made interactive graphics. The integration of iPlots with iWidgets and JGR provides a complete toolkit to build interactive, custom applications which use R as a foundation.

Extending BRugs

Andrew Thomas

BUGS is a long running software project aiming to make modern MCMC techniques based on graphical models available to applied statisticians in an easy to use package. This talk will give an overview of the structure of OpenBUGS the open source version of the BUGS software and the tools used in it creation and maintenance. Interfacing BUGS to R will also be discussed in particular the possibilities for closer coupling than currently available in the BRugs package.

Biomarker detection in LC-MALDI mass spectrometry proteomic profiles using R

Grégoire R. Thomas, Sven Degroeve, Koen De Cremer, Filip D'Hondt, Koen Kas, Luc Krols

PEAKADILLY nv, Technologiepark 4, VIB Bio-Incubator, B-9052 Zwijnaarde/Ghent – Belgium

http://www.peakadilly.com/

Reliable biomarker discovery has a dramatic impact on current and future healthcare. Peakadilly's unique proteomics platform allows biomarker discovery in the context of different biological and clinical problems and different biological samples, with an emphasis on serum biomarker discovery.

Peakadilly here describes the robust analysis of LC-MALDI MS data for the discovery of clinical biomarkers using R (http://www.r-project.org/). The biological samples are analysed using COFRADIC™, a powerful gel-free proteomics technology (Nat Biotechnol. 21:566-9, 2003). COFRADIC™ reduces the complexity of an enzymatic digest without loosing information on protein content. This allows sensitive proteome-wide profiling of complex biological and clinical samples.

For each biological sample, a COFRADIC™ analysis generates several hundred spectra each corresponding to an elution fraction from the LC system. The resulting LC-MS profiles are intensity maps representing the expression levels of proteins.

Peakadilly's biomarker discovery pipeline comprises visualisation of LC-MS profiles, feature detection, alignment of the different LC-MS profiles, generation of expression matrices, integration of tandem MS data, and data-mining. The analysis pipelines generates expression matrices derived from thousands of proteins. Candidate biomarkers are extracted by selecting the features that are differentially expressed between groups of samples.

Abbreviations: MS, mass spectrometry; LC, liquid chomatography; MALDI, matrix-assisted laser desorption/ionisation

Bringing transparency to commodity markets in India: A real-world mission-critical deployment of R

Susan Thomas* Shobhana Vyas

February 20, 2006

Abstract

A key input underlying derivatives markets is reliable spot price information. This is used by traders in their trading strategies, for operationalising cash settlement procedures at exchanges, and for risk management at clearing corporations. However, many spot markets are highly non-transparent where the spot price is not readily observed. A particularly difficult case of this nature are the commodity spot markets in India: these markets are spread over 3.3 million square kilometres, and there is little transparency about their operations.

One useful strategy for obtaining a spot price with such market microstructure, is to poll a panel of dealers. However, since dealers can form cartels and attempt to manipulate the polled price, this requires a robust location estimator based on the polled data. As an example, the LIBOR is calculated using a trimmed mean of quotes for the short-term interest rate, polled from dealers at banks.

This paper describes an effort where a system was setup in India in order to create a benchmark price on commodities that traded futures. Within this system, a million phone calls are made a year, covering over 30 commodities, from across multiple market locations, for quotes that are obtained thrice a day. These polled quotes are then used to calculate the 'adaptive trimmed mean' (ATM), which is the robust estimator used to reduce the impact of possible manipulation.

The paper describes the processes through which data is obtained from across the country, how the ATM for each commodity, and the standard deviation of the ATM, is calculated using R and the boot library, and released in near-realtime as the reference spot rate for any given commodity.

This system of reference rates has led to a substantial improvement in transparency of prices, and has made futures trading possible. These reference rates are used for cash settlement of futures traded on the National Commodity Derivatives Exchange (NCDEX), India's largest commodity futures exchange.

