technische universität
dortmund

AASC
Austrian Association
for Statistical Computing

useR! 2008

Abstracts

CRC Press
Taylor & Francis Group
CHAPMAN & HALL

Lufthansa
A STAR ALLIANCE MEMBER

REvolution
computing

MERCK & CO., INC.
Whitehouse Station, N.J., U.S.A.

mangosolutions
data analysis that delivers

statistics.com

WILEY-BLACKWELL

CAMBRIDGE
UNIVERSITY PRESS

Technische
Universität
Dortmund
SFB 475: Reduction of Complexity
in Multivariate Data Structures

StatSoft®

Springer
the language of science

Google™

ZEMENTIS
ADAPA Predictive Analytics Edition

Contents

Spatial Durbin Model for Poverty Mapping and Analysis

Thomas Achia, Atinuke Adebanji, Richard Ng'etich, John Owino and Anne Wangombe

Abstract

The use of spatial regression models for describing and explaining spatial data variation in poverty mapping has become an increasingly important tool. This study considered the spatial Durbin model (SDM) in identifying possible causes of poverty in Bari region of Somalia using Somalia settlement census data. Data properties were identified using exploratory spatial data analysis (ESDA) and the output ESDA provided input into the spatial Durbin model. Parameter estimation and hypotheses testing and assessment of goodness of fit were carried out for the specified model. Dissimilarity of neighbouring settlements in North West Somalia and similarity of neighbouring settlements in North East and South Central Somalia with respect to the variables of interest were observed using the Global and Local Moran's I test statistic. The proportion of families who cannot afford two meals per day was taken as a proxy indicator for poverty level and the implication of the findings on policy decision making for development planning are discussed.

Keywords: spatial regression, spatial Durbin model, poverty
2000 Mathematics Subject Classification: 62M30, 62D05, 62J12

Large atomic data in R: package ’ff’

Daniel Adler Jens Oehlschlägel Oleg Nenadic Walter Zucchini

A proof of concept for the **ff** package has won the large data competition at useR!2007 with its C++ core implementing fast memory mapped access to flat files. In the meantime we have complemented memory mapping with other techniques that allow fast and convenient access to large atomic data residing on disk. ff stores index information efficiently in a packed format, but only if packing saves RAM. HIP (hybrid index preprocessing) transparently converts random access into sorted access thereby avoiding unnecessary page swapping and HD head movements. The subscript C-code directly works on the hybrid index and takes care of mixed packed/unpacked/negative indices in ff objects; ff also supports character and logical indices. Several techniques allow performance improvements in special situations. ff arrays support optimized physical layout for quicker access along desired dimensions: while matrices in the R standard have faster access to columns than to rows, ff can create matrices with a row-wise layout and arbitrary ‘dimorder’ in the general array case. Thus one can for example quickly extract bootstrap samples of matrix rows. In addition to the usual `[` subscript and assignment `[<-` operators, ff supports a `swap` method that assigns new values and returns the corresponding old values in one access operation - saving a separate second one. Beyond assignment of values, the `[<-` and `swap` methods allow adding values (instead of replacing them). This again saves a second access in applications like bagging which need to accumulate votes. ff objects can be created, stored, used and removed, almost like standard R ram objects, but with hybrid copying semantics, which allows virtual `views` on a single ff object. This can be exploitet for dramatic performance improvements, for example when a matrix multiplication involves a matrix and it’s (virtual) transpose. The exact behavior of ff can be customized through global and local `options`, finalizers and more.

The supported range of storage types was extended since the first release of ff, now including support for atomic types `raw`, `logical`, `integer` and `double` and ff data structures `vector` and `array`. A C++ template framework has been developed to map a broader range of signed and unsigned types to R storage types and provide handling of overflow checked operations and NAs. Using this we will support the packed types ‘boolean’ (1 bit), ‘quad’ (2 bit), ‘nibble’ (4 bit), ‘byte’ and ‘unsigned byte’ (8 bit), ‘short’, ‘unsigned short’ (16 bit) and ‘single’ (32bit float) as well as support for (dense) symmetric matrices with free and fixed diagonals. These extensions should be of some practical use, e.g. for efficient storage of genomic data (AGCT as.quad) or for working with large distance matrices (i.e. symmetric matrices with diagonal fixed at zero).

Robust Inference in Generalized Linear Models

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

March 31, 2008

Abstract

The weighted likelihood approach is used to perform robust inference on the parameters in a generalized linear models. We distinguish the case of replicated observations of the dependent variable for each combination of the explanatory variables, common in the design of experiment framework, and the case of one observation for each combination of the explanatory variables, very common in observational studies. We provide some theoretical results on the behavior of the introduced estimators and we evaluate their performance by Monte Carlo experiment. A non exhaustive comparison with the methods already presented in the literature is presented. Illustration of the proposed methods in R is provided by examples on real datasets.

Keywords: Generalized linear models, Outliers in GLM, Residual adjustment function, Robust estimation, Weighted likelihood.

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

mvna, a R-package for the Multivariate Nelson–Aalen Estimator in Multistate Models

A. Allignol[1,2,*] J. Beyersmann[1,2] M. Schumacher[2]

[1]Freiburg Center for Data Analysis and Modelling,
Freiburg University, Germany
[2]Institute for Biometry and Medical Informatics,
Freiburg Medical Center, Germany
[*]`arthur.allignol@fdm.uni-freiburg.de`

Abstract. The multivariate Nelson–Aalen estimator of cumulative transition hazards is the fundamental nonparametric estimator in event history analysis (Andersen et al., 1993, chap. IV). However, and to the best of our knowledge, there is not yet a multivariate Nelson–Aalen R-package (R Development Core Team, 2007) available, and the same appears to be true for SAS and Stata. Therefore, we have developed the **mvna** package (Allignol et al.) with convenient functions for estimating and plotting the Nelson–Aalen estimates in any multistate model, possibly subject to independent right–censoring and left–truncation. The usefulness of this package is illustrated with two important data examples from event history analysis: competing risks and time–dependent covariates, in which displaying estimates of the cumulative transition hazards provides useful insights and straighforwardly illustrates results from standard Cox analyses.

References

A. Allignol, J. Beyersmann, and M. Schumacher. **mvna**: A R-package for the multivariate Nelson–Aalen estimator in multistate models. Submitted to *R news*.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer-Verlag, New-York, 1993.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Keywords

COMPETING RISKS, CUMULATIVE TRANSITION HAZARDS, EVENT HISTORY ANALYSIS, TIME–DEPENDENT COVARIATE

BARD: Better Automated Redistricting

Micah Altman
Institute for Quantitative Social Science
Harvard University
Micah_Altman@harvard.edu

Michael P. McDonald
George Mason University
Brookings Institution

Abstract

BARD is the first (and currently only) open source software package for redistricting and redistricting analysis. BARD is a program that makes political redistricting more accessible and understandable by providing methods to create, display, compare, edit, automatically refine, evaluate, and profile political districting plans. BARD supports both scientific analysis of existing redistricting plans and citizen participation in creating new plans. BARD facilitates map creation and refinement through command-line, gui, and automatic methods. Since redistricting is a computationally complex partitionaling problem not amenable to an exact optimization solution, BARD makes use of a variety of selectable metaheuristics, including genetic algorithms, GRASP, and simulated annealing, that can be use to refine existing or randomly-generated redistricting plans based on user-determined criteria.

Furthermore, BARD supports the ability to randomly generate redistricting plans, and to generate profiles of plans for different scoring weights. This functionality can be used both to explore trade-offs among criteria and to make inferences regarding the intent behind existing redistricting plans.

Because of the computational intensity of these methods, performance is an important criterion in the design and implementation of BARD. The program implements performance enhancements such as evaluation caching, explicit memory management, and distributed computing across snow clusters.

Keywords: redistricting, optimization.

The “deltaR” package: a flexible way to compare regression models on independent samples using a bootstrap approach

Gianmarco Altoè
Faculty of Psychology, University of Padova, Italy
e-mail: gianmarco.altoe@unipd.it

A frequently asked question in the social and behavioral sciences concerns the statistical comparison of different regression models performed on independent samples. This comparison may be useful to: (1) directly compare the goodness of fit of one or more models in independent samples; (2) explore the behavior of different models in different groups in order to conduct more complex analyses (e.g., multigroup analyses).

The aim of this paper is to present a flexible method to test the difference between explained variance (Δ R-squares) of two multiple linear regression models in two independent samples. The method is based upon a stratified, non-parametric bootstrap approach.

The consistency and efficiency of “deltaR” is illustrated via Monte Carlo simulations, and a case study based on real data will be presented. The discussion will focus on the usefulness of this method, with a special emphasis on its applications in the social and behavioral sciences.

Automatic construction of graphical outputs of common multivariate analyses with a special reference to predictive biplots

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, 4900–348 Viana do Castelo, Portugal, mruialves@estg.ipvc.pt

Predictive biplots [1] have several aspects which are very important in multivariate analyses, mainly because they enable an easier interpretation of multivariate analyses outputs by relating sample configurations to the initial, declared variables, without losing the modulation aspects so characteristic of these statistical methodologies.

The disadvantages of biplots are mostly related to software limitations causing difficulties in obtaining the final graphical solutions and to difficulties in deciding how relevant a given biplot axis is and how many plots are necessary to accurately describe available data.

The latter difficulties are also experienced by users of normal statistical methodologies when it comes to interpret multivariate outputs, i.e., how important initial variables are to explain latent variables and how many dimensions are necessary to describe data. One problem is that there is a huge degree of freedom in the way this can be done, and easy methods to help taking these decisions are very important, mainly for inexperienced statisticians.

This presentation follows our previous works on the subject [e.g., 2,3] and demonstrates how a decision on whether a biplot axis is or not drawn in two-way plots can be carried out automatically [4]. The method, based on the predictive power of variables and on a specially defined tolerance value, enables R to evaluate the interest of each of the initial variables and draw the predictive biplots automatically. Also, the final number of plots is also automatically decided by R. Since the method may be made fully automatic, inexperienced users can take profit of all the R facilities, carrying out multivariate analysis and final interpretations, and at the same time being protected from common over-fitting problems, difficulties in interpretations of multivariate outputs, etc.

The methods can be applied to several statistical methodologies, and examples are provided for principal components analysis and canonical variate analysis (in chemistry) and three-way Tucker-1 common loadings analysis (in the field of sensory analysis).

It is also shown that the method devised to produce biplots in an automatic way can be diverted to common outputs, enabling R to provide users of multivariate analyses with automatic interpretations of results, including the decision on the number of dimensions to retain and their respective interpretations.

[1] Gower, J.C.; Hand, D.J. Biplots. Chapman and Hall: London, 1996.

[2] Alves, M.R.; Oliveira, M.B. (2005). Monitorization of Consumer and Naïf Panels in the Sensory Evaluation of Two Types of Potato Chips by Predictive Biplots Applied to Generalized Procrustes and 3-way Tucker-1 Analysis. . *Journal of Chemometrics*, **19**: 564–574.

[3] Alves, M.R.; Oliveira, M.B. (2006). R algorithms for the calculation of markers to be used in the construction of predictive and interpolative biplot axes in routine multivariate analyses. User!2006 – The R user Conference, Wirtschaftsuniversität Wien.

[4] Alves, M.R.; Oliveira, M.B. A method for the production of predictive biplots applied to multivariate outputs with a complete R algorithm enabling fully automatic processes and user defined parameters. Submitted.

R in Automation: Accessing Real-time-data in PLCs

Thomas Baier*

March 29, 2008

Industrial Automation in general and in particular PLCs (Programmable Logic Controllers) and embedded devices are a rapidly growing market. Embedded devices are found in small devices, like, e.g., watches or mobile phones, are used in everyday life as for example ABS systems or engine-monitoring systems in cars. In larger applications these system are typically called PLCs and used to control assembly lines, rolling mills or power plants.

Depending on the requirements on availability of automation systems on the one hand or safety considerations on the other hadn, more and more effort is put into monitoring the system during its whole life time.

Typical aproaches for monitoring are either rule-based systems or open-loop control scnearios. In rule-based systems data is collected and processed according to statically defined rules (e.g., issuing an emergency shutdown if some safety-related devices fails). Open-loop are designed to collect data and present the results to an operator. The operator then has to decide on further actions (or if the operator fails to acknowledge an alarm message, an automatic procedure brings the whole automation system into a fail-stop or fail-safe operation mode).

In addition to these methods, we are suggesting an alternative method to capture the “big picture” of the automation device and allow to apply statistical methods on the process data. This analyses will be the input for further optimization of operation of the automation system in better planning of device/system maintenence (so-called “predictive maintenance”).

Fortunately, the automation industry has decided on standard means for data acquisition which is typically used by visualization and data collection software. OPC (formerly known as *OLE for Process Control*, nowadays OPC is marketed as *Openess, Productiviy and Collaboaration*) provides standardized mechanisms for accessing real-time data. This data can either be a PLC’s or embedded device’s internal state or “real” process data from the sensor/actor level (e.g., state of switches, valves or drives).

In our short presentation we will show how R can access this data using OPC DA (OPC Data Acquisition), which allows to connect to nearly every PLC or embedded device used in automation industry. OPC DA is based on Microsoft’s COM technology, which currently is easiest to use with R for Windows. In addition to OPC DA we will shortly discuss current developments in the field of OPC which will also enable to access OPC data from non-Windows systems.

*logi.cals by kirchner SOFT, Mailüfterlweg 1, 3124 Oberwölbling, Austria, `http://www.logicals.com/`

RSTAR: A Package for Smooth Transition Autoregressive Modeling Using R

Mehmet Balcilar
Department of Economics, Eastern Mediterranean University
Famagusta, North Cyprus
Office Tel: +90 (392) 630 1548, Fax: +90 (392) 365 1017
e-mail: ⟨mehmet.balcilar@emu.edu.tr⟩

In the last few years, numerous improvements have been made for statistical inference in threshold autoregressive models. Particularly, new tests are developed and methods are proposed for diagnostic control, forecasting, and impulse response analysis. These developments are examined in Granger and Terasvirta (1993), Terasvirta (1998), Potter (1999), and van Dijk, Terasvirta and Franses (2002). This study develops a comprehensive R package for testing, estimating, diagnostic checking, forecasting, and further analysis of smooth transition autoregressive models (STAR). The package is designed around the empirical modeling cycle for STAR models devised by Terasvirta (1994), and van Dijk, et al. (2002). This modeling approach consists of specification, estimation and evaluation stages and, thus, is similar to the modeling cycle for linear models of Box and Jenkins (1976). In the testing stage, the package emphasis LM-type test and implements all tests proposed in the literature (see Luukkonen, Saikkonen and Terasvirta (1988), Granger and Terasvirta (1993), van Dijk, et al. (2002)). The package allows estimation of logistic and exponential STAR models using analytical gradients. Very extensive diagnostic control techniques are implemented in the package. All aspects of the diagnostic tests examined discussed in Eitrheim and Terasvirta (1996), van Dijk and Franses (1999) and Lundbergh, Terasvirta and van Dijk (2000) are fully implemented. The R-STAR package allows robust estimation methods for all tests in order to guard against influence of possible outliers. R-STAR emphasizes aspects such as model evaluation by means of out-of-sample forecasting and impulse response analysis, and the influence of possible outliers on the analysis of smooth transition type nonlinearity. Forecasts and impulse responses are calculated using Monte Carlo or bootstrap methods with code highly optimized for speed. We also incorporate recently introduced extensions of the basic smooth transition model. On the programming side, R-STAR needs no programming experience. Although all commands have control over all aspects, default values are provided and only one or two options need to be passed, if needed at all. We take advantages of object-oriented programming, S4 methods, and vectorization provided by the R environment.

References

Box, G.E.P. and G.M. Jenkins (1976), Time Series Analysis; Forecasting and Control , San Francisco: Holden-Day.

Eitrheim, O. and T. Terasvirta (1996), Testing the adequacy of smooth transition autoregressivemodels, Journal of Econometrics 74, 59-76.

Franses, P.H. and D. van Dijk (2000), Nonlinear Time Series Models in Empirical Finance , Cambridge: Cambridge University Press.

Granger, C.W.J. and T. Terasvirta (1993), Modelling Nonlinear Economic Relationships , Oxford: Oxford University Press.

Lundbergh, S., T. Terasvirta and D. van Dijk (2000), Time-varying smooth transition autoregressive models, Working Paper Series in Economics and Finance No. 376, Stockholm School of Economics.

Luukkonen, R., P. Saikkonen and T. Terasvirta (1988), Testing linearity against smooth transition autoregressive models, Biometrika 75, 491- 499.

Potter, S.M. (1999), Nonlinear time series modelling: an introduction, Journal of Economic Surveys 13, 505-528.

Terasvirta, T. (1994), Specication, estimation, and evaluation of smooth transition autoregressive models, Journal of the American Statistical Association 89, 208-218.

Terasvirta, T. (1998), Modelling economic relationships with smooth transition regressions, in A. Ullah and D.E.A. Giles (eds.), Handbook of Applied Economic Statistics , New York: Marcel Dekker, pp. 507-552.

van Dijk, D., T. Terasvirta and P.H. Franses (2002), Smooth transition autoregressive models - a survey of recent developments, Econometric Reviews 21, 1-47.

Tree-based and GA tools for optimal sampling design

Marco Ballin, Giulio Barcaroli
(ballin@istat.it, barcarol@istat.it)
ISTAT, via Cesare Balbo 16, 00184 Roma Italy

Keywords: sampling design, multivariate allocation,
optimal stratification, genetic algorithm, multivariate allocation

The optimality of a sample design can be defined in terms of costs (associated to fieldwork: number of units to be interviewed) and accuracy (sampling variance related to target estimates). Bethel proposed an algorithm (Bethel, 1985) able to determine total sample size and allocation of units in strata, so to minimise costs under the constraints of defined precision levels of estimates, in the multivariate case (more than one estimate). Input to this algorithm is given by the information on distributional characteristics (total and variance) of target variables in the population strata. Under this approach, population stratification, i.e. the partition of the sampling frame obtained by cross-classifying units by means of potential stratification variables, is given. But stratification has a great impact on the optimal solution determined by Bethel algorithm and, in general, it must be defined in the first steps of a survey planning.

If a frame with a set of potential variables for stratification is available, the survey planner has to choose the "best" auxiliary variable cross product (partition of the frame). Among the possible partitions, the one with the maximum number of strata, given by the Cartesian product of all auxiliary variables, does not always yield the optimal sample size. In fact, organisational considerations, and the necessity to define a minimum amount of units per stratum, oblige not to increase the number of strata beyond a certain limit. In that case, how to determine the best partition among all partitions obtainable combining the auxiliary variables (what auxiliary variables? what values for each of them to take into consideration?) has to be considered as a part of the whole problem.

Until recently, on the contrary, the problem of determining the optimal size and allocation of units in strata has been solved considering the stratification of population as given; and, conversely, the definition of an optimal stratification has been investigated independently by the optimisation problem of sampling size and allocation.

An interesting proposal has been advanced in the recent past (Benedetti et. al 2005), offering a joint solution to both problems: it is based on a tree search in the space of possible strata configurations, solving for each visited node the corresponding multivariate allocation problem accordingly to Bethel algorithm. At each level, the node that is the best in terms of sample size reduction, is chosen as the branching node. This tree-based approach is deterministic and very fast, but it may heavily suffer for the presence of local minima and, consequently, solutions can be far from optimality.

Together with this tree-based approach, we propose a non deterministic evolutionary approach, based on the genetic algorithm (GA) paradigm. Under the GA approach, each solution (i.e. a particular partition in strata of the sampling frame) is an individual in a population, whose fitness is evaluated by calculating the sampling size satisfying accuracy constraints on the target estimates; crossover and mutation carried out along each iteration ensure an increase of average fitness.

In general, the characteristic of GA are such that the risk of local minima is lower than in the tree search, though processing time is noticeably higher. Our proposal is the following: in complex situations (characterised by a high number of stratification alternative configurations and/or a high number of target variables and domains), first the tree-based algorithm is applied, in order to individuate a solution. This solution is then introduced in the GA initial population, in order to speed its convergence to a better solution. Our experiments show an improvement of the tree-based solution, and encourage the adoption of this procedure.