*Susan Thomas is Assistant Professor at IGIDR, email:susant@igidr.ac.in, URL: http://www.igidr.ac.in/~susant on the web. Shobhana Vyas is the CIO at the Center for Monitoring Indian Economy (CMIE, URL: http://www.cmie.com), email:shobhana@cmie.com. The views and opinions expressed in this paper are the authors own and not of their employing organisations. Please forward all communication to Susan Thomas.

Robust Location and Scatter Estimators for Multivariate Analysis

Valentin Todorov
Austro Control GmbH
valentin.todorov@chello.at

Abstract: This talk reviews the most popular robust alternatives of the classical multivariate location and scatter estimates and discusses their application in the multivariate data analysis with main emphasis on the availability in R.

Keywords: Robust Estimation, Multivariate Analysis, Minimum Covariance Determinant Estimator, MCD

1 Introduction

The estimates of the multivariate location vector $\boldsymbol{\mu}$ and the scatter matrix $\boldsymbol{\Sigma}$ are a cornerstone in the analysis of multidimensional data, since they form the input to many classical multivariate methods. The most common estimators of the multivariate location and scatter are the sample mean $\bar{\boldsymbol{x}}$ and the sample covariance matrix $\boldsymbol{S}$, i.e. the corresponding MLE estimates. These estimates are optimal if the data come from a multivariate normal distribution but are extremely sensitive to the presence of even a few outliers (atypical values, anomalous observations, gross errors) in the data. If outliers are present in the input data they will influence the estimates $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ and further will worsen the performance of the classical multivariate procedure based on these estimates. Therefore it is important to consider robust alternatives to these estimators and actually in the last two decades much of effort was devoted to development of affine equivariant estimators which have also high breakdown point. Among the most widely used estimators of this type are the Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1985), for which also a fast computing algorithm was constructed -Rousseeuw and Van Driessen (1999), the S-estimators - Davies (1987), Rocke (1996) and the Stahel-Donoho estimator introduced by Stahel (1981) and Donoho (1982) and studied by Maronna and Yohai (1995). If we give up the requirement for affine equivariance, more estimators like the one of Maronna and Zamar (2002) are available and the reward is an extreme gain in speed.

Most of these estimates became available in the popular statistical packages like $S-PLUS$, SAS, $MATLAB$ as well as in R - the packages $MASS$, $rrcov$ and recently $robustbase$. The latter intends to become the "Essential Robust Statistics" R package and strives to cover the upcoming book Maronna *et al.* (2006). Substituting the classical location and scatter estimates by their robust analogues is the most straightforward method for robustifying many multivariate procedures like principal components, discriminant and cluster analysis, canonical correlation and correspondence analysis, hypothesis testing, etc. The reliable identification of multivariate outliers (covered in an other talk) which is an important task by itself, when performed by means of robust estimators, is another approach to robustifying many classical multivariate methods.

The purpose of this talk is to review the most popular estimates of the multivariate location and scatter, to present their implementation in R as well as the accompanying graphical diagnostic tools and to compare this implementation with other statistical packages in terms of functionality and computational time. Further, the application of the robust estimates in the multivariate analysis will be illustrated on several examples - robust linear discriminant analysis, stepwise linear discriminant analysis and robust Hotelling T2 test in package $rrcov$.

References

Davies P. (1987) Asympthotic behaviour of s-estimators of multivariate location parameters and dispersion matrices, *Annals of Statistics*, 15, 1269–192.

Maronna R., Martin D. and Yohai V. (2006) *Robust Statistics*, Wiley, New York.

Maronna R. and Yohai V. (1995) The behavoiur of the stahel-donoho robust multivariate estimator, *Journal of the American Statistical Assosiation*, 90, 330–341.

Maronna R. and Zamar R. (2002) Robust estimation of location and dispersion for high-dimensional datasets, *Technometrics*, 44, 307–317.

Rocke D.M. (1996) Robustness properties of s-estimators of multivariate location and shape in high dimension, *Annals of Statistics*, 24, 1327–1345.