The whole system can be thought of as a "toolkit", composed by a series of instruments, all implemented and operating in the R environment. Main scripts are:

1. strataTree.R implementing the tree-based algorithm;
2. strataGenalg.R that implements the GA approach, making use of "genalg" package (Willighagen, 2002) in a slightly modified version;
3. Bethel.R implementing Bethel algorithm.

These programs can be run directly in the R environment, but, as and an additional facility, a simple web interface has been developed using Rwui that enables the user to carry out the processing without being obliged to be acquainted with R language or even R environment.

References

Benedetti R., Espa G., Lafratta G. (2005), "A Tree-based Approach to Forming Strata in Multipurpose Business Surveys", *Discussion Paper No 5, 2005*, Università degli Studi di Trento – Dipartimento di Economia

Bethel J. (1985), "An Optimum Allocation Algorithm for Multivariate Surveys", in *American Statistical Proceedings of the Survey Research Methods Section*, pp. 209-212

Willighagen E. (2005), "*genalg: R Based Genetic Algorithm*". R package version 0.1.1. URL http:// cran.r-project.org/

ArDec: Autoregressive-based time series decomposition in R

Susana M. Barbosa

Universidade do Porto, Faculdade Ciências

The extraction of trend and periodic components from an observed time series is a topic of considerable practical importance. Most time series methods require the assumption of stationarity to be met, and therefore the removal of any trend-like or seasonal signals from the data. Furthermore, in many applications such signals are often of interest in themselves. Flexible methods are therefore required for the decomposition of a time series into physically-relevant components. The R package ArDec implements the autoregressive-based time series decomposition of West (1997). The method is based on the dynamic linear representation for an autoregressive process from which results a constructive approach for the decomposition of an observed time series into latent constituent sub-series. The approach and the usage of package ArDec are illustrated through an example of decomposition of sea-level time series.

Visualizing multivariate categorical and continuous data from epidemiologic studies: An expanded scatter plot matrix

Benjamin Barnes, Karen Steindorf

German Cancer Research Center, Heidelberg, Germany

Epidemiologic datasets often contain a mix of categorical and continuous variables. Understanding the interrelationships among these variables is vital for subsequent analysis and can be aided by graphical presentation. Scatter plot matrices, produced in R using functions such as pairs() and splom(), are useful for graphically displaying multivariate continuous data. For displaying multivariate categorical data, the Visualizing Categorical Data (vcd) package offers many flexible options, including the pairs.table() function. However, these functions are not readily compatible with one another, making visual presentation of mixed epidemiologic data difficult. With this in mind, the scope of the splom() function was expanded to include visualization of categorical data. Furthermore, a novel panel function compatible with splom() was created to visualize categorical-categorical data using a mosaic plot. Continuous-continuous data was plotted using existing scatter plot and level plot panel functions. Existing panel functions were also used to produce box-and-whisker plots for categorical-continuous data as well as stacked bar charts for categorical-categorical data. With these modifications and the new mosaic function, categorical and continuous data can be viewed in a unified plot matrix. An example of such a plot matrix was created using simulated data inspired by a study investigating the effects of lifestyle and anthropometric factors on insulin-like growth factor (IGF)-I and IGF binding protein (IGFBP)-3. These two proteins are suspected of playing a role in breast cancer development, and current research focuses on identifying modifiable lifestyle factors that influence their concentrations in blood. The expanded scatter plot matrix described here improves visualization of mixed datasets and can be further enhanced to visualize bivariate linear models, chi-squared tests, and other bivariate statistical test results.

Understanding product integration

Jan Beyersmann[1,2], Arthur Allignol[1,2] and Martin Schumacher[2]

[1] Freiburg Centre for Data Analysis and Modelling, University of Freiburg, Eckerstraße 1, 79104 Freiburg, Germany, jan.beyersmann@fdm.uni-freiburg.de
[2] Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Straße 26, 79104 Freiburg, Germany

Abstract. Product integration is a very powerful, but somewhat neglected topic in applied survival analysis: Survival data are usually incompletely observed, the most important example being independent right-censoring. This leads to survival analysis being based on hazards, because the hazard of seeing an event is undisturbed by censoring. The Kaplan-Meier estimator of the survival function is a finite product over one minus empirical hazards, and it approaches e to the negative true cumulative hazard. This result is not very intuitive, but it is much better understood using product integration: A product integral is a 'continuous time product', like a usual integral is a 'continuous time sum'. The product integral over one minus the true hazard is a 'product' over infinitesimal conditional survival probabilities, and therefore equal to the survival probability. The product integral over one minus the empirical hazard equals the Kaplan-Meier estimator. The 'e to the negative true cumulative hazard'-formula is then seen to simply be the solution of a product integral. We explore these connections in R, where one function `prodint` both approximates the true survival function arbitrarily close and results in the Kaplan-Meier estimate when applied to data based empirical hazards.

Both theory and the R implementation generalize to the matrix-valued case that is important for multistate models: Here one individual may experience a possibly random number of events, e.g. transitions between 'healthy' and 'ill' before dying. Closed formulae for the transition probabilities will in general not be available anymore, but a matrix-valued version of `prodint` may still be used for numerical approximation. It also results in the Aalen-Johansen estimator of the matrix of transition probabilities, a generalization of the Kaplan-Meier estimator, when applied to empirical transition hazards. Empirical transition hazards can be obtained in R using the `mvna`-package.

References

Gill, R and Johansen, S (1990): A survey of product-integration with a view towards application in survival analysis. *Ann Stat, 18, 1501–1555.*

FluxEs: An ‘R’ Framework for Parameter Estimation in Biological Networks

Thomas Binsl[1], Jaap Heringa[1], David Alders[2], and Hans van Beek[2]

1) Centre for Integrative Bioinformatics, VU University Amsterdam, 1081HV Amsterdam, The Netherlands

2) VU University Medical Centre Amsterdam, 1081HV Amsterdam, The Netherlands

Abstract

Parameter estimation in biological networks is a difficult task and many computer programs were developed for this purpose. However, available computer methods suffer from lack of easy implementation of new biological pathways. Hence, we have designed a new framework called *FluxEs* using an object-oriented programming approach, implemented using S4 classes in R. It particularly addresses distribution of isotopes between metabolites in a carbon-transition network useful for quantifying metabolic fluxes. The developed package provides a simple way to specify the topological information of the network as well as the precise transitions of carbon atoms between molecules in plain text files, and guides the user through the optimization process. For the purpose of parameter estimation, *FluxEs* automatically derives the mathematical representation of the formulated network, and assembles a set of ordinary differential equations (ODEs). Afterwards, it fits experimentally measured Nuclear Magnetic Resonance (NMR) multiplet intensities with the metabolic model result, by continuously solving the ODEs numerically, scanning parameter space to obtain optimal parameter estimates. A test was performed by applying *FluxEs* to fit a model of the tricarboxylic acid (TCA) cycle to simulated ^{13}C NMR data, including realistic measurement noise. Flux values could be re-estimated with significant precision. Subsequent flux estimation on experimental NMR data of animal heart biopsies showed good correspondence with independent chemical measurements.

Towards a Java Framework for Rapid Development of Graphical User Interfaces for Statistical Applications based on R

Bernd Bischl
Fakultät Statistik, TU Dortmund
bernd.bischl@uni-dortmund.de

Kornelius Rohmeyer
Institut für Biostatistik, LUH
rohmeyer@biostat.uni-hannover.de

Abstract

Many users from a non-statistical background are not programmers and often are not up to the task of using R for their statistical problems. Therefore, specific and intuitive applications need to be provided, which hide much of the complexity of the underlying R system, in order to enable the users to solve their problems at hand.

While it is possible to either control R from different programming languages or to interface Java or C++ from R, it is not very efficient to create the above mentioned applications from scratch by any of these alternatives. Because many characteristics are shared, these should also be encapsulated in shared code. Hence we believe that a toolbox is necessary which helps the developers of statistical software to conveniently and flexibly design graphical applications in their area of expertise.

Our open source and platform independent framework, which is implemented in Java and builds upon JRI (http://www.rforge.net/JRI) and Rserve (http://www.rforge.net/Rserve), aims to achieve that goal by inserting an abstraction layer between the business logic of the application and these two packages. Thereby we can create the same application as a local variant (employing the user's already installed version of R) or as a web oriented application with minimal local requirements (which is automatically installed via Java Webstart and performs all computations on an R server, thus not forcing the user to have R installed at all).

Currently, there are utility classes and respective GUI elements to import data from XLS or CSV files, create dialogs to perform statistical analysis, generate, display and save plots and print output to PDF files. We also provide a basic LaTeX support for tables.

Our framework has evolved from two major projects: One application to estimate dose-response models for the University of Copenhagen and one for the Leibniz Universität Hannover to do quality assessment and novel statistical analysis for toxicological data. The last one also includes convenience classes to generate dialogs regarding different assays of toxicology. These elements either act as a tutorial or provide a guided walk through the analysis of the user's own dataset.

In our presentation we will compare our own approach to existing frameworks for building statistical tools based on R by highlighting their general advantages and disadvantages. A short demo of the framework and the look-and-feel of an implemented application will be given. We are looking forward to receiving feedback and discussing further features and improvements with attending researchers, users and developers from the field.

Simulating Games on Networks with R. Application to coordination in dynamic social network under heterogeneity

Michał Bojanowski
ICS and Department of Sociology
Utrecht University
m.j.bojanowski@uu.nl

April 1, 2008

Most of the existing theoretical contributions to understanding mechanisms of co-evolution of social networks and individual behavior assume that actors are homogeneous (e.g. Buskens et al., 2008; Jackson and Watts, 2002). The consequences of relaxing this assumption (Galeotti et al., 2006) are not yet fully understood. Under which conditions will the differences between actors result in higher segregation levels than in the homogeneous case? In this paper we study the interrelated dynamics of social networks and behavior when actors' interests differ. As a framework for analysis we propose a baseline model in which actors simultaneously choose their behavior and manage their personal relations with others. The population of actors is composed of two types and interactions are modeled with asymmetric two-person games. The heterogeneity is represented by three elements:

1. The degree to which actors' interests behavioral options differ.
2. The severance of "mis-coordinating".
3. Complementary or substitutable character of relations with actors of the other type.

To address the posed problems and evaluate the role of the three above mentioned components we employ both analytical and computer simulation methods.

This paper presents the results of computer simulation study prepared and executed in R. The implementation relies on the framework proposed in package simecol (Petzoldt and Rinke, 2007) which was fine-tuned for use in our setting.

The results identify stable network architectures that emerge if actors actively try to improve their position by making behavioral and relational choices. We also investigate the dynamics of selected structural characteristics which, among other, include network segregation, centralization and transitivity. Examples of the dynamics of the system are shown with network visualizer SoNIA[1].

References

Vincent Buskens, Rense Corten, and Jeroen Weesie. Consent or conflict: Coevolution of coordination and networks. *JPR*, 45(2), 2008.

Andrea Galeotti, Sanjeev Goyal, and Jurjen Kamphorst. Network formation with heterogeneous players. *GEB*, 54:353–372, 2006.

M. Jackson and A. Watts. On the formation of interaction networks in social coordination games. *Games and Economic Behavior*, 41:265–291, 2002.

Thomas Petzoldt and Karsten Rinke. simecol: An object-oriented framework for ecological modeling in r. *Journal of Statistical Software*, 22(9):1–31, 2007. ISSN 1548-7660. URL http://www.jstatsoft.org/v22/i09.

[1] http://www.stanford.edu/group/sonia/

Using R for time series analysis and spatial-temporal distribution of global burnt surface multi-year product

Jedrzej Bojanowski
`jedrzej.bojanowski@jrc.it`

César Carmona-Moreno
`cesar.carmona-moreno@jrc.it`

European Commission - Joint Research Centre
Institute for Environment and Sustainability
Global Environment Monitoring Unit
TP 440, 21020 - Ispra (Va), Italy

Fires are one of the most significant components in the workings of the global ecosystem. There is no doubt that the global fires regime has a major influence on climate, carbon cycle, pollution, etc. Modeling those phenomena has been complicated because of the lack of exhaustive databases concerning past fires distribution. For this reason, JRC is working on the concatenation of two existing independent global multi-year burnt area products: GBS (1982-1999) and L3JRC (2000-2007). Since both time series are produced using different satellite data with different spatial and temporal resolution and algorithms, the main objective is to develop a statistically coherent database. Combining GBS and L3JRC products requires dissecting both of them - analyzing their variations in time and space.

In this paper, we present a few R applications which were applied in our research. `RNetCDF` package was implemented to handle the big amount of spatial data. Afterwards, we applied different predefined methods for time series analysis, using those from `stats` package, as well as from the `zoo`, `tseries` or `lmtest` packages. We introduce several decomposition theorems, tests, stochastic models as well as some graphics dedicated to time series objects.

The principal components analysis allows us the description of the differences in fires regimes derived from GBS and L3JRC algorithms. Based on this, we propose a visualization technique to evaluate the spatial temporal coherence of the 26 years of these global burnt area products. `Rgl` package was used to present the data in principal components space.

We also carried out some new methods to present spatial-temporal distribution of the data. We used variogram analysis and kriging method based on `spatial` package.

A Maximum Likelihood estimator of a Markov model for disease activity in chronic diseases that alternate between relapse and remission, for annually aggregated partial observations.

Sixten Borg. The Swedish Institute for Health Economics (IHE), Box 2127, 220 02 Lund, Sweden. Email: sb@ihe.se.

Background

Crohn's disease (CD) and ulcerative colitis (UC) are chronic inflammatory bowel diseases that have a remitting, relapsing nature. Relapses are treated with drugs or surgery. No drug can be considered a curative treatment. In CD, surgery is not curative, and may need to be performed many times, since the disease may reappear. For UC, curative surgery is possible after which it cannot relapse again.

We needed a discrete-time Markov model for the disease activity of relapse and remission with a cycle length of one month, in order to study the effect of shortening or post-poning relapses. Our data consisted of yearly observations of the individual patients. Each year, the number of relapses and surgical operations were recorded. There were no data on the time points at which relapses started or ended.

Method

The disease activity model is a Markov chain with four states: 1) first month of remission, 2) subsequent months of remission, 3) first month of relapse, and 4) subsequent months of relapse. A period of remission is defined as an unbroken sequence of cycles spent in states 1 and/or 2. A relapse is defined as an unbroken sequence of cycles spent in state 3 and/or 4. Surgery can occur in states 3 and 4.

An exact maximum likelihood estimator was used, that translated the yearly observations into monthly probabilities of transition between remission and relapse, and surgery. The probability of remission depends on time since start of relapse, as does the probability of relapse since the start of remission, due to the model structure. The parameters themselves do not change over time in our context.

The initial implementation of the estimator was slow, counting through all possible pathways of the model. Many paths have a zero likelihood, and not all are unique in how their likelihood depend on the parameters. We created a list of *profiles*, with the values necessary to evaluate the likelihood of each unique pathway, given the parameter values. We thus optimized our estimator.

Results

The maximum likelihood estimator appears to work well. Simulated training datasets result in reasonable estimates. The estimator initially took over three hours to complete. Optimization reduced this time to around one minute.

The estimated disease activity model fits well to observed data and has good face validity, in the absence of curative surgery. Presence of curative surgery imposes a transient nature to the disease which makes the disease activity model unsuitable.

Conclusions

The disease activity model and its estimator work well. Presence of curative surgery calls for further development of the model, the estimator and its use of profiles.

MCPMod – An R Package for the Design and Analysis of Dose-Finding Studies

Björn Bornkamp José Pinheiro Frank Bretz

In this presentation the **MCPMod** package for the R programming environment will be introduced. It implements a recently developed methodology for dose-response analysis that combines aspects of multiple comparison procedures and modeling approaches (Bretz et al., 2005, Pinheiro et al., 2006). The MCPMod package provides tools for the analysis of dose finding trials as well as a variety of tools necessary to plan an experiment to be analysed using the MCP-Mod methodology. Both design and analysis capabilities of the package will be illustrated with examples.

References

Bretz F, Pinheiro J, Branson M (2005). Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies. *Biometrics, 61*, 738–748.

Pinheiro J, Bornkamp B, Bretz F (2006). Design and Analysis of Dose Finding Studies Combining Multiple Comparisons and Modeling Procedures. *Journal of Biopharmaceutical Statistics, 16*, 639–656.

Use R! for estimating forest parameters based on Airborne Laser Scanner Data

Johannes Breidenbach
Forstliche Versuchs- und Forschungsanstalt Baden-Württemberg
Abteilung Biometrie und Informatik
Johannes.Breidenbach@forst.bwl.de

1 Abstract

Forest parameters such as timber volume, diameter distributions, tree height and tree species are important information for a sustainable forest management and planning issues in the wood-working industry. Additionally, the amount of carbon stocks in woody biomass has become an crucial parameter due to international reporting commitments (e.g., the Kyoto-protocol). Conventionally, these information are surveyed in sample plot inventories.

However, terrestrial sample plot inventories usually cannot provide estimates on the stand-scale[1]. Furthermore, as a result to their high costs, they are repeated in a decennial cycle. Therefore, one aim of the research project MatchWood (`www.matchwood.de`) is to develop methods to regionalize forest parameters based on remotely sensed data. Since many variables of interest are correlated with the structural characteristics of the canopy, airborne laser scanning data were used as auxiliary variable.

Airborne laser scanning (ALS) or light detection and ranging (lidar) is an active remote sensing technique that comprises scanning and navigation units. In an ALS system, a laser pulse is projected on a scanning mirror and sent to the surface. Since the position and orientation of the aircraft is known, the time-of-flight of the laser pulse can be used to determine the position of the reflection on the earth's surface. ALS provides a high resolution 3D representation of the canopy and the terrain surface in one overflight.

R was used to derive height- and density metrics of the lidar-derived vegetation height for inventory plots and to develop statistical models for the response

[1] A usual forest stand in southern Germany has an area of 1-3 ha and comprises of trees with more or less the same species and age.

variables. The presentation will show the application of different R methods and libraries for estimation and regionalizing the above mentioned forest parameters. For example:

- Calling external command-line tools (FUSION) for handling the huge amount (about 500.000 returns km^{-2}) of lidar raw data.
- Mixed-effects models (library nlme) for estimating timber volume and biomass by accounting for the spatial correlation of the inventory plots and heteroscedasticity.
- Generalized additive models for location, scale and shape (library GAMLSS) for estimation of the Weibull distributed response variable diameter.
- RandomForests for non-parametric estimation of timber volume by species.
- Generating maps using the maptools library.

Tricks and Traps for Young Players UseR! 2008 Dortmund, August 12-14 2008

Ray D Brownrigg
Statistical Computing Manager
School of Mathematical and Computing Sciences
Victoria University of Wellington
Wellington, New Zealand
ray@mcs.vuw.ac.nz

March 18, 2008

Abstract

This presentation will illustrate for new users to R some of its very useful features that are frequently overlooked, and some frequently misunderstood features. Emphasis will be on achieving results efficiently, so there may be some value for (moderately) seasoned users as well as beginners. Many of the features discussed will be illustrated by following the development of an actual simulation project.

Issues to be discussed include:

- Using a matrix to index an array
- Vectorisation
  - user-defined functions (using curve(), optimi[zs]e())
  - pseudo vectorisation
  - multi-dimensional
- Matrices, lists and dataframes, which are most efficient?
- Local versions of standard functions
- Resolution of pdf graphs
- .Rhistory
- get()
- file.choose()
- sort(), order() and rank()

Exploring Financial System Convergence in 8 OECD countries by means of the plm package.

Giuseppe Bruno

Bank of Italy, Economic Research and International Relations

Keywords: Financial systems, β convergence, σ convergence.

Abstract

The relevance of financial systems for collecting resources from saving households to funds constrained firms is widely recognized.