Rousseeuw P. (1985) Multivariate estimation with high breakdown point, in: *Mathematical Statistics and Applications Vol. B*, W.Grossmann G.Pflug I. and W.Wertz, eds., Reidel Publishing, Dordrecht, 283–297.

Rousseeuw P. and Van Driessen K. (1999) A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41, 212–223.

Stahel W. (1981) Breakdown of covariance estimators, Research Report 31, ETH Zurich, fachgruppe fuer Statistik.

Construction of Statistical Models for Hospital Management

Yuko Tsumoto*1 and Shusaku Tsumoto*2
1 Department of Fundamental Nursing,
Shimane University, School of Nursing
9-1 Enya-cho, Izumo 693-8501 Japan
tsumotoy@med.shimane-u.ac.jp
2 Department of Medical Informatics,
Shimane University, School of Medicine
89-1 Enya-cho, Izumo 693-8501 Japan
tsumoto@med.shimane-u.ac.jp

Abstract

It has passed about twenty years since clinical information are stored electronically as a hospital information system since 1980's. Stored data includes from accounting information to laboratory data and even patient records are now stared to be accumulated: in other words, a hospital cannot function without the information system, where almost all the pieces of medical information are stored as multimedia databases[1]. Especially, if the implementation of electronic patient records is progressed into the improvement on the efficiency of information retrieval, it may not be a dream for each patient to benefit from the personal database with all the healthcare information, "from cradle to tomb". However, although the studies on electronic patient record has been progressed rapidly, reuse of the stored data has not yet been discussed in details, except for laboratory data and accounting information to which OLAP methodologies are applied. Even in these databases, more intelligent techniques for reuse of the data, such as data mining and classical statistical methods has just started to be applied from 1990's[2, 3].

Human data analysis is characterized by a deep and short-range investigation based on their experienced "cases", whereas one of the most distinguished features of computer-based data analysis is to enableus to understand from the different viewpoints by using "cross-sectional" search. It is expected that the intelligent reuse of data in the hospital information system provides us to grasp the all the characteristics of university hospital and to acquire objective knowledge about how the hospital management should be and what kind of medical care should be served in the university hospital. This paper focuses on the following two points for analysis. One is what kind of knowledge (statistical models) can be extracted by statistical methods from the datasets stored for about twenty years in Chiba University Hospital. The other is how these pieces of knowledge are useful for the future hospital management and decision support. For construction of statistical models, we applied R to large hospital data because

R gives a wide variety of statistical model construction with nice visualization interface..

The analysis gives interesting results: (1) malignant neoplasm is the first major category which determines the profitability of Chiba University Hospital, which is stable for twenty years. (2) In a global view, the length of stay is the principle factor for the revenue of the hospital, whose distribution follows the log-normal distribution. (3) Treatment method may be a secondary factor to determine the distribution of the length of stay for each disease, which may be correlated with the property that the length of stay follows log-normal distribution for each minor division in total. (4) Treatment without a surgical operation should be more examined by additional information, which is also important to evaluate the profitability of the university hospital.

References

[1] Institute of Medicine Committee on Improving the Patient Record (1997) The Computer-based Patient record: An Essential Technology for Health Care. National Academy Press, Washington DC.

[2] Tsumoto S (2000) Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. Inf. Sci, 124(1-4): 125-137.

[3] Tsumoto S (2001) Chapter G5: Data mining in medicine, In: Kloesgen W, Zytkow J, editors. Handbook of Data Mining and Knowledge Discovery, pp.798–807, Oxford University Press, Oxford.