Taking advantage of a dataset covering the financial accounts for 8 of the main OECD economies we run some experiments of $\beta-$ and $\sigma-$ convergence for the main components of the household financial assets. These experiments have been carried out with the **R** package equipped with the plm package for panel data estimation. The empirical literature on $\beta-$ and $\sigma-$ convergence is typically based on regression models where the average growth rate of per capita income is assumed dependent on its initial level and possibly on other exogenous variables used to control for country idiosyncracies:

$$\frac{1}{T}log(y_{i,t+T}/y_{i,t}) = \alpha + \beta y_{i,t} + \gamma x_{i,t} + \epsilon_{i,t} \tag{1}$$

In this model we say there is *conditional $\beta-$convergence* if we find $\beta < 0$. In other words, in presence of $\beta-$convergence poor economies tend to grow faster, and therefore to catch up richer countries. Sala-i-Martin (1996) proposed the concept of $\sigma-$convergence defined as follows: a group of economies satisfy $\sigma-$convergence if the dispersion of their per capita income levels decreases over time: $\sigma_{t+T} < \sigma_t$ where $\sigma_t = \Sigma_{i=1}^{N}(log(y_{i,t} - \bar{y}_t)^2$. Using a dataset on financial accounts produced in 2007 by the OECD, Pioneer G.A.M. and some National Central Banks, we have carried out a thorough convergence analysis in the **R** environment. In this paper we explored the behaviour of the total financial assets held by household and four of their main components: **currency and deposits**, **securities other than shares**, **shares and other equities**, **insurance technical reserves**.

The main economic conclusions drawn from the analysis are:

a) it is found evidence of $\beta-$ and $\sigma-$ convergence for the household total financial assets, shares and other equity, and insurance product;

b) often no convergence is found for currency and deposits and securities other than shares;

c) the intensity of banking disintermediation for deposits shows marked difference among the OECD countries.

These kind of empirical applications are usually carried out with commercial econometric packages. In this work we compared the numerical results produced by **R** with those achieved with two well know packages such as E-Views and LIMDEP. Some interesting results can be drawn from this comparison:

a) the coefficients estimates do always agree among the different packages;

b) some differences arise among the numerical values of the standard errors;

c) model definition might be improved with the help of an inline symbolic lag/lead operation.

Computationally Tractable Methods for High-Dimensional Data

Peter Bühlmann

Many applications nowadays involve high-dimensional data with p variables (or covariates), sample size n and the relation that $p \gg n$. We focus on penalty-based estimation methods which are computationally feasible and have provable statistical and numerical properties. The Lasso (Tibshirani, 1996), an ℓ_1-penalty method, became very popular in recent years for estimation in high-dimensional generalized linear models. Extensions to other models or data-types call for more flexible convex penalty functions, for example to handle categorical data or for improved control of smoothness in additive models. The Group-Lasso (Yuan and Lin, 2006) and a new sparsity-smoothness penalty are general and useful penalty functions for many high-dimensional models beyond GLM's. Fast coordinatewise descent algorithms can be used for solving the corresponding convex optimization problems which allow to easily deal with large dimensionality p (e.g. $p \approx 10^6, n \approx 10^3$).

The talk includes: (i) a review of Lasso-type methods; (ii) new flexible penalty functions, fast algorithms (`R` package `grplasso`) and some comparisons with boosting; and (iii) some illustrations for bio-molecular data.

An Automatic Recommendation System using R: Project Thank You eMail

Christopher Byrd*

March 31, 2008

Throughout the guest experience, at an IHG brand hotel, chances are you will receive an email communication that contains an offer. For example, “Earn 5,000 miles on your next stay”. Using only Open Source software, resident statisticians begin the task of building an automatic recommendation system, with R as a core component. The pilot project for this endeavor is named, “Thank You eMail”. The name is fitting since the plan is to deliver offer recommendations through the Thank You email; which are distributed 24-72hrs after a guest has checked out. Given the high volume of email transactions the team must meet standard IT constraints e.g. $>$ 1msec response time. This presentation will show the role of R in the enterprise wide solution, and how Java it used to integrate it with other very popular and easily accessible tools.

Packages: R Sessions, FlexMix (latent class regression), Bayesm, and more

*InterContinental Hotels Group, USA

washAlign: a GC-MS Data Alignment Tool Using Iterative Block-Shifting of Peak Retention Times Based on Mass-Spectral Data

Minho Chae[1, 3], John J. Thaden[3, 5], Steven F. Jennings[2], and Robert J. Shmookler Reis[3, 4, 5]

[1]UALR/UAMS Joint Graduate Program in Bioinformatics, University of Arkansas at Little Rock, Little Rock, AR 72204
[2]Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR 72204
[3]Department of Geriatrics, and [4]Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, AR 72205
[5]Central Arkansas Veterans Healthcare System LRVA-151, 4300 W. 7th Street, Little Rock, AR 72205

Email addresses: MC: mxchae@ualr.edu, JT: jthaden@uams.edu, SJ: sfjennings@ualr.edu, and RJSR: rjsr@uams.edu

In GC-MS, a gas chromatograph (GC) resolves chemicals by time of elution from a coated capillary through which gas flows; a mass spectrometer (MS) resolves ions (produced upon fragmentation of eluates) by mass/charge (m/z) ratio; and an acquisition program records ion intensity as a function of m/z and elution, yielding spectra and chromatograms, respectively. A problem when comparing records in an experiment is that elution times will vary. washAlign has been developed in R to address this problem. It warps regions between peaks that it has shifted, thereby aligning those peaks to spectrally matched peaks in a reference chromatogram while preserving their shape and area. Through pair-wise comparisons of all records to one arbitrarily selected reference record, all records in a large experiment can be aligned for subsequent processing, *e.g.*, by three-way methods, including those such as PARAFAC that assume mathematical trilinearity.

In washAlign, (a) ion chromatograms are extracted for a subset of those m/z channels with the five highest ion intensities in any of the consecutive MS scans that define "a region of the sample and reference chromatograms that exhibit a peak on the total intensity chromatogram"; (b) peaks are detected in them, and key peaks are matched between sample and reference through a procedure involving iterative localization with spectral correlation, to produce for each sample and the reference a peak list for alignment; and (c) the key sample peaks are shifted toward the matching peaks in the reference run, and nonpeak regions are warped, i.e., linearly interpolated, to join the shifted peak regions.

Users can visually inspect the chromatograms before and after alignment of a pair of chromatograms, through an interactive selection of matched peaks. Taking an iterative block-shift approach makes it possible to not only reveal strongly matching peaks at early stages but also to reduce the risk of mismatching chemically different peaks.

Scaling and Robustifciation of ARMA Models with GARCH/APARCH Errors Using R/Rmetrics

Yohan Chalabi, ETH Zürich and Finance Online GmbH Zürich
Michal Miklovic, Charles University Prague
Diethelm Würtz, ETH Zürich

Abstract

This presentation explores concepts and methods to implement extensions of ARMA models with GARCH/APARCH errors introduced by Ding, Granger and Engle.

It is nowadays common to estimate GARCH/APARCH models of financial time series. They play an essential role in risk management and volatility forecasting. Although these models are well studied, numerical problems may arise in the estimation of series with extreme events.

In this talk, we present how to explore the different behavior in the upper and lower tails of the financial return series distribution. Generalized hyperbolic skew Student's t-distributions can explain extreme polynomial losses and exponential decaying gains. We follow ideas from robust estimation, appropriate parameter scaling from optimization, and present their implementation in R/Rmetrics.

References:

Ding, Granger and Engle, A long memory property of stock market returns and a new model, Journal of Empirical Finance 1, 1993, 83

R/Rmetrics Core Team, R/Rmetrics fGarch Package, www.rmetrics.org and r-forge.r-project.org

Submitted to "The R User Conference 2008"
August 12-14, Technische Universität Dortmund, Germany

tdm - A Tool of Therapeutic Drug Monitoring in *R*

Miao-ting Chen[a], Yung-jin Lee[b]

[a]Department of Hospital Pharmacy, Kaohsiung Veteran General Hospital,
[b]College of Pharmacy, Kaohsiung Medical University, Kaohsiung, Taiwan

Introduction Therapeutic drug monitoring (TDM) aims to optimize individual patient's drug therapy through monitoring the plasma/serum concentrations of the target drug, as well as the observed clinical responses. However, there are usually only few blood samples that can be collected and analyzed. Usually there is even only one single blood sample available. Therefore, it becomes very important to accurately estimate individual pharmacokinetic (PK) parameters with limited observations. Bayesian estimation is a very suitable algorithm for this situation. In contrast to minimizing an objective function, Bayesian estimation with Markov-chain Monte-Carlo (MCMC) simulation (integration) using Gibbs sampler technique (BUGS) might be worth to implement and apply. Hence, the objective of this study was to develop a TDM tool using *BUGS* for R. **Methods and Materials** We chose *OpenBUGS*, an open-source version of BUGS for Windows (through its *R* interface package *BRugs*), to develop this tool under *R*. Each drug model was divided into two parts: the probability distribution of population PK parameter (as priors), and the probability distribution of observed drug serum/plasma concentration or observed clinical response (as the conditional probability or the likelihood function). This tool was validated with simulated data obtained from the published PK parameters within the range of 2*s.d.. And the accuracy of PK parameters was evaluated with percent prediction error (PE %). **Results and Discussion** We named this tool as *tdm*. Seventeen drug models including one PK/PD model (warfarin) and sixteen PK models were built in *tdm*. It can be used to estimate individual PK/PD parameters with one or more observations obtained from a single subject, as well as multiple subjects at the same time. Other than one drug, imatinib, PK or PD parameters of all other drugs are estimated at their steady-state. Furthermore, *tdm* also provides dosage adjustment function. Based on the results of estimation validation, we found PEs of PK parameters of built drugs were similar to those using nonlinear regression obtained from other computer software, *JPKD*. **Conclusion** *tdm* has been released on Nov. 2006 and can be downloaded and installed from *R* mirror websites. The latest version is 2.2.1. Currently *tdm* is only available for Windows, because *BRugs* has not been available for other platforms yet.

Title: The Virtual R Workbench, towards an open platform for R based e-Science

Author: Karim Chine, Internet Center – Imperial College London

Abstract. Biocep frameworks and tools make it possible to use R as a Java object-oriented toolkit or as an RMI server. Calls to R functions from java locally or remotely cope with local and distributed R objects. Stateless and stateful JAX-WS web services can be generated and deployed on demand for R packages. An infrastructure with a large number of R servers running on an heterogeneous set of machines can be deployed and used for multithreaded web applications and web services, for distributed and parallel computing, for thin web clients dynamic content generation including graphics and for R virtualization in a shared computation resources context. The virtualization is based on a universal advanced GUI for R (virtual R workbench) that can be used also to control self-managed R servers. A dedicated HTTP gateway enables the control of R servers running behind firewalls. The workbench includes a powerful and easy-to-use docking framework, advanced script editors, spreadsheet views fully connected to R, R objects inspectors views, data storage views, a highly interactive zooming system for exploring complex visual data and several new R Graphics interactors. It can run as an applet, via Java Web Start or as a cross-platform desktop application.

The virtual workbench is capable of creating R servers on any remote machine having R accessible from the command line without any extra pre-installation/pre-configuration. It enables collaborative R Sessions (one session, multiple simultaneous users, console and devices content broadcasting). It has built-in distributed computing facilities accessible via the API or directly from the R Console. The functions available are similar to what has been defined within the snow package (makeCluster, clusterEvalQ, clusterExport, clusterApply, stopCluster..) and do not require any configuration.
Biocep has built-in Python scripting facilities both on server and on client sides. The bridging of R and Python is bidirectional, R objects can be exported to Python and Python objects imported to R. Scripting with R as a component becomes easier than ever by using the Biocep API or from within the workbench via the R / Python Consoles and via the embedded jEdit based script editor.

The virtual workbench is designed to be an open platform: on one hand, it allows users to acquire an R computational resource in different ways either by creating an R server on intranet machines or by connecting to public grids exposing a virtualized infrastructure via a HTTP or a SOAP front-end. On the other hand, it has a plugin architecture that enables the integration of new GUIs designed for end users as new views and perspectives. The creation of those views can be done programmatically (Java/Swing) or visually via a bean builder (Netbeans Matisse) and various Java beans are available as GUI components mapping standard R objects and devices. The plugin architecture handles the notification and the synchronization of the views with the R objects, changes done to the data in the views become effective within the R session and changes made on R objects are visible on real time in the different views. Several available interactive statistical software for data analysis (KLIMT, iPlots, Mondrian..) would become in the future plugins among others available on a web-accessible central repository.

The virtual workbench would enhance the user experience and the productivity of anyone working with R directly or indirectly. The openness would leverage the range of software available for statistical computing and statistical data visualization/exploration. The interoperability coupled with a large-scale deployment of virtualization infrastructures on various grids would democratise R based HPC and enable users from within their browsers to compute with R and visualize data with unprecedented flexibility and performance.

Biocep is a project hosted by R-Forge and it is released under the Apache 2.0 License
Biocep home: www.biocep.co.uk

Statistical cartoons

Ewan Crawford Adrian Bowman

The advent of interactive controls in R has allowed researchers to construct very convenient mechanisms to explore data and models but has also allowed lecturers to produce animated graphics to explain ideas. This talk focusses on the latter activity and aims to draw associations with both meanings of the word ‘cartoon’. From one perspective these are sketches of the real thing which aid experimentation and understanding while, from another, animated drawings have the potential to raise a smile in the classroom. Both of these are helpful.

In between the command line interface of R and the gui interfaces such as R Commander (Fox, 2005), panel controls offer a very useful mode of control in both classroom and laboratory settings. This talk aims to illustrate and discuss ‘cartoons’ of this type, using a variety of illustrations from the context of teaching and learning at elementary, advanced and practical levels. The illustrations have been built using the rpanel package which was designed to provide a quick and easy means of building control panels. The basic design of the package, which is based on the rtcltk (Dalgaard, 2001) extensive set of gui tools, will also be outlined.

An Alternative Package for Estimating Multivariate Generalised Linear Mixed Models in R

Rob Crouchley[1], Damon Berridge[2], Dan Grose[1]

[1]Centre for e-Science, Lancaster University, Lancaster, LA1 4YT
[2]Centre for Applied Statistics, Lancaster University, Lancaster, LA1 4YF

Contact email: r.crouchley@lancaster.ac.uk

Abstract

There are several packages at [1] that have been specially written for estimating Generalised Linear Mixed Models in R, these include, lme4 [2] and npmlreg [3]. There are also commercial systems that have algorithms for the same class of models, see e.g. Stata [4], gllamm [5] and SAS [6]. In this presentation we compare the performance of these systems with our alternative (sabreR, to be available from [7]) on some standard small to medium sized data sets and show that our alternative is very much faster. We also present a grid enabled version of the software (SabreRgrid), this shows how easy it has become to submit grid jobs from the desktop PC and the extra speed up that can be obtained by going parallel on a High Performance Computer on the grid or otherwise. This extra speed up is particularly important for estimating complex models on large and very large data sets.

SabreR is a program for the statistical analysis of multi-process event/response sequences. These responses can take the form of binary, ordinal, count and linear recurrent events. The response sequences can also be of different types (e.g. linear (wages) and binary (trade union membership)). Such multi-process data is common in many research areas, e.g. in the analysis of work and life histories from the British Household Panel Survey or the German Socio-Economic Panel Study where researchers often want to disentangle state dependence (the effect of previous responses or related outcomes) from any omitted effects that might be present in recurrent behaviour (e.g. unemployment). Understanding of the need to disentangle these generic substantive issues dates back to the study of accident proneness (Bates and Neyman, 1952) and has been discussed in many applied areas, including consumer behaviour (Massy et al, 1980) and voting behaviour (Davies and Crouchley, 1985)

SabreR can also be used to model collections of single sequences such as may occur in medical trials, e.g. headaches and epileptic seizures (Crouchley and Davies, 1999, 2001), or in single equation descriptions of cross sectional clustered data such as the educational attainment of children in schools.

We call the class of models that can be estimated by sabreR, Multivariate Generalised Linear Mixed Models. These models have special features added to the basic models to help them disentangle state dependence from the incidental parameters (omitted or unobserved effect). The incidental parameters can be treated as random or fixed, the random effects models being estimated using normal Gaussian quadrature or Adaptive Gaussian quadrature. ‘End effects' can also be added to the models to accommodate ‘stayers’ or ‘non susceptibles’. The fixed effects algorithm we have developed uses code for large sparse matrices from the Harwell Subroutine Library, see [8].

SabreR and SabreRgrid also includes the option to undertake all of the calculations using increased accuracy. This is important because numerical underflow and overflow often occur in the estimation process for models with incidental parameters. This feature does not seem to be available is other similar software [2, 3, 4, 5, 6].

References

Bates, G.E., and Neyman, J., (1952), Contributions to the theory of accident proneness, I, An optimistic model of the correlation between light and severe accidents, II, True or false contagion, *Univ Calif, Pub Stat*, 26, 705-720.

Crouchley, R. and Davies, R.B., (1999), A comparison of population average and random effect models for the analysis of longitudinal count data with base-line information, *Journal of the Royal Statistical Society, Series A,* 162, 331-347

Crouchley, R. and Davies, R.B., (2001), A comparison of GEE and random effects models for distinguishing heterogeneity, nonstationarity and state dependence in a collection of short binary event series, *Statistical Modelling*, 1, 271-285

Davies, R.B. and Crouchley, R., (1985), The determinants of party loyalty: a disaggregate analysis of panel data from the 1974 and 1979 General Elections in England, *Political Geography Quarterly*, 4, 307-320.

Massy, W.F., Montgomery, D.B., and Morrison, D.G., (1970), *Stochastic models of buying behaviour*, MIT Press, Cambridge, Mass.

URL Links

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

[2] http://cran.r-project.org/web/packages/lme4/index.html

[3] http://cran.r-project.org/web/packages/npmlreg/index.html

[4] http://www.stata.com/

[5] http://www.gllamm.org/

[6] http://www.sas.com/

[7] http://sabre.lancs.ac.uk/

[8] http://www.cse.scitech.ac.uk/nag/hsl/

igraph – a package for network analysis

Gábor Csárdi

Department of Medical Genetics, University of Lausanne, Switzerland.

The igraph R package is an interface to the C library with the same name, developed for implementing graph algorithms. As many graph algorithms are already included in igraph, it is also a handy tool for (exploratory) network analysis.

Main igraph features:

- igraph uses a simple, flat data structure for graph representation, this allows handling graphs with millions of edges and/or vertices.
- It is possible to assign attributes to the vertices or edges of the graph, or to the graph itself, the attributes can be arbitrary R objects.
- Graph visualization, both interactive and non-interactive, using 1) traditional R graphics, 2) Tcl/Tk or 3) OpenGL via rgl.
- A variety of classic and recent graph algorithms are implemented in igraph: ○ Shortest paths and shortest path based measures, e.g. diameter. ○ Weakly and strongly connected components, biconnected components and articulation points. ○ Maximum flows and minimum cuts, edge and vertex connectivity. ○ Various centrality measures: degree, closeness, betweenness, Burt's constraints, Page Rank, eigenvector centrality, Kleinberg's hub and authority scores. ○ Fast graph and subgraph isomorphism algorithms. ○ Cliques and independent vertex sets. ○ Graph motifs. ○ Community structure detection based on many recently published heuristics. ○ K-cores, transitivity, minimum spanning trees, toplogical sorting, etc.
- Graphs can be created in various ways: ○ From data frames, edge lists, adjacency matrices, from a simple R formula notation. ○ From a list of famous graphs, predefined structures like rings, stars, trees, etc. or from the Graph Atlas. ○ Using random graph models, like preferential attachment, or the small-world model.
- igraph supports many commonly used file formats for storing graphs, like GraphML, GML or the format used by Pajek.

In this lecture I will show several practical examples on how to turn data into igraph graphs, how to calculate various graph properties: vertex centrality and community structure, and graph visualization.

Quantitative approach to Entropy weighting methodology in MADM

Mohammad Ali Dashti

It seems that those MADM matrices whose scattering of the alternative values distribution have different importance. The Entropy technique would give completely irrelevant weights in comparison to other techniques such as external weighting. We propose a method to resolve this apparent disagreement. In this method we shall first convert all quantitative values to qualitative values using DM judgment and then we will apply the Entropy technique as before. An example is presented to illustrate the proposal.