BAYESIAN COVARIANCE SELECTION IN HIERARCHICAL LINEAR MIXED MODELS

Regina Tüchler, Department of Statistics and Mathematics, WU Vienna, Austria
regina.tuechler@wu-wien.ac.at
Sylvia Frühwirth-Schnatter, Department of Applied Statistics, JKU Linz, Austria
Sylvia.Fruehwirth-Schnatter@jku.at

We present an MCMC algorithm to parsimoniously estimate the random-effects covariance matrix in hierarchical linear mixed models. The definite structure of zero and non-zero elements in the variance-covariance matrix is chosen in a data-driven manner in the course of the modeling procedure. Thereby model selection with regard to fixed versus random effects is automatically included. We specify a straightforward MCMC scheme for joint selection of elements of the random-effects covariance matrix and parameter estimation.

We write the model in the non-centered parameterization, see e.g. [3], which is based on the Cholesky decomposition of the random-effects covariance matrix: $Q = C \cdot C'$, with a lower triangular C. The structure of this model representation allows to identify zeros in the Cholesky factors by common variable selection methods, [2]. This approach is related to ideas of [4], who introduced covariance selection for multivariate normal data.

We contribute to on-going research about random-effects models in various respects. The non-centered parameterization with the above Cholesky decomposition allows us to choose a conditionally conjugate normal prior for C and automatically leads to non-negative definite covariance matrices. A straightforward Gibbs sampling scheme may easily be derived but contrary to the common inverted Wishart prior our new prior is less influential on posterior inference. Existing approaches to estimate a parsimonious variance-covariance matrix are by [4], who select *off-diagonal* elements and by [1], who determine *whole rows and columns* as zero or non-zero. Our choice of the Cholesky decomposition makes it possible to determine zeros and non-zeros for *each* element of the variance-covariance matrix. The ability to specify the finer structure of the random-effects covariance matrix turned out to be of hight importance in real applications with higher dimensional parameters.

New R-code is developed for this method and a real-data example from marketing is given as an illustration.

References

[1] Z. Chen and D. Dunson. Random effects selection in linear mixed models. *Biometrics*, 59:762–769, 2003.

[2] E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection., *Statistica Sinica*, 7:339–373, 1997.

[3] X.-L. Meng and D. van Dyk. Fast EM-type implementations for mixed effects models. Journal of Royal Statistical Society B, 60:559–578, 1998.

[4] M. Smith and R. Kohn. Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97:1141–1153, 2002.

gnm: a Package for Generalized Nonlinear Models

Heather Turner and David Firth

Department of Statistics, University of Warwick, UK

Contact: Heather.Turner@warwick.ac.uk

This talk will introduce the *gnm* package which provides functions for the specification, estimation and evaluation of generalized nonlinear models. The class of generalized non-linear models extends that of generalized linear models by allowing non-linear terms in the predictor. Examples include models with multiplicative interaction terms such as the row-column association models from sociology and the GAMMI models from crop science; stereotype models for an ordered categorical response, and diagonal reference models for dependence on a square two-way classification.

The main functions in *gnm* have been patterned on the *base* function `glm` and its methods (for generalized linear models), so the package integrates well into R and useRs should find it straight-forward to pick up. The package includes some functions that may be used in the context of (generalized) linear models, in particular functions for setting up structured linear interactions between factors. However the major contribution of the package is to provide facilities for the specification and estimation of nonlinear terms. From a user's perspective, this is achieved through two functions: `Mult` for the common case of multiplicative terms and `Nonlin` for any other differentiable non-linear term.

The generality of *gnm* is made possible by two features of the package. First `Nonlin` can be used to specify any differentiable nonlinear term through the use of "plug-in" functions, a number of which are provided by the package and which may also be user-defined. Second an over-parameterized representation of models is used throughout, so that rules for applying constraints do not need to be defined. A set of tools is provided by *gnm* so that estimable parameter combinations and their standard errors can be obtained after a generalized nonlinear model has been fitted.

Although the number of user-level functions in *gnm* is small, the functionality of the package is large and this talk will only provide a snapshot of its capabilities. A more detailed overview is provided by the vignette which is available on the *gnm* webpage (http://www.warwick.ac.uk/go/heatherturner/gnm) or as part of the package itself, which may be downloaded from CRAN (http://cran.r-project.org).