RiDMC: an R package for the numerical analysis of dynamical systems

Antonio, Fabio Di Narzo[*] and Marji Lines[†]

February 5, 2008

Abstract

RiDMC is an R package for the numerical analysis of discrete- and continuous-time dynamical systems. With RiDMC the user can easily encode a model in the simple, interpreted LUA language, and immediately perform numerical analysis with a variety of algorithms. The LUA language gives maximum flexibility in model specification, and allows for the introduction of stochastic components in a very natural way. Or if the user wants to work with an existing model, he may choose from a large number of well-known dynamical systems already available in the package.

Once a model is loaded, the user can compute trajectories, bifurcation diagrams, Lyapunov spectra, basins of attraction and periodic cycles. For each analytical routine there is an associated plotting function, with reasonable default settings (axes labelling, font sizes, etc.), so that publication-quality plots can be produced directly with almost no additional effort. Moreover, plots are based on the grid system, so that full plot customization, manipulation and reuse is possible for more expert R users.

RiDMC uses the idmclib C library for interpreting user-supplied models and for doing core numerical computing. The idmclib library, released with sources under the GPL-v2 license, is small, well-documented and easy to understand for anyone desiring a closer look at the internal numerical algorithms.

A set of interesting case studies is presented as a demonstration of the package capabilities.

[*] Dipartimento di Scienze Statistiche, Università degli studi di Bologna, via Belle Arti 41, 40126 Bologna, Italia, e-mail: antonio.dinarzo@unibo.it

[†] Dipartimento di Scienze Statistiche, Università degli studi di Udine, via Treppo 18, 33100 Udine, Italia, e-mail: marjilines@gmail.com

Created by USNACracking the Nut

Making the best computing and most sophisticated methods accessible to the Social Science Undergrad

This is a paper presents strategy for introducing R to a social science department not necessarily ready to embrace it. It about not simply teaching R, but finding a mechanism to insert it into the core of a department. It suggests the naïve assumption of R's 'self-sellability' invites enormous frustration and almost certain failure. It argues that springing R on a department should be a campaign based on principles of military planning. It draws its occasionally offbeat lessons from the generally successful effort to integrate R in the political science department at the United States Naval Academy.

The key point is this: a successful R introduction does not resemble the entry of conquering heroes to the exuberant welcome of a liberated population. It is far more like an insurgency, fought fiercely behind the scenes.

For accomplished R users this is mystifying. We enthusiastically celebrate R's computational power, its magnificent graphics, and its explosive increase in functionality through tailored add-on libraries. Sometimes, though, we forget how intimidating it all looks to a novice uncertain about even loading the data.

Social science generally has been revolutionized by advances in sophisticated methods, and the disciplines have been forever changed. However, these developments have not altered the type of undergraduate selecting social science as a discipline. Nor have they necessarily spread to established faculty whose training predates this revolution, or whose interests are not easily addressed by a data set.

As a result, scholars selling R are greeted not by adoring throngs seeking leadership into the world of cutting edge social science, but rather by intensely skeptical stakeholders with strong interest in preserving the status quo. Frontal assaults on this position are futile; the push-back is overwhelming.

As a model for a successful introduction, this paper suggests following guidelines for military planning. Generations of the best military minds have devised strategies for asymmetrical match ups. Students in war colleges and service academies are challenged to examine lessons of the past when facing contemporary threat environments. They are encouraged to consider their strengths, and those of the enemy, and to bring maximum force to bear on a problem with supreme economy of effort.

This paper suggests that the problem of introducing R are partly tactical (using graphs and simulated quantities of interest to make complex findings accessible), and partly strategic (building alliances, scoring and exploiting visible public victories, and when all else fails, deceiving) until victory belongs not to the entrenched, but to the deserving. This paper tells how it was done here, and offers it as a model for others facing similar struggles.

Dubyak, William G, US Naval Academy

dynGraph: interactive visualization of “factorial planes” integrating numerical indicators

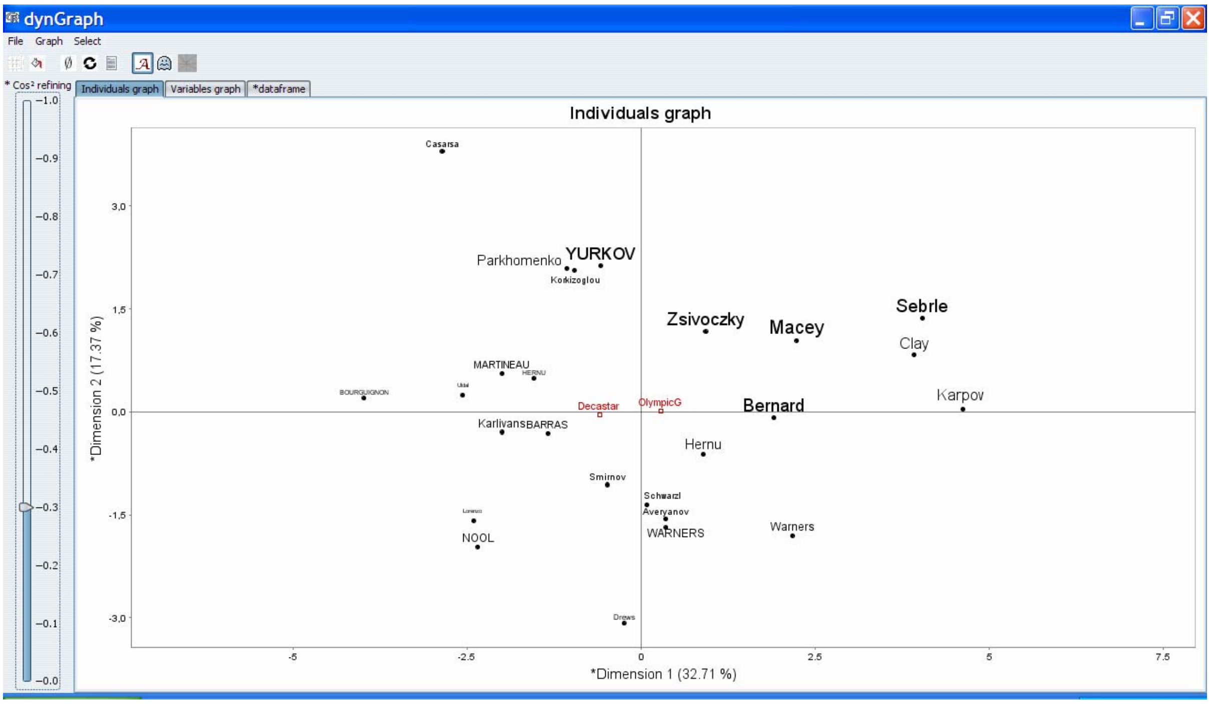
J. Durand, J. Josse, F. Husson and S. Lê*
Agrocampus Rennes, 65 rue de St Brieuc, CS 84215, 35042 Rennes Cedex, France
sebastien.le@agrocampus-rennes.fr; fax 02 23 48 58 93; phone 02 23 48 58 81

dynGraph is a visualization software that has been initially developed for the FactoMineR package, an R package dedicated to multivariate exploratory methods such as principal components analysis, (multiple) correspondence analysis and multiple factor analysis (http://factominer.free.fr); dynGraph has been extended to allow the visualisation of data frames. The main objective of dynGraph is to allow the user to explore interactively graphical outputs provided by multidimensional methods by visually integrating numerical indicators.

The first basic feature of dynGraph is the connecting line that appears whenever the user moves a label associated with an object, *i.e.* an individual, a variable or a category. Labelling of the different objects displayed on the graph can be easily set. Colours can be assigned to individuals according to a categorical variable of interest.

One of the main features of dynGraph is the way objects are displayed. Objects are displayed according to their quality of representation, by default above the threshold of 0.8 with respect to a maximum of 1. Of course the amount of information to be displayed can be easily set by the user with a cursor: graphical outputs can be analyzed interactively from the most general piece of information to the most relevant one. Moreover, the font size of each label associated with an object is proportional to the importance of the object in the analysis which facilitates tremendously the interpretation of the results. Besides, different criteria can be used to assess the importance of an object and this information is calculated via R and the FactoMineR package.

Finally, by clicking on one of the dimensions provided by the analysis, the user gets a list of the variables that may explain the dimension significantly that will help him to interpret the data.



A Graphical User Interface for Environmental Statistics

R. Dutter[1]

[1] Vienna University of Technology, Wiedner Hauptstr. 7, A-1040 Vienna, Austria

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

Abstract

We report on a package called DAS+R under development using a graphical user interface which should ease the application of more or less sophisticated methods. The basis of the graphical user interface comes from the R Commander (see Fox, 2004). It uses Tcl/Tk programming tools (Welch and Jones, 2003). The emphasis is on the analysis of spatially depending uni- or multivariate data, particularly on problems of geochemical data.

Three special properties of DAS+R should be stressed:

- Interactive definition of data subsets (numerically or graphically) together with set operations. Usage of these subsets in almost all graphics and computations.
- Intensive use of possible relations between the geographical information with the values of data in the statistical and graphical analysis.
- The strong requirement of fast reproducibility and repeatability with small variations in the analysis.

For specified subsets many simple graphics can be generated in an easy way by a few mouse clicks (histograms, boxplots, xy-, ternary plots, scatterplot matrices). These nevertheless can become very sophisticated by using the provided advanced options where almost all options of the usual R commands can be specified by clicking graphical icons.

The geographical information is used by generating different kinds of maps. Different symbol sets can be used for representing the values in space. Surface maps may be produced by simple interpolation algorithms or by sophisticated geostatistical methods as kriging.

All these graphical displays may be produced in any specified scale on a user defined worksheet which can be interactively splitted into arbitrary frames which are provided for the different graphics.

Finally many multivariate methods like principal component and factor analysis, cluster and discriminance analysis, are available.

The package is also meant as a companion to the book recently published by Reimann et al. (2008). We describe in short the system and illustrate the usability on some geochemical data sets.

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, P. Filzmoser, R.G. Garrett, and R. Dutter (2008). *Statistical Data Analysis Explained: Applied Environmental Statistics with R. Wiley, New York.*

B.B. Welch and K. Jones (2003). *Practical Programming in Tcl and Tk. Prentice Hall PTR, New York.*

Scripting with R in high-performance computing: An Example using littler

Dirk Eddelbuettel

High-Performance Computing with R often involves distributed computing. Here, the MPI toolkit is a popular choice, as it is well supported in R by the Rmpi and snow packages. In addition, resource and and queue managers like slurm help in allocating and managing computational jobs across compute nodes and clusters.

In order to actually to execute tasks, we can take advantage of a scripting frontend to R such as r (from the littler package) or Rscript. By discussing a stylized yet complete example, we will provide details about how to organise a task for R by showing how to take advantage of automated execution across a number of compute nodes while being able to monitor and control its resource allocation.

Management and Analysis of Large Survey Data Sets Using the memisc Package

Martin Elff
Universität Mannheim

March 25, 2008

Abstract

One of the aims of the memisc package is to make life easier for useRs who have to work with (large) survey data sets. It provides an infrastructure for the management of survey data including value labels, definable missing values, recoding of variables, production of code books, and import of (subsets of) SPSS and Stata files. Further, it provides functionality to produce tables and data frames of arbitrary descriptive statistics and (almost) publication-ready tables of regression model estimates. Also some convenience tools for programming and simulation are provided, as well as some miscellaneous probability distributions, statistical models, and graphics.

Based on an example analysis of the cumulated ALLBUS 1980-2004 data set (ZA-No. 4243), it is demonstrated how even large data sets can be handled without much pain using the memisc package. The cumulated ALLBUS comprises data of 44,526 respondents and 1,141 (!) variables. The proposed presentation shows the workflow of analysis of such a large data set: First, variables that are relevant for the analysis are loaded selectively into the workspace, thus minimizing the overall memory footprint. Second, attributes of variables in such a data set, like variable labels, value labels and user-defined missing values are retained and used for data management conducive for typical social science data analysis. Third, tables of descriptive statistics are produced for preliminary or exploratory analyses using the genTable function of the package. Fourth, estimates of statistical models are formatted in a way suitable for publication in social science journals using the mtable function.

The bigmemoRy package: handling large data sets in R using RAM and shared memory

John Emerson
⟨john.emerson@yale.edu⟩

Michael Kane

Multi-gigabyte data sets challenge and frustrate R users even on well-equipped hardware. C programming provides memory efficiency and speed improvements, but is cumbersome for interactive data analysis and lacks R's flexibility and power. The new package **bigmemoRy** bridges this gap, implementing massive matrices in memory (managed in R but implemented in C) and supporting their basic manipulation and exploration. It is ideal for problems involving the analysis in R of manageable subsets of the data, or when an analysis is conducted mostly in C.

In a Unix environment, the data structure may be allocated to shared memory, allowing separate R processes on the same computer to share access to a single copy of the data set; mutual exclusions (mutexes) are provided to avoid conflicts. This opens the door for more powerful parallel analyses and data mining of massive data sets.

Exploratory and Inferential Analysis of Benchmark Experiments

Manuel J. A. Eugster and Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München,
Ludwigstrasse 33, 80539 München, Germany,
firstname.lastname@stat.uni-muenchen.de

Abstract. Benchmark experiments produce data in a very specific format. The observations are drawn from the performance distributions of the candidate algorithms on resampled data sets. `benchmark` is the comprehensive `R` toolbox for the setup, execution and exploratory and inferential analysis of these experiments. The package introduces an additional layer of abstraction (using S4 mechanisms) representing the elements of benchmark experiments. This allows the integration of all statistical learning algorithms available in the `R` system and a consistent way for developing new ones. The consequence of this slight extra work is a standardized setup and analysis of benchmark experiments. The package provides wrapper methods for common learning algorithms available in `R`.

In this presentation we introduce the elements of benchmark experiments and show how to combine them into a flexible framework. The usage is illustrated with exemplary benchmark studies based on common learning algorithms on one or several popular data sets, respectively. We present new visualisation techniques, show how formal test procedures can be used to evaluate the results, and, finally, how to sum up to an overall ranking.

Hedging interest rate risk with the dynamic Nelson/Siegel model

Robert Ferstl Josef Hayden
Department of Finance, University of Regensburg

An accurate forecast of the yield curve is an important input for the pricing and hedging interest-rate-sensitive securities. Diebold and Li (2006) formulate the widely-used Nelson and Siegel (1987) model in a dynamic context and provide a factor interpretation of the estimated parameters as level, slope and curvature. This model can be used to forecast the future yield curve.

We implement the dynamic Nelson/Siegel model in R by extending the CRAN package `termstrc`, which allows us to efficiently use market data from coupon bonds (see Ferstl and Hayden, 2008). Further, we test the performance of bond portfolio and interest rate risk management problems, where the dynamic Nelson/Siegel yield curve is used for pricing and hedging the underlying securities. We compare our results to common strategies in practice, e.g. duration hedging, duration vector models.

References

Diebold, F. X. and C. Li (2006, February). Forecasting the Term Structure of Government Bond Yields. *Journal of Econometrics 130*(2), 337–364.

Ferstl, R. and J. Hayden (2008). Zero-Coupon Yield Curve Estimation with the Package termstrc. Working Paper.

Nelson, C. and A. Siegel (1987, October). Parsimonious Modeling of Yield Curves. *The Journal of Business 60*(4), 473–489.

Sweave or how to make 286 customized reports in two clicks

Delphine Fontaine

The R environment is normally used to perform statistical analyses and then people making statistics usually make a report. To make this report, we can either copy and paste the R output in a text editor or use Sweave.

Sweave is an R tool created by Friedrich Leisch which allows the insertion of the R code in LaTeX code in such a way that statistical analysis and statistical reports are compiled at the same time. The purpose of Sweave is to create dynamic reports which are automatically modified when data or analysis change (Leisch, 2002).

Sweave allows one to quickly update a report if data changed. It also can be used to make several reports with different data but all having the same structure (same sections, same text, same graphs, same tables...). With some automation and two clicks, it is possible to use Sweave to make a vast number of reports, each with a different data subset.

In clinical development, doctors participating in a study usually receive a report with the general results of the study. Sweave can be used to make this report. But one can go beyond this. Sweave allows the customisation of a report for each doctor using the data collected in his site and to compare the resulting customized statistical analyses with the overall study results.

References

Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, Compstat 2002 - Proceedings in Computational Statistics, pages 575-580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9.

The Past, Present, and Future of the R Project – Social Organization of the R Project

John Fox

Much of the work in the social sciences on the development of open-source software focuses on the issue of motivation: Why do individuals or organizations participate in open-source projects? Is their participation rational? Voluntary activity is, however, a natural part of social life, and I find it more interesting to ask how an open-source project, such as the R Project for Statistical Computing, is organized, and how its social organization contributes to the success – or lack of success – of the project. After all, anarchic voluntary cooperation does not, on the face of it, seem a promising approach to developing a complex product such as statistical software. My investigation, based partly on interviews with members of the R Core team and with other individuals closely associated with the R Project, suggests that the success of R is due to a number of factors. Some of these factors, such as the implementation of a division of labour (albeit an informal one), are common to most organizations; other factors, such as the clever social use of technology (e.g., version control and package systems), are specifically adapted to the development of software; and still others, such as the adoption of the S language, which was already in wide use prior to the introduction of R, are particular to R.

R4X : Simple XML Manipulation for R

Romain François – *Mango Solutions*

useR! 2008. Dortmund.

Abstract

Data transfer is an important component in many multi-technology applications. The eXtensible Markup Language (XML) is a medium of choice for exchanging various sources of data. Recent developpements at Mango Solutions have justified the production of an R package to provide convenient manipulation of XML structures.

Based on the powerful parsing facilities of the `XML` package[4] and templating abilities of the `brew`[3] package, `R4X` gives R users a simple mechanism to create, read and manipulate XML structures. The functionality of the package is conceptually based on the E4X[2] standard which promotes XML as a core data-type of the javascript language. In order to create a seamless integration of XML into R, much of the functionality of E4X has been ported to R4X.

`R4X` provides a convenient environment for the *creation* of XML structures, through the single generic `xml` function. `R4X` also features simple *manipulation* of XML structures via the usual R slicing operators (`[ and [[`) combined with a syntax close to XPATH in order to extract arbitrarily nested content from an XML structure.

This presentation will describe key features of the `R4X` package and discuss anticipated extensions of the functionality. Examples will be used to demonstrate the use of `R4X` to build a simple Rich Site Summary (RSS) reader, generate a tag cloud of the description of current CRAN packages in xHTML, create Scalable Vector Graphics (SVG) and a custom RUnit[1] protocol report based on the Mozilla XML User Interface Language (XUL).

References

[1] Matthias Burger, Klaus Juenemann, and Thomas Koenig. *RUnit: R Unit test framework*, 2007. R package version 0.4.17.

[2] The ECMAScript Group. *Standard ECMA-357. ECMAScript for XML (E4X) Specification*, 2nd edition edition, 2005.

[3] Jeffrey Horner. *brew: Templating Framework for Report Generation*, 2007. R package version 1.0-2.

[4] Duncan Temple Lang. *XML: Tools for parsing and generating XML within R and S-Plus.*, 2007. R package version 1.93-2.

XML-based Reporting Application

Romain François, David Ilsley – *Mango Solutions*

useR! 2008. Dortmund.

Abstract

Several software projects recently developed at Mango Solutions require the production of fully styled reports in several output formats, mainly HTML, PDF and RTF.

Many existing systems were considered by Mango Consultants but were discounted due to restrictive licences, inflexibility of input data format, overcomplex or simplistic feature sets. The *Mango Report Generator* is a software component written in Java that has been developed to respond to the demands of producing flexible reports from multiple data sources.

The system is based on XML descriptions of the content of the report — currently covering graphics, tables and styled text report items — and the XML description of the actual layout of the report. The report layout is associated with the report items, and styled using Cascading Style Sheets (CSS)[2] to produce fully styled reports suitable for browsing using XHTML, printing or further editing in mainstream word processors using XSL-FO [1] and Apache FOP.

The input and output streams of the Report Generator are XML-based which makes it straightforward to create report items and layouts via any third party application. A proof-of-concept R package has been created as part of the project to demonstrate the ease of integration of content from other systems.

This presentation will highlight the challenges that occured during the developement of the component and a demonstration of the typical workflow of the system by creating reports by amalgamating content from R as well as a commercial implementation of the S language.

References

[1] Dave Pawson. *XSL-FO*, 2002.

[2] Dave Shea and Molly E. Holzschlag. *the Zen of CSS design*, 2005.