Abstract

Automated Lag Order Selection and Forecasting in VAR modeling
(Svetlana Unkuri)

Estimation and forecasting of univariate time series by means of ARIMA models is a standard tool in nearly every statistical package. Similarly, automated model selection and forecasting for a great deal of time series within an ARIMA framework is straightforward implemented. Focussing on the modeling of multivariate time series, there are also a broad range of statistical software tools for estimating and forecasting multivariate vector autoregressive (VAR) processes (for a given set of underlying time series). Within this talk we present a bundle of R-functions which allow to automate the selection, estimation and forecasting process of VAR models for a large number of different sets of time series. The choice of the lag order can be based on information criteria (e.g. Akaike Criterion, Schwarz Criterion and Final Prediction Error) or on minimizing the forecast errors. Finally, we demonstrate our function by means of economic data.

Key Words: vector autoregressive models, lag order selection, automated lag order determination, automated forecasts.

Estimating survival from Gray's flexible model

Zdenek Valenta

European Center for Medical Informatics, Statistics & Epidemiology
Institute of Computer Science, Academy of Sciences of the Czech Republic

Flexible extension of the Cox proportional hazards (PH) model was introduced by Robert J. Gray in 1992[1]. This extension relies on the inclusion of penalized splines in the Cox PH model which allows for deviating from the assumption of proportionality via incorporating time-varying regression coefficients. In this context the piecewise-constant penalized splines were shown to exhibit more desirable estimation properties than their quadratic or cubic counterparts[1].

We illustrate the use of R function "gsurv.R" for estimating survival from Gray's piecewise-constant time-varying coefficients model[2] which is now part of the "coxspline" R package developed by Gray. The R package is available from the author's website (http://biowww.dfci.harvard.edu/~gray/) and is compatible with the 2.2.1 release of R. We show a few examples of estimating survival from the Gray's model in R using real and simulated data. Using a simulation study we assess the performance of survival estimators based on Cox PH, Aalen's linear and Gray's model under different modeling assumptions[3].

The work on this project was partially supported by the Institutional Research Plan AV0Z10300504.

References:

[1] Gray, Robert J (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association*, 87, 942-951.

[2] Valenta, Zdenek and Weissfeld, Lisa (2002). Estimation of the survival function for Gray's piecewise-constant time-varying coefficients model. *Statistics in Medicine*, 21, 717-727.

[3] Valenta, Zdenek, Chang, Chung-Chou H. and Weissfeld, Lisa A. (2002). Model misspecification effect in univariable regression models for right-censored survival data. *ASA Proceedings of the Joint Statistical Meetings*, 3541-3546.

Accelerating Any EM Algorithm without Sacrificing Simplicity and Stability

Ravi Varadhan[1], Christophe Roland[2], and Hormuzd Katki[3]

The EM algorithm is a ubiquitous computational approach for obtaining maximum likelihood estimates in incomplete data problems. Primary reasons for its popularity are simplicity and stability. However, the stability of EM is usually achieved at a high cost of slow, linear convergence. This limits the usefulness of EM in problems with complex statistical models, high-dimensionality, and large scale data. We have recently (Varadhan and Roland, 2004; Roland and Varadhan 2005) developed two new classes of iterative schemes, Steffensen-type methods for EM (STEM) and squared iterative methods (SQUAREM), to accelerate the convergence of EM. SQUAREM schemes, obtained by ``squaring" the underlying STEM methods, are faster and more efficient. The proposed methods are completely general as they can accelerate any linearly convergent fixed point iteration, and hence, any EM-type algorithm. SQUAREM schemes exhibit either fast-linear or superlinear convergence (3 to 30-fold in our examples). We will illustrate the superior convergence behavior of SQUAREM with a simple binary Poisson mixture example. We will also demonstrate the usefulness of SQUAREM schemes on an important problem in population genetics – the reconstruction of haplotypes from population genotype data.