SIMSURVEY— a tool for (geo-) statistical analyses with *R* on the web

Mario Gellrich *ETH Zurich, Institute of Terrestrial Ecosystems, gellrich@env.ethz.ch,*
Rudolf Gubler,*Terraplan Gubler, gubler@terraplan.ch,*
Andreas Schönborn, *armadillo media gmbH, schoenborn@armadillo-media.ch,*
Andreas Papritz *ETH Zurich, Institute of Terrestrial Ecosystems, papritz@env.ethz.ch*

Geostatistical methods are used in many branches of environmental research and applications for the statistical analyses of spatially referenced measurements and for the interpolation and mapping of data measured at a limited number of locations in a study domain. Courses in geostatistics are therefore part of the curriculum in environmental sciences and engineering at many universities. However, experience shows that geostatistics is a rather difficult subject to teach. Apart from the mostly limited prior knowledge in statistics, a lack of flexible, but at the same time easy-to-use software adds to the problems many students have with this topic. Commercially available statistics and GIS software either offers no or only limited geostatistical functionality, or it is expensive (and in addition often quite demanding to use). *R* includes several powerful packages for geostatistical analyses, but as a script-based programming language, *R* is difficult to use in introductory courses.

To mend this deficiency we developed *SIMSURVEY*, a graphical user interface (GUI) for geostatistical analyses with *R*. Unlike other *R* GUIs no software (apart from a browser) is required *as SIMSURVEY* runs on a web server (http://bolmen.ethz.ch/~simsurvey/simsurvey/simProto.html). Currently, *SIMSURVEY* offers the following functionality:

- Data transformation and management,
- exploratory analysis of spatial data,
- linear regression analysis of spatial data and analysis of variance,
- estimation and modelling of variograms, and
- universal kriging.

Various kinds of graphical tools are available for all these tasks. All these analyses can be run by using the GUI. For experienced *R* users, SIMSURVEY contains in addition a command window. For educational purposes, *SIMSURVEY* allows one to sample and to analyse simulated soil pollution data.

SIMSURVEY was implemented by an interplay of *Adobe Flash*, *PHP* and *R*. The GUI by which a user interacts with *R* is a *Flash* animation in a browser window. The dynamically changing structure of the GUI is largely controlled by *XML* code. The actions of a user are passed to *R* by *PHP*. Based on template *R* code *PHP* dynamically generates complete *R* scripts that are processed by *R* processes running permanently on the web server. To improve the performance *PHP* and *R* communicate with each other by a socket connection. The output that *R* generates (text and graphic files) are then routed back to the Flash animation by *PHP* and are then presented in the browser to the user.

Thanks to its modular architecture, *SIMSURVEY* can be easily modified and extended. To this aim the following steps are required:

- Define the new items of the GUI by adding to the *XML* code. To facilitate this task predefined elements for text input fields, radio buttons, check boxes etc. can be used.
- Write the template *R* code for the new tasks.
- Extend the *PHP-R* interface to pass the required information from the GUI to *R* (by dynamically generating *R* scripts) and route the *R* output back to the GUI.

This architecture provides a novel and flexible framework for general computations with *R* on a web server.

In our presentation we shall demonstrate the use of *SIMSURVEY,* and we shall show by an example how SIMSURVEY can be extended for new tasks.

Bayesian generalized linear models and an appropriate default prior

Andrew Gelman
Columbia University

Many statistical methods of all sorts have tuning parameters. How can default settings for such parameters be chosen in a general-purpose computing environment such as R? We consider the example of prior distributions for logistic regression.

Logistic regression is an important statistical method in its own right and also is commonly used as a tool for classification and imputation. The standard implementation of logistic regression in R, `glm()`, uses maximum likelihood and breaks down under separation, a problem that occurs often enough in practice to be a serious concern. Bayesian methods can be used to regularize (stabilize) the estimates, but then the user must choose a prior distribution. We illustrate a new idea, the "weakly informative prior", and implement it in `bayesglm()`, a slight alteration of the existing R function. We also perform a cross-validation to compare the performance of different prior distributions using a corpus of datasets.

ChainLadder: Reserving insurance claims with R

Markus Gesmann

Libero Ventures Ltd
markus.gesmann@libero.uk.com

May 19, 2008

Abstract

One of the biggest liability items on an insurance company's balance sheet is the reserves for future claims payments. This reserve is an estimate of the amount an insurance company expects to pay for reported and unreported claims. Based on historical incurred claims and payment patterns, methods have been developed to forecast future payments.

The *ChainLadder* package provides the Mack-chain-ladder and Munich-chain-ladder methods to estimate reserves. The implementation in R allows both methods to be seen in a linear model context and therefore makes heavy use of the `lm` function in R.

The *ChainLadder* package grew out of presentations the author gave at the *Stochastic reserving and modelling seminar, 29 - 30 November 2007 at the Institute of Actuaries*.

Keywords: Claims reserving, Mack-chain-ladder, Munich-chain-ladder, linear models

Time Series Database Interface[*]

Paul D. Gilbert
Department of Monetary and Financial Analysis, Bank of Canada,
234 Wellington Street, Ottawa, Canada, K1A 0G9
pgilbert@bank-banque-canada.ca

Abstract

This presentation describes a package that abstracts an interface to time series databases, and a related group of packages that implement interfaces to SQL databases and to Fame through the PADI protocol. The TSdbi package, which implements the abstraction, imports the DBI namespace and DBI functions that support many SQL databases. For these cases there is limited need for code specialized to the specific database. This has been implemented in packages TSMySQL and TSSQLite, which require packages RMySQL and RSQLite respectively. It should also be possible to use the abstraction with RODBC (in progress but untested at the moment). TSdbi can also be used to interface to other time series databases, but this will typically require more database specific code below the abstraction. A working interface to Fame is implemented in the package TSpadi. It should also be possible to implement a more direct connection using the fame package.

Time series databases are typically simple in the sense that series are named with a unique identifier, and queries are limited to lookups using this key. From this perspective an SQL database is hardly needed. Apart from the abstraction, which is useful to make other code independent of the database implementation, the the main advantages are to use the database's client/server protocol, the ability to handle endian issues, and security features. However, when an SQL database is used, additional features can be added: it is possible to have vintage and panel dataset, with the same identifier used for different release dates and/or different panel members.

The package also (potentially) allows choice of the R representation to use for the time of the series data points. The default is ts where applicable, and zoo otherwise. (At the moment, only the default is working.)

The structure of the back-end SQL data bases and some utilities for implementing them will also be discussed.

[*] The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada.

The BLCOP package: an R implementation of the Black-Litterman and copula opinion pooling models.

Francisco Gochez – *Mango Solutions*

useR! 2008. Dortmund.

Abstract

In the early 1990s Fischer Black and Robert Litterman devised a framework for smoothly blending analyst views on the mean of the distribution of financial asset returns with a market "official equilibrium" distribution. The model has generated substantial interest since then, though it is limited by its assumptions of normality in market and analyst view distributions, as well as by vagueness in the meaning of certain parameters and their determination. In late 2005 Attilio Meucci of Lehman Brothers proposed the "copula opinion pooling" (COP) method as a generaliztion that overcomes all of these limiations, though at the cost of greater complexity. The BLCOP package is an implementation of both of these models. The emphasis of the package is on ease of use, flexibility, and allowing the user to easily analyze the impact of his or her views on the market posterior distribution.

References

[1] Meucci, Attilio. *Beyond Black-Litterman: Views on Non-normal Markets.* November 2005. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=848407

FAiR: A Package for Factor Analysis in R

Ben Goodrich (goodrich@fas.harvard.edu)

The primary objective of FAiR is to provide functionality that is not available in closed-source factor analysis software, but FAiR also strives to integrate the various tools for factor analysis that are already available in R packages and to provide a reasonably user-friendly GUI (based on gWidgets) so that people who have more experience with factor analysis than with R can readily estimate their models. The first version of FAiR was released in February 2008, and the second version will have been released in April 2008.

FAiR is unique in that it utilizes a genetic algorithm (rgenoud) for constrained optimization, which permits new approaches to exploratory and confirmatory factor analysis (EFA and CFA) and also straightforwardly leads to a new estimator of the common factor analysis model called semi-exploratory factor analysis (SEFA). The common factor analysis model in the population can be written as $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Psi}$, where $\mathbf{\Sigma}$ is a covariance matrix among n observable variables, $\mathbf{\Phi}$ is a correlation matrix among r common factors, $\mathbf{\Lambda}$ is a $n \times r$ matrix of factor loadings, and $\mathbf{\Psi}$ is a (typically diagonal) covariance matrix among n unique variances. However, $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ are not separately identified unless additional restrictions are imposed. For example, CFA requires the analyst to specify which cells of $\mathbf{\Lambda}$ are zero *a priori*. SEFA differs by requiring the analyst to specify the *number* of zeros in each column of $\mathbf{\Lambda}$ but does not require the analyst to specify *where* the zeros occur. SEFA thus uses a genetic algorithm to maximize the fit to the data over the locations of these exact zeros in $\mathbf{\Lambda}$ and the values of the corresponding non-zero parameters.

FAiR also differs from all other factor analysis software in that the analyst can impose a wide variety of (non-linear) inequality restrictions on (functions of) parameters in SEFA and CFA models and also during the transformation stage of EFA models. For example, Louis Thurstone — who was the father of exploratory factor analysis with multiple factors — proposed a criterion for factor transformation in 1935 that had never been implemented by any factor analysis software, largely due to its perceived computational difficulty. Optimizing with respect to Thurstone's criterion is implemented in FAiR, which is fairly easy and very reliable due to power of the underlying genetic algorithm.

FAiR utilizes S4 classes, which not only facilitates post-estimation analysis but also provides a framework that allows rapid development and easy integration of other R packages with FAiR. My goals for useR are to attract the interest of additional developers and to expose attendees (and their colleagues at home) to the new ways of thinking that are embodied in FAiR but are unavailable in traditional factor analysis software.

Statistical Modeling of Loss Distributions Using **actuar**

Vincent Goulet
École d'actuariat, Université Laval, Québec, Canada

actuar is a package providing additional Actuarial Science functionality to the R statistical system. The current version of the package contains functions for use in the fields of loss distributions modeling, risk theory (including ruin theory), simulation of compound hierarchical models and credibility theory. This talk will present the features of the package most closely related to usual statistical work, namely the modeling of loss distributions. Among other things, we introduce a number of probability laws functions, handling of grouped data, minimum distance estimation methods and functions to compute empirical moments. If time allows, we will also present the function to simulate data from compound hierarchical models.

Distributed Computing using the multiR Package

Daniel J Grose[1]

[1]Centre for e-Science , Lancaster University , United Kingdom

March 31, 2008

Abstract

There exist a large number of computationally intensive statistical procedures that can be implemented in a manner that is suitable for evaluation using a parallel computing environment. Within this number there exists a class of procedures, often described as "course grained parallel" or "embarrassingly parallel". The defining characteristic of these procedures is that they can be reduced to a number of sub-procedures that are independent of each other and require little or no inter-procedure communication i.e. they can be executed concurrently. Initially, it might be thought that this class is too small to warrant significant attention, however this is far from being the case. For example, methodologies such as bootstrapping, cross-validation, many types of Markov Processes (including MCMC), and certain optimisation and search algorithms are of this type. Importantly, the increase in availability of High Throughput Computing (HTC) environments, consisting of large numbers of interconnected computers, has made employing such procedures particularly attractive, leading to a significant increase in the amount of research being undertaken using HTC, notably in the areas of biochemistry, genetics, pharmaceuticals, economics, financial modelling and the social sciences.

A High Throughput Computing environment provides a means for processing a large number of independent (non-interacting) tasks simultaneously. In the simplest case, the HTC environment may employ only a single multi-processor system. At the other extreme, the HTC environment might comprise a large number of systems with different operating systems and hardware located across a number of different institutions and administrative domains. When this is the case the environment may be said to provide High Throughput Distributed Computing (HTDC).

HTC on a single multiprocessor system is relatively straightforward. Typically the user has an account on the system (can be identified to the system by a user name and password) and can submit the tasks for processing by using the software tools available on that system. Higher level means of submitting tasks exist, such as the **snow** package for R [1]. This package allows functions defined in R or installed R packages to be invoked multiple times with varying argument signatures and executed on a number of processors simultaneously. In [1] it is noted that the functionality offered by **snow** could be extended to use the GRID, which by its nature provides a HTDC environment. Some of these extensions have been addressed within the GridR system [3], which is similar in principle to **snow** but provides some of the technical requirements necessary

for using GRID based resources which it achieves by employing the COG toolkit [2].

However, there are a number of important considerations which arise when using generalised HTDC (GRID based or otherwise) not all of which have been encapsulated in either **snow** or GridR. These considerations are

1. A client session may terminate before all tasks have been processed. For instance, the results of the completed tasks may need to be collected in a future client session, possibly from a different system.
2. The systems employed to process the tasks may be multi-fold and reside in different administrative domains, thus it is not practical for a client to have to obtain and manage accounts on all (potentially hundreds or even thousands) of these systems. Consequently, a single means of identifying the client is required.
3. The client system must employ a secure channel for communication.
4. Host systems are typically shared by many clients and have scheduling systems to allocate resources, thus the execution time and the order in which tasks are processed may vary.
5. Individual tasks may fail to complete (this is quite common on certain systems, such as Condor pools).
6. The client interface should be independent of the nature of the distributed systems used for undertaking the computation.

All of these considerations have been well studied in many varied contexts and the design pattern most associated with realising the above design criteria is a three-tier client server employing the public key infrastructure for authentication and security. The technologies required for implementing such an architecture to host a HTDC service for R are readily available and have been used to develop servers which expose an interface for use within a client R session. The **multiR** package contains an implementation of a client interface for use in R which is similar in many respects to that of **snow** and GridR in that it extends the **apply** family of functions (available in the base package) for submitting multiple function invocations in a distributed environment. **multiR** also provides the functionality required to generate certificate based proxy credentials, manage active jobs and harvest results when they become available. Importantly, the interface provided by **multiR** is independent of the many different types of hardware and software systems employed within a HTDC environment and requires no additional software components (Globus, CoG and so on) to be installed before it can be used.

The full presentation of this work demonstrates how **multiR** is installed and used using several example applications which include bootstrapping, calculating multivariate expectation values and function optimisation. For each of the examples the benefits of using **multiR** are examined, with particular reference to the reduced time required to compute them.

References

[1] A. J. Rossini, Luke Tierney, and Na Li. Simple parallel statistical computing in R. *Journal of computational and Graphical Statistics*, 16(2):399–420, June 2007.

[2] Gregor von Laszewski and Mike Hategan. Workflow concepts of the Java CoG kit. *Journal of Grid Computing*, 3(3–4):239–258, 2005.

[3] Denis Wegener, Thierry Senstag, Stelios Sfakianakis, Stefan Ruping, and Anthony Assi. GridR: an R-based grid-enabled tool for data analysis in ACGT clinico-genomic trials. In *Third IEEE International Conference on e-Science and Grid Computing (e-Science 2007), Bangalore, India.*, pages 205–212. IEEE, 2007.

FlexMix: Flexible fitting of finite mixtures with the EM algorithm

Bettina Grün
Wirtschaftsuniversität Wien

Friedrich Leisch
Ludwig-Maximiliansuniversität München

Finite mixtures are a flexible model class for modelling unobserved heterogeneity or approximating general distribution functions. The R package **flexmix** provides infrastructure for fitting finite mixture models with the EM algorithm or one of its variants. The main focus is on finite mixtures of regression models and it allows for multiple independent responses and repeated measurements. Concomitant variable models as well as varying and constant parameters for the component specific generalized linear regression models can be fitted.

The main design principles of the package are easy extensibility and fast prototyping for new types of mixture models. It uses S4 classes and methods and exploits features of R such as lexical scoping. The implementation of the package is described and examples given to illustrate its application.

SimpleR: Taking on the “Evil Empire” by Developing Applications for Non-statistical Users

The “Standard Model” for R software development – the R package – assumes:

1. Many independent users;
2. Some degree of statistical understanding by users;
3. Access and functionality within the overall R environment;
4. Persistence over time;
5. (Usually) A command line interface.

This model serves serious data analysts well, and R provides many tools to facilitate such development. We argue here that R can also serve another and potentially much larger community of engineers and scientists for whom Excel® (or similar software) is now the primary tool for data analysis and statistical graphics. These folks need:

1. Narrow, often single purpose “template” analyses;
2. Rapid development/rapid discard – needs disappear as soon as a project is completed or change radically when technology changes;
3. Software customized for a few – even one – users;
4. A simple GUI interface requiring minimal documentation and learning;
5. Graphs as the primary output.

Excel is their default because (a) it’s there and they know it; (b) they don’t know about or can’t implement better methods.

R can change this state of affairs. R is open source, has superb graphics, and is easily embedded into web served applications using R2HTML, Rpad, Rserve, Rzope, etc.. Alternatively, it can be modified for single use applications through a GUI such as Rcmdr, gWidgets, or by simply modifying the R menu structure. The key is to present the user with a simple interface and readily interpretable output, even if the underlying analysis is complex.

We discuss our strategy for developing such applications, which relies on the global workspace as the software environment, thus avoiding the unnecessary (for us) overhead of packages. We give an example in actual use at Genentech and discuss the pros and cons of this approach.

Estimation in classic and adaptive group sequential trials

Niklas Hack[1] and Werner Brannath[2]

[1] Section of Medical Statistics, Medical University of Vienna, Vienna, Austria
niklas.hack@meduniwien.ac.at
[2] Section of Medical Statistics, Medical University of Vienna, Vienna, Austria
werner.brannath@meduniwien.ac.at

Abstract. We present a R-package for estimation in classic and adaptive group sequential trials. We will give an overview of classic and adaptive group sequential designs and will present two methods for the calculation of p-values and confidence intervals. The first method is based the repeated approach of Jennison and Turnbull (1984), which was extended by Mehta, Bauer, Posch and Brannath (2006) to the adaptive setting. The second method is based on the stage-wise ordering of Tsiatis, Rosner and Mehta (1989), which was extended by Brannath, Mehta and Posch (2008) to the adaptive setting. The key idea of both methods, based on the method of Müller and Schäfer (2001), is to preserve the overall type I error rate after a possible design adaptation, by preserving the null conditional rejection probability of the remainder of the trial at each time of an adaptive change. The implementation and the application of these methods in R (available in package AGSDest) will be illustrated.

References

Brannath, W and Mehta, CR and Posch, M (2007): Exact confidence bounds following adaptive group sequential tests, submitted.

Jennison, C and Turnbull, BW (1984): Repeated confidence intervals for group sequential clinical trials, *Contr. Clin. Trials, 5, 33-45*

Mehta, CR,Bauer, P, Posch, M and Brannath, W (2006): Repeated confidence intervals for adaptive group sequential trials, *Statistics in Medicine, 26, 5422-5433*

Müller, HH and Schäfer, H (2001): Adaptive group sequential design for clinical trials: Combining the advantages of adaptive and of classic group sequential approaches, *Biometrics, 57, 886–891.*

Tsiatis, AA and Rosner, GL and Mehta CR (1984): Exact confidence intervals following a group sequential test, *Biometrics, 40, 797–804.*

Keywords

ADAPTIVE GSD, CONFIDENCE INTERVALS, POINT ESTIMATES

Introducing BioPhysConnectoR

Kay Hamacher[1]*, Franziska Hoffgaard, Phillip Weil*
Technische Universität Darmstadt
AG Bioinformatics and Theoretical Biology, Fachbereich Biologie
Schnittspahnstr. 10, 64287 Darmstadt, Germany

The biggest challenge for systems biology and bioinformatics in the post-genome area is the integration of countless experimental data such as sequence information, gene expression data, physio-chemical values, phylogenetic relationships, or physiological data.

With R researchers command over an efficient framework for statistical modelling and accordingly R became – with the event of Bioconductor at the latest – the major platform for analyzing biostatistical data. Up to now much effort has been invested in the statistical modelling and subsequent implementation of *information-driven* packages, and protocols. This allowed tremendous progress in understanding *information* contained within e.g. biological sequence data. Experiments are nowadays guided to a large extend by the knowledge gained from such protocols.