The proposed schemes are extremely easy to implement, since they work solely with the EM updates. Auxiliary quantities such as complete or observed data log-likelihood, gradient, and hessian are not required. Most importantly, SQUAREM schemes achieve impressive gains in speed without sacrificing the stability of EM. Although the proposed schemes are generally non-monotone, we have developed simple globalization strategies that leverage the base EM iteration to provide enhanced stability. These combined attributes of simplicity, stability, speed, and generality, make SQUAREM methods highly attractive as an off-the-shelf accelerator for any EM algorithm.

We will discuss important computational issues involved in the implementation and use of SQUAREM techniques. An implementation of the new methods is available for general use as an R function.

---

[1] Assistant Professor, School of Medicine, Johns Hopkins University, rvaradhan@jhmi.edu

[2] Post-Doctoral Fellow, Laboratoire Paul Painleve, UFR de Mathematiques Pures et Appliquees-M3, Universite des Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq cedex, France, christophe.roland@math.univ-lille1.fr

[3] Mathematical Statistician, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, katkih@mail.nih.gov

R as integrated engine in blended learning environment

Voirin P., Abou Khaled O., Senn T.

University of Applied Sciences of Western Switzerland, Fribourg, Pérolles 80, CP 32,
1705 Fribourg, Switzerland
(pascale.voirin@hefr.ch)

This paper presents the ongoing work[1] related to the realization of an introductory statistics course produced by a blended learning authoring environment. The main challenge of this work is applying a blended learning approach in the field of statistics using R as a main engine behind all related pedagogical activities which belong to the course materials (self-training exercises, homework, guided exercises, quizzes, etc.). Moreover the production of the course materials using this environment will be open, sharable and compliant with SCORM standard.

The authoring environment supports the tutor/professor during the whole production steps:

1) The design of the guided thread of the course structure: requirements, pedagogical goals, chapter elements, other materials and hyperlinks, etc. using the pacemaker system[2]
2) The realization of related R based interactive self training exercises, homework, guided exercises, and quizzes, based on Rpad
3) The easy deployment of the full course on a course management platform, Moodle in the present case.

This work will be validated, from a pedagogical and technical point of view, through a part of an introductory statistics course for beginners, presenting the concept of hypothesis testing. It will take place during spring semester 2006.
The leading idea is to investigate the pedagogical impact of the use of R engine in such integrated, easy to follow, interactive blended environment, which has mainly to stimulate the motivation of beginner students, not familiar with probability notions.

[1] project granted by Cyberlearn/ n° HES-SO 16043
[2] http://www.eif.ch/projets/eFBS/

Open Source Software in Pharmaceutical Research

Gregory R. Warnes
Max Kuhn
James Rogers

Pfizer Inc., USA
gregory.r.warnes@pfizer.com

2006-02-08

Abstract

Open-Source statistical software is being used with increasing frequency for the analysis of pharmaceutical data, particularly in support of "omics" technologies within discovery. While it is relatively straightforward to employ open-source tools for basic research, software used in any regulatory context must meet more rigorous requirements for documentation, training, software life-cycle management, and technical support.

We will focus on R, a full-featured open-source statistical software package. We'll briefly outline the benefits it provides, as seen from the perspective of a discovery statistician, show some example areas in which it may be used, and then discuss the documentation, training, and support required for this class of use.

Next we will discuss what is needed for organizations to be comfortable with employing open-source statistical software for regulatory use within clinical, safety, or manufacturing. We will then talk about how well or poorly R meets these requirements, highlighting current issues. Finally, we will discuss options for third-party commercial support for R, and evaluate how well they meet the requirements for use of R within both regulated and non-regulated contexts.

The R Genetics Project

Bioconductor foR Genetics

Gregory Warnes[1,2] Scott Chasalow[4] Giovanni Montana[5]
Michael O'Connell[6] David Henderson[6] Nitin Jain[1] Weiliang Qiu[3]
Junsheng Cheng[7] Ross Lazarus[3]

[1]Statistical Applications, Pfizer, Inc.
[2]Department of Computer Science, Yale
[3]Channing Laboratory, Harvard University
[4]Bristol-Myers Squibb
[5]Department of Mathematics, Imperial College
[6]Insightful, Inc.
[7]University of Chicago

Abstract

The R Genetics Project is a collaborative effort to develop a complete set of tools for storing, accessing, manipulating, and analyzing genetics data, from small candidate gene studies consisting of a few genetic markers to large whole genome studies containing hundreds of thousands of markers. The initial goal is to provide a foundation of efficient data structures and easy to use manipulation functions. We intend this foundation to allow methods developers to quickly and easily develop packages implementing their own techniques, while maintaining interoperability. This will reduce the burden on both method developers and applied data analysis, who must currently move data between numerous packages and data formats.

The foundation R Genetics packages, GENETICSBASE, has reached sufficient maturity for introduction to the R community. This talk will describe the R Genetics project, provide an outline of the data structures and features within GENETICSBASE. We will give a brief demo of some of these features, as well as mentioning several additional packages which are building upon this common base, including FBAT, and GENETICSPED.

Supervised Self-Organising Maps

Ron Wehrens, Egon Willighagen,
Willem Melssen and Lutgarde Buydens
Radboud University Nijmegen,
The Netherlands

Self-organising maps (SOMs) provide a means to project a collection of possibly high-dimensional data points to a two-dimensional grid, preserving topology. This means that objects mapped in the same region are similar. SOMs have been used for many different applications, mapping anything from a hundred to hundreds of millions of objects. Typically, the mapping is done in an unsupervised fashion. An example from chemistry is the analysis of the Cambridge Structural Database, which contains almost 400,000 crystal structures. Using a specially designed similarity function it is possible to project these (or a relevant subset) to a SOM. This mapping can be used in several different ways: it can help in identifying the crystal structure of unknown compounds using experimental data, it provides a visual and appealing way of inspecting the database, it has applications in quality control, etcetera.

In this example, the crystal structures are represented by X-ray powder patterns. In the natural sciences one typically has a situation where additional information, either class information or additional continuous variables, is available. In the example of the crystal structures, extra variables are the symmetry class of the crystal structure, the molecular volume, and others. We show how to combine these in a straightforward and consistent way, and present two new forms of the self-organising map that can efficiently incorporate this information in the mapping. A specific advantage of this kind of supervised mapping is the interpretability of the results.

iPlots 2.0

Tobias Wichtrey, Alexander Gouberman,
Martin Theus and Simon Urbanek

The iPlots package brings interactive graphics to R. It was introduced in its first version at the DSC meeting in 2003. Version 0.9 supported only the most elementary plots like scatterplots, histograms and barcharts. The next generation iPlots 2.0 implements the full suite of interactive statistical graphics including multivariate plots such as parallel coordinates and mosaic plots. All plots are fully linked and offer interactive features like sorting/reordering, alpha-blending or custom queries with a user interface consistent across all plots.

iPlots 2.0 are Java-based and the modular approach allows us to offer multiple rendering back-ends. The user can choose to use either pure AWT, Swing or OpenGL. The latter makes interactive analysis feasible even with very large datasets. The programming interface on the R side has further been improved and enables the use of interactive plots as easily as the integrated plots in R.

An implementation of the grammar of graphics in R: ggplot

Hadley Wickham

The R package ggplot is an implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: conditioning and shared axes are handled automatically, while maintaining the ability to build up a plot step by step from multiple data sources. It also implements a more sophisticated multidimensional conditioning system and a consistent interface to map data to aesthetic attributes.

ggplot is built up of four basic building blocks: aesthetic mapping functions, graphic object functions, scale objects and the plot object. Aesthetic mapping functions transform data in aesthetics, graphic object (grob) functions turn lists of aesthetics into grobs, and scale objects ensure that each mapping function operates on a common domain, as well as producing plot guides. The plot object brings these building blocks together to convert a data frame and plot specification into a complete graphic.