The *information* contained within biological sequences reflects the whole evolutionary history of the organism under investigation (including external selective pressure such as drugs and resistance development). The selection step of every evolutionary process is, however, an event in the *physical realm* as selection tests the physiochemical properties of molecules involved in relevant processes.

Therefore, to construct molecular interaction networks [1], there is a pressing need to connect *information* (the evolutionary memory) with *the physical realm*, its forces, the molecular dynamics and mechanics (the selective „horizon“).

We achieve this with our ongoing efforts [2] in integrating standard sequence/statistical-model-driven methodologies with new reduced-molecular-models derived from biophysical interaction theories [3,4], eventually bridging the gap between bioinformatics and molecular dynamics simulations/molecular biophysics. We developed an R-package (BioPhysConnectoR) to this end. With this package we connect the information space and the physical space – thus allowing for functional annotation of sequence data and systematic *in silico* experiments. Additional useful functions for dealing with sequences and matrices are provided within the package.

We integrated C-code with R-routines and found that regarding the run-time efficiency our packages compares perfectly with our original code in C/FORTRAN. Due to the abstraction offered by R and leveraging the power of the packages Rmpi and papply, we were able to implement the package in a massively parallelized fashion.

As it is possible in R to interactively examine the results of the computations, this allows for both large-scale screening and high-throughput-scans on the one hand and online, interactive method development and hypothesis testing on the other. We discuss future research directions.

References

[1] K. Hamacher, J. Trylska, J.A. McCammon. *Dependency Map of Proteins in the Small Ribosomal Subunit.* **PLoS Comput. Biol.** 2(2): e10, 2006

[2] K. Hamacher. *Information Theoretical Measures to Analyze Trajectories in Rational Molecular Design*, **J.Comp.Chem.**, 28(16):2576-2580, 2007.

[3] K. Hamacher, J.A. McCammon. *Computing the amino acid specificity of fluctuations in biomolecular systems*, **J.Chem.Theo.Comp.** 2:873, 2006.

[4] K. Hamacher, *Relating Sequence Evolution of HIV1-Protease to Its Underlying Molecular Mechanics,* Gene, accepted, 2008

[1] *corresponding author, hamacher(AT)bio.tu-darmstadt.de, http://www.kay-hamacher.de*

Modelling biodiversity in R: the untb package

Robin Hankin

The distribution of abundance amongst species with similar ways of life is a classical problem in ecology.

The Unified Neutral Theory of Biodiversity (UNTB), due to Hubbell, states that observed population dynamics may be explained on the assumption of per capita equivalence amongst individuals. One can thus dispense with differences between species, and differences between abundant and rare species: all individuals behave alike in respect of their probabilities of reproducing and death.

It is a striking fact that such a parsimonious theory results in a non-trivial dominance-diversity curve (that is, the simultaneous existence of both abundant and rare species) and even more striking that the theory predicts abundance curves that match observations across a wide range of ecologies.

The UNTB, being a statistical hypothesis, is well-suited to simulation using the R computer language. Here I discuss the untb package for numerical simulation of ecological drift under the unified neutral theory. A range of visualization, analytical, and simulation tools are provided in the package and these are presented with examples and discussion.

JavaStat: a Java-based R Front-end

E. James Harner Dajie Luo Jun Tan

Architectures are described which allow a Java-based front-end to run R code on a server. The front-end is called JavaStat (`http://javastat.stat.wvu.edu`), a Java application. JavaStat is a highly-interactive program for data analysis and dynamic visualization with data management capabilities. The objective is to bring the high-level functions of R to JavaStat without excessive duplicative development work. Results returned from R are wrapped and then displayed using dynamic graphics in JavaStat.

The principal idea is to use RMI (Remote Method Invocation) to communicate with a Java server program (JRIServer), which in turn communicates with R using JRI (Java/R Interface). Two versions have been implemented. The first architecture maintains a connection between the client and server in order to return the results from R. This is suitable for small to moderate data sets in which statistical models are run. The second architecture queues the requests and uses polling to fetch the results. It is suitable for large data sets and complex models, e.g., those encountered in genomic studies.

Models for Replicated Discrimination Tests: A Synthesis of Latent Class Mixture Models and Generalized Linear Mixed Models

Rune Haubo Bojesen Christensen & Per Bruun Brockhoff
DTU Informatics, Statistics Section, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark, rhbc@imm.dtu.dk

Discrimination tests are often used to evaluate if individuals can distinguish between two items. The tests are much used in sensory and consumer science to test food and beverage products, and in psychophysics to investigate the cognitive strategies of the mind. Signal detection theory, experimental psychology and medical decision making are other areas, where the tests are applied. The basic idea is to use humans as instruments to measure attributes or differences between products (eg. Lawless and Heymann, 1998). In sensory and consumer science a panel of judges or a sample of consumers are employed, but humans are difficult to calibrate and much variation remains between individuals.

Often respondents perform the test several times and because subjects tend to have different discriminal abilities, this leads to overdispersion in the data. Traditionally this is handled by marginal models where the amount of overdispersion is estimated in order to adjust standard errors.

Commonly used discrimination tests can be identified as generalized linear models (GLMs) with the so called psychometric functions (Frijters, 1979) as inverse link functions (Brockhoff and Christensen, 2008). This makes generalized linear mixed models (GLMMs) available to model the variation between subjects.

The inverse psychometric functions maps the probability of a correct answer in the discrimination test to a measure of discriminal ability, which becomes an intercept parameter in a GLM or GLMM. Since the discriminal ability is a non-negative quantity, the random effect distribution in a GLMM consists of a point mass at zero and a continuous positive part. The resulting model can be seen as a synthesis of a latent class mixture model and a generalized linear mixed effect model. We have implemented functions that will fit the proposed model in R.

Interest is often in characterizing the variation between subjects and in obtaining estimates of individual discriminal abilities. Both sets of quantities are available from the proposed model as a variance component and posterior modes respectively. Also available is an estimate of the proportion of discriminators in the population as well as an estimate of the probability that each individual is a discriminator.

This presentation will introduce models for replicated discrimination tests, show how to fit them in R and consider important properties of the models. We end with an example from sensory science showing how to interpret the results.

References

Brockhoff, P. B. and R. H. B. Christensen (2008). Thurstonian model as generalized linear models. *Food Quality and Preference* . In prep.

Frijters, J. E. R. (1979). Variations of the triangular method and the relationship of its unidimensional probabilistic models to three-alternative forced-choice signal detection theory methods. *British Journal of Mathematical and Statistical Psychology 32*(229-241).

Lawless, H. T. and H. Heymann (1998). *Sensory evaluation of food.* Chapman and Hall, London.

SpRay - an R-based visual-analytics platform for large and high-dimensional datasets

J. Heinrich[1,2], J. Dietzsch[1], D. Bartz[2] and K. Nieselt[1]

[1]Center for Bioinformatics, University of Tübingen Sand 14, 72076 Tübingen, Germany

email: {juheinri, dietzsch, nieselt}@informatik.uni-tuebingen.de

[2]ICCAS/VCM, University of Leipzig, Germany

email: dirk.bartz@medizin.uni-leipzig.de

March 31, 2008

Recently developed high-throughput methods produce increasingly large and complex datasets. For instance, microarray-based gene expression studies generate data for several thousands of genes under numerous different conditions, yielding large, heterogeneous, potentially incomplete or conflicting datasets. From both technical and analytical points of view, extracting useful and relevant information - known as the knowledge discovery process - from these large data sets is a challenge. While the technical capacity to collect and store such data grows rapidly, the ability to analyze it does not advance at the same pace. The extraction of relevant information from large and high-dimensional data is very difficult and requires the support of automated extraction algorithms based on statistical computing. Unfortunately, the unsupervised application of these statistical measures does not guarantee the successful extraction of relevant information, but requires critical consideration itself. Hence, the use of interactive visualization methods for the simultaneous evaluation of the applied statistical models is of central relevance and plays therefor a key role in the emerging field of visual analytics.

The aim of the work is to combine statistical methods with modern visualization techniques in an extendable, hardware-accelerated visual-analytics framework. We are currently developing SpRay (viSual exPloRation and AnalYsis of high-dimensional data), which provides for the explorative analysis of large, high-dimensional datasets in accordance with the visual-analytics paradigma. Similar to GGobi [SLBC03], the statistical backend is provided through R, as a plugin. The performance-oriented design of SpRay, which uses hardware-accelerated graphics (OpenGL), C++ and Qt, also allows very large datasets to be explored with greatly reduced response times. The use of modern GPUs (OpenGL) further accelerates the application of different transparency-modulations and color maps to the currently implemented plugins, such as refined parallel coordinates and scatterplots. All plugins (currently: parallel coordinates, scatterplots, TableLens, TableView, Histogram, R-Console, Brushing) are linked by means of a common data model which is particularly useful to tightly integrate R along with all its extensions via packages. Hence, adequate statistical values may be defined and interactively visualized together with the raw data, providing an iterative, interactive and integrated approach to the analytical reasoning process as proposed by the visual-analytics-paradigm. The benefit of the currently implemented features has succesfully been demonstrated with different gene-expression datasets [DHNB06, DHNB08].

References

[DHNB06] J. Dietzsch, J. Heinrich, K. Nieselt, and D. Bartz. Poster: Extended parallel coordinates for bioinformatical applications. In *German Conference on Bioinformatics (GCB)*, Tübingen, 2006.

[DHNB08] J. Dietzsch, J. Heinrich, K. Nieselt, and D. Bartz. Visual Analysis of Microarray Data from Bioinformatics Applications. Technical Report WSI-2008-1, ISSN 0946-3852, Dept. of Computer Science (WSI), University of Tübingen, 2008.

[SLBC03] D. Swayne, D. Lang, A. Buja, and D. Cook. GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization. *Computational Statistics & Data Analysis*, 43:423–444, 2003.

NADA for R. A contributed package for censored environmental data

Dennis R. Helsel Lopaka Lee
US Geological Survey

Trace contaminants in air, water biota, soils, and rocks often contain data recorded only as a "nondetect", or less than a detection threshold. These left-censored values cause difficulties for environmental scientists, as no single number can be validly assigned to them. The typical solution of substituting one-half the detection limit and proceeding with regression, t-tests, etc., has repeatedly been shown to be inaccurate. Instead, these data can be effectively interpreted using survival analysis techniques more traditionally applied to right-censored data. Methods for calculating descriptive statistics, testing hypotheses, and performing regression, both parametric and nonparametric, are available using the contributed package NADA. Methods include censored maximum likelihood (ML), Kaplan-Meier, and the Akritas version of Kendall's robust line that is applicable (unlike ML) to doubly-censored data. Methods such as censored boxplots and residuals plots that can graph data containing nondetects are also included. The NADA package complements the first author's textbook, *Nondetects And Data Analysis: Statistics for censored environmental data* (Wiley, 2005).

High Performance Computing with NetWorkSpaces for R

David Henderson (dnadave@revolution-computing.com)
Stephen Weston
Nicholas Carriero
Robert Bjornson

Increasingly, R users have access to multiprocessor machines or multiple-core CPUs. However, base R does not natively support parallel processing; this can force R users to wait while computationally intensive work is done on a single processor or core and other processors or cores lie idle. NetWorkSpaces for R (NWS-R) was developed at Scientific Computing Associates, the predecessor to REvolution Computing. It is a Python-based coordination system that is portable across virtually all popular computing platforms. NWS-R includes a web interface that displays the workspaces and their contents; this is helpful when debugging or developing a program, or monitoring the progress of an application. NWS-R is easy to learn, accessible from many development environments, and deployable on ad hoc collections of spare CPUs. The server and client for NWS-R are available at SourceForge (nws-r.sourceforge.net); the client is also available at CRAN (cran.r-project.org/web/packages/nws/). We will present NetWorkSpaces for R and demonstrate the web interface.

Providing R functionality through the OGC Web Processing Service

Katharina Henneböhl and Edzer Pebesma
Institute for Geoinformatics (IfGI)
University of Münster

The new Web Processing Service (WPS) 1.0 standard, recently released by the Open Geospatial Consortium (OGC, opengeospatial.org), specifies how GIS calculations can be made available via the Internet, as web services. Operations can be as simple as adding two map layers or as complex as running a full hydrological model. This opens the possibility of providing the spatial analysis functionality available in R also through this interface. In this paper we will show how a connection between the open-source Java-based WPS reference implementation from 52North (52north.org) and R can be established, and how R functionality can be exposed through an OGC-compliant web service. The question how to exchange spatial datasets between Java and R is of special interest.

Estimation of Theoretically Consistent Stochastic Frontier Functions in R

Arne Henningsen, University of Kiel

March 31, 2008

Conventional econometric analysis in the field of production economics generally assumes that all producers always manage to optimize their production process. Least squares-based regression techniques attribute all departures from the optimum exclusively to random statistical noise (Kumbhakar and Lovell, 2000). However, producers do not always succeed in optimizing their production. Therefore, the framework of "Stochastic Frontier Analysis" (SFA) has been developed that explicitly allows for failures in producers' efforts to optimize their production (Kumbhakar and Lovell, 2000).

Stochastic frontier analysis is generally based on production, cost, distance, or profit functions. Microeconomic theory implies several properties of these functions. Sauer *et al.* (2006) show that consistency with microeconomic theory is important especially for estimating efficiency with frontier functions. Although theoretical consistency is required for a reasonable interpretation of the results, these conditions are not imposed in most empirical estimations of stochastic frontier models — probably because the proposed procedures to impose these conditions are rather complex and laborious. Recently, a much simpler three-step procedure that is based on the two-step method published by Koebel *et al.* (2003) has been proposed by Henningsen and Henning (2008). We show how theoretical consistent stochastic frontier functions can be estimated in R using this new procedure. This is illustrated by estimating a stochastic frontier production function with monotonicity imposed at all data points.

References

Henningsen A, Henning CHCA (2008). "Estimation of Theoretically Consistent Stochastic Frontier Functions with a Simple Three-Step Procedure." Unpublished manuscript, Department of Agricultural Economics, University of Kiel.

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.

Kumbhakar SC, Lovell CK (2000). *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge.

Sauer J, Frohberg K, Hockmann H (2006). “Stochastic Efficiency Measurement: The Curse of Theoretical Consistency.” *Journal of Applied Economics*, **9**(1), 139–165.

Metabolome data mining of mass spectrometry measurements with random forests

Chihiro Higuchi [*]and Shigeo Takenaka [†]

Metabolome analysis is expected to become a leading technology for rapid discovery of novel biomarkers, which are key components for successful drug development. Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are frequently employed as effective tools for metabolome measurements, and when it comes to analysis, the principal component analysis and the partial least square methods have been the methods of choice for mining of metabolome data.

In the present study, we have investigated the application of the random forests machine learning method (Breiman 2001) for analysis of metabolme data. The data comprised FT-ICR-MS measurements of urine from rats, which have been administered the antiarrhythmic agent amiodarone. Amiodarone treated rats will exhibit lipidosis and phenylacetylglycine (PAG) can be measured in the urine.

Unsupervised classification applied to these data with the random forests approach clearly separated the groups, that is before and after amiodarone treatment, and the separation was superior to that of the principal component analysis method. The supervised classification with the random forests approach furthermore suggested several class discriminating MS peaks, which were selected by the importance value generated by the random forests machine learning method. These MS peaks were assigned biomarker candidates and ranked by the loading values from the principal component analysis.

This analysis was carried out with the randomForest, amap and Heatplus packages of R 2.4.1 on Linux (kernel 2.6.21) operating system.

[*] Genomic Science Laboratories, Dainippon Sumitomo Pharma Co., Ltd., Suita-city, Osaka 564-0053, Japan

[†] Graduate School of Life and Environmental Sciences, Osaka Prefecture University, Sakai-city, Osaka 599-8531, Japan

RcmdrPlugin.epack: A Time Series Plug In for *Rcmdr*

Erin Hodgess and Carol Vobach
Department of Computer and Mathematical Sciences,
University of Houston - Downtown, Houston, TX

ABSTRACT

In many statistics courses, R has excellent facilities but the learning curve can be somewhat daunting for undergraduates. Fox(2005) has overcome some of these hurdles with the *Rcmdr* package, which provides menu-driven options for regression in particular. *Rcmdr* also provides options for most functions found in basic statistics classes and is supplemented by Heiberger and Holland (2007), with their *RcmdrPlugin.HH* package.

R has nearly all of the typical functions used in undergraduate time series courses. Even with these functions available from the command line, students still balk at command line use. This new package, *RcmdrPlugin.epack*, provides sets of submenus for student use in an undergraduate time series courses. *RcmdrPlugin.epack* promotes ease of use and permits students to use their efforts to understanding concepts rather than programming. Students can develop models for both explanatory and forecasting purposes.

Key Words: time series, menus, *Rcmdr*, R

Modelling and surveillance of infectious diseases - or why there is an R in SARS

Michael Höhle[1,2]

[1] Department of Statistics, University of Munich, Germany
michael.hoehle@stat.uni-muenchen.de
[2] Center for Mathematical Sciences, Munich University of Technology, Germany

Abstract. This talk will focus on how R could assist in two aspects of the continuing efforts to better understand and control infectious diseases - be it in human, plant or veterinary epidemiology.

Firstly, stochastic modelling is an important tool in order to better understand the dynamics of infectious diseases. A key epidemic model in this process is the stochastic susceptible-exposed-infectious (SIR) model. The R package `RLadyBug` contains a set of functions for the simulation and parameter estimation in spatially heterogeneous SIR models. Simulation is based on the Sellke construction or Ogata's modified thinning algorithm, while estimation is based on maximum likelihood or - when the disease is only partially observed - Markov Chain Monte Carlo.

Secondly, routine surveillance of public health data often boils down to the on-line detection of change-points in time series of counts. Surveillance has hence a close connection to problems from statistical process control. The R package `surveillance` contains an implementation of some of the most common surveillance methods such as the Farrington procedure or cumulative sums. Data and results can be temporally and - in case of multiple time series - spatio-temporally visualized.

Both packages are introduced and their use is illustrated by means of examples and R-code.

References

HÖHLE, M. (2007): surveillance: An R package for the surveillance of infectious diseases, *Computational Statistics, 22(4), pp. 571-582.*

HÖHLE, M. and FELDMANN, U. (2007): RLadyBug – An R package for stochastic epidemic models. *Computational Statistics and Data Analysis, 52, 680-686.*

Keywords

infectious disease epidemiology, SIR model, outbreak detection

Variable Selection and Model Choice in Survival Models with Time-Varying Effects

Benjamin Hofner, Thomas Kneib & Torsten Hothorn

Institut für Statistik, Ludwig-Maximilians-Universität München, Munich, Germany;
benjamin.hofner@stat.uni-muenchen.de

Abstract: Flexible hazard regression models based on penalised splines allow to extend the classical Cox-model via the inclusion of time-varying and nonparametric effects (Kneib & Fahrmeir 2007). Despite their immediate appeal in terms of flexibility, these models introduce additional difficulties when performing model choice and variable selection.

Boosting (cf. Bühlmann & Hothorn, 2008, and Tutz & Binder, 2006) supports model fitting for high-dimensional data. By using component-wise base-learners, variable selection and model choice can be performed in the boosting framework.

We introduce a boosting algorithm for survival data that permits the inclusion of time-varying effects in a parametric form or in a flexible way, using P-splines. Thus we can fit flexible, additive hazard regression models and have a fully automated procedure for variable selection and model choice at hand.

The properties and performance of the algorithm are investigated in simulation studies. In an application, we present the analysis of retrospective data of surgical patients with severe sepsis were the aim was to build a flexible prognostic model.

References

Kneib, T. & Fahrmeir, L. (2007) A mixed model approach for geoadditive hazard regression. *Scand. J. Statist.*, **34**, 207–228.

Bühlmann, P. & Hothorn, T. (2008) Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, Accepted.

Tutz, G. & Binder, H. (2006) Generalized additive modelling with implicit variable selection by likelihood-based boosting. *Biometrics*, **62**, 961–971.

Keywords

boosting, survival analysis, time-varying effects, P-splines, model choice, variable selection

The Past, Present, and Future of the R Project – Development in the R Project

Kurt Hornik

The development of R is a multi-tiered process, with a core team providing a base system only, and even key statistical functionality available via contributed extension packages. We review some of the basic milestones of this development process, discuss current patterns, and speculate on what the future might have in store. Particular emphasis is given to the fact that the number of available R packages keeps growing at amazing speed, making it increasingly challenging for both users and developers to deal with the size and the complexity of the R project. Given the complexity of the social network underlying R, we emphasize that no single information technology solution can satisfy the R community's desire for information, and discuss how and by whom the community might be served.

Good Relations with R

Kurt Hornik and David Meyer

April 7, 2008

Relations are a very fundamental mathematical concept: well-known examples include the linear order defined on the set of integers, the equivalence relation, notions of preference relations used in economics and political sciences, etc. A k-ary (finite) relation is defined by its *domain*, a k-tuple of sets, and its *graph*, a set of k-tuples. Package **relations** provides data structures along with common basic operations for relations and relation ensembles (collections of relations with the same domain). In doing so, it builds on the infrastructure for sets and tuples provided by package **sets**. Package **relations** also features various relational algebra-like operations, such as projection, selection, and joins. Finally, it contains algorithms for finding suitable consensus relations for given relation ensembles, including the constructive approaches of Borda, Condorcet and Copeland, as well as optimization-based methods which minimize the aggregate symmetric difference distance between the ensemble members and their consensus. We show how relations can be obtained and manipulated, and how the functionality in the package can be employed to rank the results of benchmarking experiments.

An extension of the coin package for comparing interventions assigned by dynamic allocation

Johannes Hüsing*

Restricted randomisation or algorithm-based allocation procedures enjoy some popularity among clinical researchers, promising a lower variance of the treatment effect estimate and balanced subgroups for exploratory analysis. They have met criticism because classical asymptotics don't hold and the argument for a random distribution may be less soundly based. This has led to a statement of mistrust in the form of a guideline issued by pharmaceutical regulators.

Permutation tests give rise to analysis strategies which incorporate the allocation strategy in order to generate more realistic null distributions. The plethora of published allocation algorithms calls for a common framework which can be used regardless of the algorithm employed. The package `coin` currently offers complete randomisation and balanced block randomisation as alternative procedures.

An extension of `coin` is introduced which allows users to consistently write new allocation procedures. The interface of the coin extension is defined so that algorithms can be used to be used both in treatment allocation service programs and in the reallocation procedure. Following this requirement, algorithms should be formulated in an incremental way, returning only the next allocation instead of the whole vector.

Algorithms should accept as parameters all previously allocated treatments and the common distribution of all factors the allocation decision is based on. It is passed as a data frame which contains all factors and the treatments. Treatment is null for the last observation, which is subject to the current allocation. The completed data frame is returned.

The interface to coin is confined to the `ApproxNullDistribution` method. Two additional arguments, `algorithm` (defaulting to "full permutation") and `shuffle` (sampling from alternative accrual sequences, defaulting to "identity") are passed. an engine is started which applies the (incrementally formulated) algorithm sequentially to the set of patients, ie. the (possibly shuffled) `x` slot of the `IndependenceTestStatistic` object.

It is hoped that the introduction of a common interface may encourage the use of dynamic allocation methods, and increase the acceptance for the results gained from appropriate analyses of data obtained this way.

*Koordinierungszentrum für Klinische Studien, Universität Heidelberg

R for climate research

Thomas Jagger James Elsner

We demonstrate using examples from our recent research papers that the R statistical language and its packages are excellent tools for climate research. The development of our expertise in R is based on the need to perform statistical analysis on climate data in research and industry. We show examples based on our work with hurricane activity and climate. Each example uses analytical and graphical functions. We demonstrate the use of

1. glm and associated functions for exploring the relationship between climate and hurricane activity.
2. analysis and graphing functions from the ismev package for exploring the role of climate on hurricane intensity.
3. graphical functions developed for selecting hurricane tracks and local wind maximums.
4. quantile regression functions from the quantreg package for exploring the relationship between increased sea surface temperature and global tropical storm intensity.
5. functions from the BRugs package for accessing OpenBugs used for analyzing the relationship of insured losses due to hurricanes to global climate covariates.

In each case we explain the advancements made in understanding the role that climate plays in the nature of tropical storm activity and insured losses from these storms.

All demonstrations will be available on our Hurricane Climate website at: `http://garnet.acns.fsu.edu/~jelsner/www/`

The Execution Engine: Client-server mechanism for remote calling of R and other systems

John James, Fan Shao – *Mango Solutions*

useR! 2008. Dortmund.

Abstract

A recent project highlighted the need to execute and manage remote jobs over a range of servers (including via a Grid API). The software systems on which jobs were to be managed included R, NONMEM and openBUGS (as well as other internal systems). In order to meet this demand, Mango created a software component we call the *Execution Engine*.

This presentation will discuss the design challenges in the development of the execution engine, with particular focus on the execution of R for model-based reporting. This will include interfaces for allowing control over R command options, and a developer interface in order to fully customize the application for other users.

This developed component can be thought of as a general client-server framework for R and other tools. It has now been used in a number of projects and is continuing to be developed.

ROMP – an OpenMP binding for R

Ferdinand Jamitzky

A binding of the parallel application programming interface OpenMP for the R-Interpreter is presented. Fortran code is generated and compiled on the fly by the toolkit and the OpenMP directives are inserted. The toolkit consists of a family of special `apply` routines together with reduction routines like `sum`, `mean`, `product` which generate parallel OpenMP code. The toolkit can be used for easy parallelization of parts of an R program without a steep learning curve for the user. Examples are presented which implement a systolic loop in RMPI and in the ROMP-toolkit.

Rapid Application Deployment with R

Wayne R. Jones

Shell Global Solutions (UK)
Shell Technology Centre Thornton,
P.O. Box 1, Chester CH1 3SH,
United Kingdom

Marco Giannitrapani

In this presentation we discuss our experiences of deploying R client based solutions. We demonstrate that R packages such as “R-(D)Com” and “RODBC” that allow communication with other applications (e.g. Excel) together with R-Gui packages such as “rpanel” and “tcltk” makes for a very powerful combination of tools for building customised statistical applications. Thanks, in part, to the very concise nature of R-programming such applications can be very quickly developed with extreme ease making previously unviable consultancy projects due to cost/benefit or manpower constraints achievable. The scope for using R based applications within Shell is enormous and we have already deployed numerous diverse solutions across all areas of the business including: Monte Carlo simulation tools, Customised Data Visualisation Tools, Forecasting toolbox, Groundwater monitoring application, automatic report generation and curve fitting to name but a few.

EpiR: a graphic user interface oriented to epidemiological data analysis

Washington Leite Junger, Antonio Ponce de Leon, Elizabeth Maciel de Albuquerque, Reinaldo Marques, Leonardo Costa

The use of R is rapidly growing among Brazilian graduate students as well as academic researchers, especially in public health sciences. Several post graduate programs have recently replaced proprietary software by R. In addition, the personnel from official epidemiological surveillance services is being trained to use R in the routine analyses, together with other software, like Epi-Info, as part of the current Brazilian government effort towards open source solutions. Despite R's power and flexibility, the absence of a graphic user interface (GUI) still refrains from adopting R as the main environment for data analysis, hence creating a demand for the development of an R GUI oriented to some applied statistical analysts. The aim of the Epi-R project is to fill this gap in public health.

The interface is being developed as both a package and a standalone application. The functions library is separated from the GUI, so commands can be issued either by point and click or command line. The GUI is developed over RGtk2 package and built with *libglade*. GTk widgets look nice and it is fairly stable running on any operating system. There are four main modules designed for data management (which also include a front end for ODBC connections and a recycle bin), data description and statistical modelling, graphical display and Epidemiology specific analyses. The library core relies on the functions available from several existing R packages as well as some homemade ones. A plug-in API is also being developed so the GUI may be easily extended and to keep the code light and clean. Besides the usual R help pages for the functions, an alternative help system for the GUI is available and information about the resources available in an open window can be obtained directly from such a window. The development of a Portuguese version of EpiR is supported by the Brazilian Ministry of Health. The EpiR package will be submitted to CRAN as of the acceptance of this paper.

A Toolbox for Bicluster Analysis in R

Sebastian Kaiser and Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München,
Ludwigstrasse 33, 80539 München, Germany,
firstname.lastname@stat.uni-muenchen.de

Abstract. Over the last decade, bicluster methods have become more and more popular in different fields of two way data analysis and a wide variety of algorithms and analysis methods have been published. In this paper we introduce the R package `biclust`, which contains a collection of bicluster algorithms, preprocessing methods for two way data, and validation and visualization techniques for bicluster results. For the first time, such a package is provided on a platform like R, where data analysts can easily add new bicluster algorithms and adapt them to their special needs.

Keywords: Biclustering, Two-Way-Clustering, Software, R

References

BARKOW, S., BLEULER, S., PRELIC, A., ZIMMERMANN, P., and ZITZLER, E. (2006): Bicat: a biclustering analysis toolbox. *Bioinformatics, 22,1282–1283.*

CHENG, Y. and CHURCH, G. M. (2000): Biclustering of expression data. In: *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 1,93–103.

KLUGER, Y., BASRI, R., CHANG, J. T., and GERSTEIN, M. (2003): Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research, 13,703–716.*

MADEIRA, S. C. and OLIVEIRA, A. L. (2004): Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1(1),24–45.*

VAN MECHELEN, I. and SCHEPERS, J. (2006): A unifying model for biclustering. In: *Compstat 2006 - Proceedings in Computational Statistics*, 81–88.

MURALI, T. and KASIF, S. (2003): Extracting conserved gene expression motifs from gene expression. In: *Pacific Symposium on Biocomputing*, 8,77–88.

PRELIC, A., BLEULER, S., ZIMMERMANN, P., WIL, A., BÜHLMANN, P., GRUISSEM, W., HENNING, L., THIELE, L., and ZITZLER, E. (2006): A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics, 22(9),1122–1129.*

SANTAMARIA, R., THERON, R., and QUINTALES, L. (2007): A framework to analyze biclustering results on microarray experiments. In: *8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07)* ,Springer, Berlin, 770–779.

TURNER, H., BAILEY, T., and KRZANOWSKI, W. (2005): Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics and Data Analysis, 48,235–254.*

SURVIVAL MODELS BUILT FROM GENE EXPRESSION DATA USING GENE GROUPS AS COVARIATES

Kai Kammers, Jörg Rahnenführer

Fakultät Statistik, Technische Universität Dortmund,
44221 Dortmund, Germany

Email: kammers@statistik.uni-dortmund.de

Abstract. We present prediction models for survival times built from high dimensional gene expression data. The challenge is to construct models that are complex enough to have high prediction accuracy but that at the same time are simple enough to allow biological interpretation.

Typical univariate approaches use single genes as covariates in survival time models, multivariate models perform dimension reduction through gene selection. Analysis of time-dependent ROC curves and the area under the curves (AUC) can be used to assess the predictive performance (Gui and Li, 2005).

We present models with higher interpretability by combining genes to gene groups (biological processes or molecular functions) and then using these groups as covariates in the survival models. The hierarchically ordered ”GO groups” (Gene Ontology) are particularly suitable. Cox models are used for detecting covariates that are significantly correlated with survival times. Based on these models statistical shrinkage procedures like Lasso-Regression are applied for variable selection. We make use of the R package penalized (Goeman, 2008) that provides algorithms for penalized estimation in generalized linear models, including linear regression, logistic regression and the Cox proportional hazards model.

Our aim is the combination of methods for survival prediction with biological a priori knowledge. First, we compare the prediction performance of models using single genes as covariates with models using gene groups as covariates on several real gene expression datasets. First results indicate that models built with gene groups alone have decreased prediction accuracy since many genes are not yet annotated to their corresponding functions. However, adding gene groups as covariates to models built from single genes improves interpretability while prediction performance remains stable.

In a next step, we integrate GO graph structure in the models (Alexa, Rahnenführer and Lengauer, 2006) in order to cope with the high correlations between neighboring GO groups.

References

Gui, J. and Li, H. (2005): Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21(13), 3001–3008.

Goeman, J.J. (2008): An efficient algorithm for L1 penalized estimation, submitted.

GO Consortium (2004): The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32:D258-D261. Oxford University Press.

Alexa, A., Rahnenführer, J. and Lengauer, T. (2006): Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22(13), 1600–1607.

Keywords

Microarray, Survival Analysis, Gene Ontology

Agreement analysis method in case of continuous variable

Kulwant Singh Kapoor

In clinical and epidemiological studies research are very much interested to know the inter - observer variation in a continuous variable or two measurement techniques.

Example. Measurement of blood pressure with pulse oximetry and ausculatory method or measurement of PEFR respiratory diseases by wright peak flow meter and mini wright meter in other case pulse rate of patient measure by two nurse or doctor.

The conventional Statistical method applied for studying the agreement between two method of measuring a continuous variable is computing the Correlation Coefficient (r), but many times this is misleading for this purpose. A change of scale of measurement does not alter r but affect the agreement . In order to overcome this difficulty we will apply five test and in case three will come out to be true we can say that there is good agreement exist between two rater or techniques

1. r – should be very high [$r > .80$]
2. r'' – should be very low [$r'' < .20$]
3. ICC – should be very high [$ICC > .80$]
4. b – should not be different from 1.
5. d – Bias should not be different from zero and limit of agreement and their 95% C.I. should be within acceptable range.

Using R as enterprise-wide data analysis platform

Zivan Karaman

In this paper we consider the suitability of R to serve as a core tool for enterprise-wide data analysis platform that would be used by both expert and occasional users.

Different requirements for such a tool are examined, including:

- Scope of built-in data analysis functions
- Graphics (including interactive plots)
- Extendibility
- Development environment
- Deployment facilities
- Database and files system connectivity
- Integration with other software
- User interface
- Web deployment capabilities

Overall, R provides an excellent platform for delivering data analytical functions enterprise-wide, including some quite unique features: the broad spectrum of statistical methods that are included, highly flexible graphics, ease of extending existing code with algorithms developed in both R/S language and in other languages (Fortran or C), great database and file system connectivity and nice built-in facilities for package updates.

However, we have also identified some aspects where significant improvements could be done. These include standard, multi-platform IDE (Integrated Development Environment), at least some form of Graphical User Interface for standard data analysis procedures (for mid-level users - expert user can use command-line interface, low-level users need completely packaged applications) to be part of R core system, and some enhanced features for Web-based deployment.

✉ Zivan Karaman
Biostatistics Unit
Limagrain Verneuil Holding, Research Department
B.P. 173, 63204 Riom Cedex, France.
zivan.karaman@limagrain.com

Design and analysis of follow-up studies with genetic component

Juha Karvanen
National Public Health Institute, Helsinki, Finland

In gene-disease association studies, the cost of genotyping makes it economical to use a two-stage design where only a subset of the cohort is genotyped. At the first-stage, the follow-up data along with some risk factors or non-genetic covariates are collected for the cohort and a subset of the cohort is then selected for genotyping at the second-stage. The case-cohort design and the nested case-control design are examples of two-stage designs that are commonly used in epidemiological follow-up studies. The data from a two-stage study can be analyzed as a missing data problem where the genotype data are missing by design for the majority of the cohort. The parameters of the data model, typically logistic model or proportional hazards model, can be estimated by maximizing the full likelihood of the data, which in general case becomes an integral over the missing observations. When dealing with single nucleotide polymorphism (SNP) data, the integrals are replaced by sums over the possible genotypes. As a consequence, the likelihood can be directly maximized by numerical optimization, e.g. by R function `optim`.

The straightforward implementation of full likelihood analysis makes it possible to consider alternative designs for the second stage. One such alternative is the extreme selection where cases and non-cases are selected for genotyping starting from those with largest and smallest covariate values. Another alternative is the D-optimal design, which maximizes the determinant of the Fisher information matrix of the parameters. The determination of the D-optimal design requires the use of heuristic algorithms, which is illustrated in Figure 1.

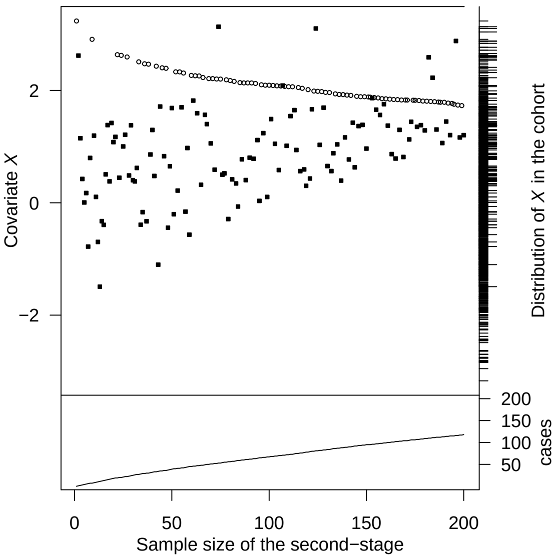


Figure 1: Sequential selection of observations to be included in the second-stage when the response is time-to-event. In the upper panel, sample size on the x-axis indicates the order in which the observations are included. Non-cases are marked by circles and cases are marked by squares. The y-axis on the left presents the covariate values of the selected observations. The tick-marks on the y-axis on the right present the distribution of the covariate values in the whole cohort. The longer tick-marks correspond to cases and the shorter tick-marks correspond to non-cases. In the lower panel, the number of cases selected is shown as a function of the second-stage sample size.

References

J. Karvanen, S. Kulathinal, D. Gasbarra (2008). Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Computational Statistics & Data Analysis*, doi:10.1016/j.csda.2008.02.010.

VARIABLE SELECTION IN REGRESSION USING- R

D. N. Kashid
Department of Statistics,
Shivaji University, Kolhapur
Dnkashid_in@yahoo.com

Abstract:

Variable selection problem is one of the important problems in regression analysis. Over the years, several variable selection methods are proposed in the literature and some frequently used methods are Mallow's Cp-statistic, Forward and Backward, Stepwise selection method etc. All these methods assume that the error distribution is normal and present software packages offer some of these methods for variable selection in regression. It is well known that in the absence of normality or absence of linearity assumption or outlier(s) presence in the data, the classical subset selection methods perform poorly. Such situations demand alternative approaches.

In the last decade, a few methods are developed in the literature based on different situation mentioned above. Ronchetti and Staudte (1994) have proposed robust version of Mallow's Cp called RCp for outlier data. Kashid and Kulkarni (2002, 2003) suggested variable selection techniques to deal the situation mentioned above.

Since these methods are computationally intensive, so it is difficult to select a set of variables without using the software. The implementation of these methods is possible by using R-software. In this article, we exploit use of R in variable selection problem in regression.

References:

Kashid and Kulkarni (2002), A More General Criterion for Subset Selection in Multiple Linear Regression. Communication in Statistics-Theory & Method, 31(5), 795-811.

Mallow's (1973), Some Comments on Cp. Technometrics, 15, 661-665.

Ronchetti and Statutdte (1994), A Robust version of Mallows Cp. JASA, 89(246), 550-559.

Specification of Landmarks and Forecasting Water Temperatur

Göran Kauermann
University of Bielefeld

Thomas Mestekemper
University of Bielefeld

25th March 2008

Abstract

We present and analyse a data set containg water and air temperature in the river Wupper in the northern part of Germany. The analysis pursues two concrete aspects. First, it is of interest to find so called landmarks, these are regularly occuring timepoints at which the temperature follows particular pattern. These landmarks will be used to assess whether the current year is running ahead or behind the "average" seasonal course of a year. Secondly, we focus on forecasting water temperature using smooth principal components. The latter approach is also used for bootstrapping temperatur data, which allows to assess the variability of the specified landmarks.

The implications of our modelling exercise are purely economic. The data trace from a larger project which aims to develop a temperature management tool for two power plants along the river Wupper. These use river water for cooling purposes and to preserve natural wild life in the river there is a strict limit of the maximal temperature of the water. The latter constraints the possible production range of the power plant. More accurate forecasts therefore mean a higher potential of energy production.