Using R as a Web Service

Douglas Wood, Solomon Henry,
David Chang and Balasubramanian Narasimhan,
Data Coordinating Center

In this paper, we describe an implementation of R as a web service using the R-Java/JRI interface. The web service is essentially a Tomcat webapp that listens for requests and acts on them producing output using XML as the lingua-franca for consumption by other web applications. This architecture has been successfully used with several projects at Stanford for generating statistical analyses and graphs on the fly for medical research. We will conclude with some ideas for future work and integration with scalable messaging standards such as JMS, SOAP/SAAJ, and UI enhancements via AJAX.

A framework for heteroskedasticity-robust specification and misspecification testing functions for linear models in R

Achim Zeileis Giovanni Millo

28th February 2006

Specification search strategies in econometric regression modelling are based on zero-restrictions testing on a maintained model (or, in the case of non-nested models' comparison, on an encompassing one). Also in the model validation stage many diagnostic tests may be seen as restriction testing on an auxiliary model derived from the maintained one. Thus, software implementation of both restriction tests and of restrictions-based diagnostic tests may rely on the same *computing engine*. We focus first on robustness of the latter under deviations from the classical normal linear regression model, then on a flexible implementation framework giving rise to a number of functions for mainstream tests.

Heteroskedasticity, a frequent concern in econometrics and most notably in cross-sectional data, invalidates standard restriction testing procedures. Asymptotic tests based on heteroskedasticity-consistent (HC) estimates of the covariance matrix (see White, [3]) are consistent w.r.t. test size, but have poor small-sample properties unless appropriately corrected. Versions of HC restriction tests reliable for use in small samples have been available for about twenty years by now, together with experimental evidence on their performance in an empirical setting and recommendations on their use (McKinnon and White, [2]). Unfortunately, as Long and Ervin ([1]) find out, 15 years later that advice went largely unheeded by practitioners, and the situation is unlikely to have changed much since. These suboptimal habits may be rooted in the unavailability of the appropriate versions of test procedures in many statistical packages, at least without *ad hoc* programming. We discuss the implementation of a range of appropriate testing functions in package `lmtest`.

Base R provides for `lm` objects: a `summary()` method performing partial t-tests and an `anova()` method carrying out F-tests for nested model comparison. Unfortunately, variance-covariance matrix estimates other than the standard (i.e. assuming spherical errors) cannot be plugged in. The functions `coeftest()` and `waldtest()` overcome this problem allowing to plug in estimators, e.g. from package `sandwich`, which provides a general framework for specifying HC and HAC estimators in linear regression models (Zeileis, [4]). In addition, `waldtest()` implements several convenience options for specifying the models to be compared. The computational tools for tests that are based on testing a zero-restriction on an auxiliary model can in turn reuse `coeftest()` and `waldtest()`.

This modular implementation of the general framework allows the researcher

to choose his computing tool at every step of the process consistently with the theory. Moreover, the approach can be easily extended reusing the components for other tests.

We briefly present some montecarlo evidence assessing the performance of a range of Wald, LM and LR zero-restriction tests in small samples in a variety of settings characterized by different degrees of heteroskedasticity and departure from normality in the error terms, designed as in Long and Ervin ([1]). We show that in small samples substantial size bias, usually in the sense of overrejection, may affect some of the HC test versions. The implementation is based on just three functions, the standard `waldtest()` and two slight modifications for LR and LM testing, capable of reproducing all the relevant versions of these test statistics present in the literature by adjusting the optional parameters. We conclude with a sketch of the work in progress.

References

[1] J.S. Long and L. Ervin. Using heteroskedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54:217–224, 2000.

[2] J.G. MacKinnon and H. White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29:305–325, 1985.

[3] H. White. A heteroskedasticity consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica*, 48:817–838, 1980.

[4] A. Zeileis. Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004.