Toward Fully Bayesian Computing: Manipulating and Summarizing Posterior Simulations Using Random Variable Objects

Jouni Kerman
Statistical Methodology Group
Novartis Pharma AG
Basel, Switzerland
jouni.kerman@novartis.com

Andrew Gelman
Department of Statistics
Department of Political Science
Columbia University
New York, USA
gelman@stat.columbia.edu

March 28, 2008

Abstract

Bayesian data analysis involves Bayesian inference (model fitting), but also requires post-fitting tasks that include summarizing and manipulating inferences numerically and graphically, and doing model-checking tasks and forecasting using predictive inference. Since Bayesian inference is based on computing and summarizing probability distributions, to do Bayesian data analysis efficiently and conveniently, we need a computing environment that enables us to work with random variables as easily as we do with numerical variables.

We propose a computing environment that defines random variables as natural extensions of traditional numerical objects, which can be regarded as random variables with zero variance. Each numeric vector variable in this environment has a hidden dimension of uncertainty, which is represented by a number of simulation draws from the joint distribution of its components. The random variables can be manipulated transparently, in the same fashion as we do numeric vectors and arrays.

We present an R package, 'rv', that implements this new computing paradigm in R by introducing a new simulation-based random variable class, along with numerous mathematical, statistical, and graphical functions. By converting posterior simulations into random variable objects, they can be manipulated and summarized intuitively and efficiently. We illustrate this by several practical examples.

The Dataverse Network

Gary King

We introduce the Dataverse Network project for data archiving, distribution, and statistical analysis. Via web application software, data citation standards, and integration with R, the Dataverse Network project increases scholarly recognition and distributed control for authors, journals, archives, teachers, and others who organize, produce, or analyze data; facilitates data access and analysis; and ensures long-term preservation whether or not the data are in the public domain. With a few minutes work, you can put a "dataverse" (a full service virtual archive and data analysis engine with your view of the universe of data) on your web page, branded completely as yours, without any local installations or need for maintenance or backups. In addition, any R statistical package can be automatically included in the Dataverse statistical analysis GUI by writing a few simple bridge functions (for the R package Zelig) that describes your package and methods. See the project homepage at `http://TheData.org/`.

Generalized count data regression in R

Christian Kleiber and Achim Zeileis

Fitting functions for the basic Poisson and negative binomial regression models have long been available in base R and in the well-known **MASS** package, respectively. More recently, a number of modified or generalized regression models for count data have become available. Specifically, there now exist functions for fitting hurdle and zero-inflation models in package **pscl** (see Zeileis, Kleiber and Jackman, forthcoming), for fitting Poisson-inverse Gaussian mixtures in package **gamlss** (Stasinopoulos and Rigby, 2007) and for fitting finite Poisson mixtures in package **flexmix** (Leisch, 2004).
The talk will present an overview of the available methods along with empirical illustrations. It will also present functions for some further generalized count data models of recent interest that are not yet publicly available, and suggest directions for future work.

References:

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8).

Stasinopoulos, D.M., and Rigby, R.A. (2007). Generalized additive models for location, scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7).

Zeileis, A., Kleiber, C., and Jackman, S. (forthcoming). Regression models for count data in R. *Journal of Statistical Software*.

Presenter:

Christian Kleiber
Dept. of Statistics and Econometrics
Universität Basel
CH-4051 Basel, Switzerland
christian.kleiber@unibas.ch

Using R as an environment for automatic extraction of forest growth parameters from terrestrial laser scanning data

Hans-Joachim Klemmt[a]

[a] *Chair of Forest Growth and Yield, Technische Universität München, Am Hochanger 13, 85354 Freising*
(h-j.klemmt@lrz.tum.de)

Keywords: Terrestrial laser scanning, TLS, forest inventory, R, software, framework

Laser scanning becomes a more and more important measurement technology in forests. Meanwhile the applicability of airborne laser scanning systems (ALS) for forestry measurement purposes is far advanced [3, 7]. So far ALS-systems mainly concentrate on the extraction of tree height parameters. To describe the structure of forests additional parameters are needed. Terrestrial laser scanning provides very quick information on the structure of forests in form of 3D-point clouds, which are processed to gain such taxation features as the number of trees in a stand, geoposition of individual trunks, diameters at breast height (DBH), crown base height and height of trees [2, 9].

So far unfortunately no software solution exists which extracts the requested parameters automatically from terrestrial 3D data. To eradicate this flaw at the Chair of Forest Growth and Yield at Technische Universität München a working group has been installed which is concerned with this topic. This working group uses R [8] to extract the relevant parameters from 3D-data. R is used because it is a GPL-licensed, Open Source solution for statistical computing which is well-resourced with various packages for clustering-purposes as well as for image processing and visualization. One big advantage of R is also the connectivity with other software like WEKA [5].

To this day a system is developed which separates automatically data sets, which belong to the ground or soil layer, from potential vegetation points. Trees are detected in vegetation point cloud by application of several cluster algorithms. Forest parameters like DBH are calculated by application of Hough-Transformation. Visualization in R is done by the use of standard output as well as by the use of the OpenGL-extension in the package RGL [1].

Although R is an interpreted computer language, which seems to be a big disadvantage for this aim because of the huge number of data sets to process [6], the promising results of the development have shown that it is possible to extract automatically forest growth parameters with a high accuracy and a high level of accordance to manual measurements by the use of this language. Further work aims on the development of a R-based software framework in combination with a JAVA visualization component for the automatic extraction of forest growth parameters from terrestrial laser scanning data.

REFERENCES

[1] Adler, D., Murdoch, D., 2008: RGL (http://cran.r-project.org/web/packages/rgl/index.html)

[2] Bienert, A., Scheller, S., Keane, E., Mullooly, G., Mohan, F., 2006: Application of terrestrial laser scanners for the determination of forest inventory parameters. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 36, Part 5.

[3] Heurich, M., Persson, A., Holmgren, J., Kennel, E., 2004: Detecting and measuring individual trees with laser scanning in mixed mountain forest of Central Europe using an algorithm developed for Swedish boreal forest conditions. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVI-8/W2, p. 307-312.

[4] Holmgren, J. 2003. Estimation of Forest Variables using Airborne Laser Scanning. Doctoral thesis, Swedish University of Agricultural Sciences, Umea.

[5] Hornik, K., 2008: The RWekaPackage. (http://cran.r-project.org/web/packages/RWeka/index.html)

[6] Ligges, U., Programmieren mit R, Springer, 2. ed., 247 pp.

[7] Reitberger, J., Krzystek, P. Heurich, M., 2006: Full-waveform analysis of small footprint airborne laser scanning data in the Bavarian Forest National Park for tree species classification, Proceedings of workshop on 3D remote sensing in forestry, 14th-15th February 2006, Vienna, pp. 1-10

[8] R Development Core Team, 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[9] Wezyk, R., Koziol, K., Glista, M., Pierzchalski, M., 2007: Terrestrial laser scanning versus traditional forest inventory. First results from the Polish forests. ISPRS workshop on Laserscanning 2007 and SiliviLaser 2007, Espoo, September 12-14, 2007, Finland, p. 424-429.

Rfit: An R Package for Rank Estimates

John Kloke

In the nineteen seventies, Jureckova and Jaeckel proposed rank estimation for linear models. Since that time, several authors have developed inference and diagnostic methods for these estimators. These estimators and their associated inference are robust to outliers in response space. The methods include estimation of standard errors, tests of general linear hypotheses, confidence intervals, studentized residuals, and measures of influential cases. Unfortunately, these methods are not implemented in main stream software, and hence are not widely used. For this presentation I will highlight the main features of an R package I am developing which implements these methods. The package uses standard linear models syntax and includes many of the main inference and diagnostics functions (e.g. `anova`, `summary`, `rstudent`, `influence.measures`).

sfCluster/snowfall: Managing parallel execution of R programs on a compute cluster

Jochen Knaus, Institute of Medical Biometry and Medical Informatics University Medical Center Freiburg, Germany

Modern bioinformatics applications require a huge amount of computing resources. To adress these, techniques such as MPI, available through the R packages `Rmpi` and `snow`, allow the bundling of single machines into compute clusters. However, management of cluster resources has to be performed manually, resulting in problems, when several, potentially unexperienced users access the same cluster pool.

As a solution for this problem we developed *sfCluster* and the `snowfall` R package based on the `snow` package and LAM/MPI. Both are designed for easy and safe usage, hiding cluster setup and internals from end users, who only see a clean `snow`-like API.

sfCluster is a Unix tool for management of parallel R programs, which assigns resources dynamically in a reasonable way, sets up the LAM cluster and monitors the execution of the parallel R program as well as controlling the cluster itself.

sfCluster features various execution modes: it can run an R interactive shell, raw batch mode or a visual monitoring mode, which allows process and logfile control during runtime directly on the terminal using Curses. Memory observation, process control and cluster session shutdowns even work if the LAM cluster itself died or some machines went offline or network problems occurred.

`snowfall` is the corresponding R package, which connects to *sfCluster*, but can also be used without it. In contrast to the `snow` package it provides easy switching between sequential and parallel execution, which eases development on machines without cluster environment. The package also features basic intermediate saving of results (with restore), so not all results are lost in case of a cluster stop.

The use of these advanced tools will be illustrated with application scenarios from our department, where several users can now perform demanding bioinformatics simulation studies at the same time.

mboost - Componentwise Boosting for Generalised Regression Models

Thomas Kneib & Torsten Hothorn
Department of Statistics
Ludwig-Maximilians-University Munich

In recent years, boosting has emerged into a widely applied technique for fitting various types of generalised regression models. The main reason for its popularity is that it is surprisingly simple in requiring only iterative fitting of some (potentially simple) base-learning procedure such as (penalised) least-squares to working residuals. Moreover, boosting allows to define various types of regression situations by formulating them in terms of a suitable loss function. From a theoretical perspective, boosting then equals a functional gradient descent algorithm for solving the empirical risk minimisation problem and the working residuals are given by the negative gradient of the loss function.

While boosting has mainly been used to fit completely nonparametric black box models in a prediction-oriented framework first, recent research has shown that it can actually be used to estimate structured regression models. Therefore the base-learning procedure is separated into several components and only the best-fitting component is updated in each iteration. For example, when fitting a generalised linear model, each base-learner might correspond to a single covariate and only the effect of the best-fitting covariate is updated in each boosting iteration. Applying a suitable stopping rule to the boosting iterations yields an adaptively regularised model fit that also provides a means of variable selection and model choice.

The package **mboost** provides implementations for the most common types of univariate exponential family responses where the negative log-likelihood provides the loss-function, but also for other types of regression situations such as robust regression based on Huber‘s loss function. Further extensions, for example to survival modelling are currently being investigated.

A wide range of componentwise base-learning procedures is available based on (penalised) least squares fits, for example

- parametric linear effects as in generalised linear models,
- penalised splines for nonparametric effects and varying coefficient terms,
- penalised tensor product splines for interaction surfaces and spatial effects,
- ridge regression for random intercepts and slopes,
- stumps for piecewise constant functions.

Through its modular formulation, boosting allows to define models consisting of arbitrary combinations of these effects and we will illustrate the versatility of the resulting model class in a spatio-temporal regression model for the analysis of forest health.

Paper Proposal / Rüya Gökhan Koçer- University of Amsterdam (r.g.kocer@uva.nl)

Believing by Seeing before Seeing by Believing: Visualizing the Gaussian Regression Model by the SIM.REG package for intuitive teaching

The Gaussian Regression Model is one of the fundamental techniques in econometric theory. There are several crucial assumptions of this model such as homoskedastic error variance, no-autocorrelation between error distributions, no high multicolinearity between independent variables, and identification of correct functional form. When the model is used for hypothesis testing the normality of error distributions too should be added to the list. These fundamental assumptions are of course always mentioned in the regression courses and several statistical tests are introduced to diagnose possible violations. Unfortunately, students, especially those with social science background, quite often fail to internalize the importance of these assumptions neither do they really appreciate the BLUE (Best Linear Unbiased Estimator) property of The Gaussian Regression Model. However, it is of crucial importance for the students to develop an intuitive understanding of the weaknesses and strengths of this basic approach in order to be able comprehend the premises of and need for more sophisticated modeling techniques. In other words students first need to 'believe' by 'seeing' to be able to 'see' more advanced theorems and models by 'believing'.

In this paper, a package programmed by the author in R language to conduct Monte Carlo simulations, the SIM.REG (Simulated Regression) is introduced and its visual and analytical strength in depicting ins and outs of The Gaussian Regression Model is demonstrated.

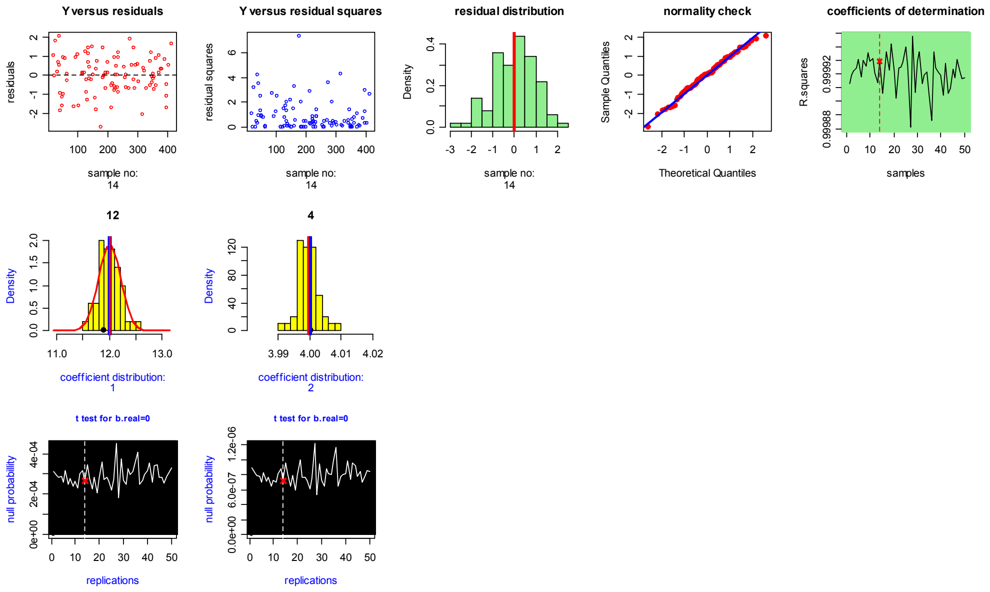
The SIM.REG allows creating, testing and visualizing data sets which contain desired degrees of heteroskedasticity, autocorrelation, multicolinearity and non-normality in order to reveal the isolated or simultaneous impact of these violations on the Gaussian regression model. Similarly, the SIM.REG also allows revealing the impact of wrong functional forms on the model. Indeed SIM.REG contains an option which allows seeing, for example, the impact of increasing level of autocorrelation on inference structure as an animation. Moreover under the SIM.REG one can also visually depict the BLUE property of the Gaussian Regression Model by making all assumptions hold. In this way students can be visually shown the degree to which violation of assumptions damage the capacity of the model to make reliable estimations and correct inference.

Moreover, it is also possible to use the SIM.REG to generate data sets which allow testing the sensitivity of statistical tests and indicators, such as rank correlation test, Durbin-Watson test, condition index etc... In other words by using the SIM.REG one can create data sets which contain an isolated problem such as heteroskedasticity and then scrutinize the ability of various statistical tests to identify the problem. In this way one can enable the students to develop a critical approach to various tests, that is, when and why to disregard the positive or negative test outcomes about particular problems. More importantly one can also create several problems simultaneously in order to reveal how these simultaneous violations may render some instruments of diagnosis ineffective.

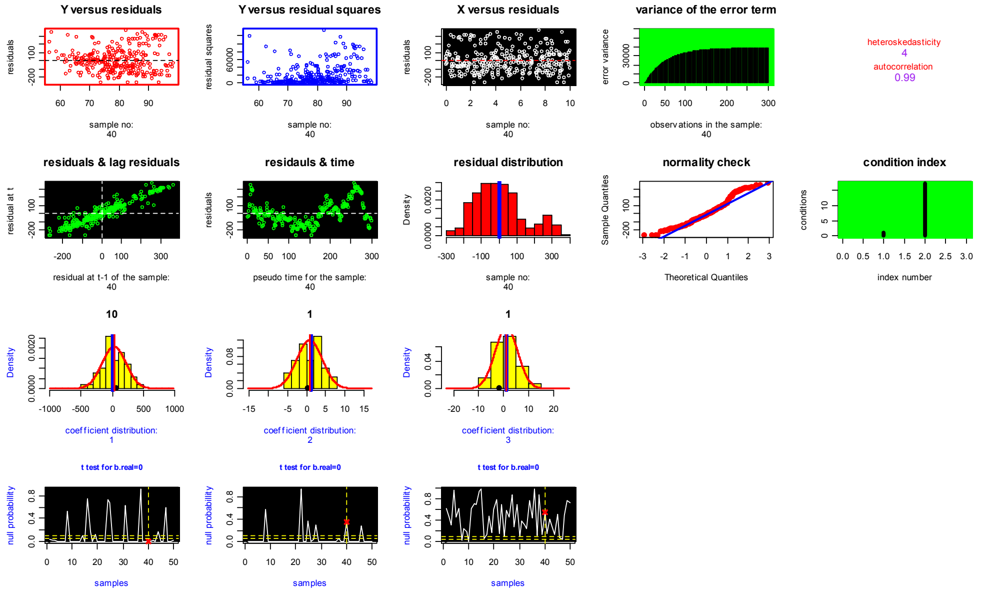
The main purpose of the paper is, by sketching out various applications of the SIM.REG package, to show how it can be effectively used for teaching purposes in MA level regression courses.

Paper Proposal / Rüya Gökhan Koçer- University of Amsterdam (r.g.kocer@uva.nl)

Example 1: the SIM.REG output for simple linear regression with full compliance after 50 simulations



Example 2: the SIM.REG output for multiple linear regression with heteroskedasticity, severe autocorrelation and mild multicolinearity after 40 simulations



R-Packages for Robust Asymptotic Statistics

M. Kohl and P. Ruckdeschel[1]

[1] Mathematik VII: Stochastics, University of Bayreuth, D-95440 Bayreuth, Germany

We present a family of R-packages designed for a conceptual adaptation of an asymptotic theory of robustness.

Package RobAStBase provides the basic S4 classes and methods for optimally robust estimation in the sense of Rieder (1994). That is, we consider L_2 differentiable parametric models in the framework of infinitesimal (shrinking at a rate of $\sqrt{n}$) neighborhoods. The combination of RobAStBase with our R packages distr, distrEx and RandVar enables us to implement **one** algorithm which works for a whole class of various models, thus avoiding redundancy and simplifying maintenance of the algorithm.

Package ROptEst so far covers the computation of optimally robust influence curves for all(!) L_2 differentiable parametric families which are based on a univariate distribution. With the Kolmogorov and the Cramér von Mises minimum distance estimators which are implemented in our R package distrMod and which serve as starting estimators, we are able to provide optimally robust estimators by means of k-step constructions ($k \geq 1$).

Package RobLox includes functions for the determination of influence curves for several classes of robust estimators in case of normal location with unknown scale; cf. Kohl (2005). In particular, the function roblox, computes the optimally robust estimator for normal location and scale as described in Kohl (2005). In contrast to package ROptEst, in which we aim for generality, the function roblox is optimized for speed.

Package ROptRegTS contains the extension of the asymptotic theory of robustness to regression-type models like the linear model and certain time series models (e.g., ARMA and ARCH).

Finally, package RobRex provides functions for the determination of optimally robust influence curves in case of linear regression with unknown scale and standard normal errors where the regressor is random. Analogously to package RobLox the functions in package RobRex are optimized for speed.

References

M. Kohl (2005). *Numerical contributions to the asymptotic theory of robustness*. Dissertation, Universität Bayreuth, Bayreuth.

R Development Core Team (2008). R: A language and environment for statistical computing. http://www.r-project.org.

H. Rieder (1994). *Robust asymptotic statistics*. Springer.

Customer Heterogeneity in Purchasing Habit of Variety Seeking Based on Hierarchical Bayesian Model

Fumiyo Kondo Teppei Kuroda

This research presents a model which expresses product choice behavior in terms of 'inertia' or 'variety seeking' for each customer by using a mixture normal-multinomial logit model in a hierarchical Bayesian framework.

A product choice behavior is called as 'inertia' if a customer chooses the same product as the previously purchased and 'variety seeking' if it is a different product from the previous one. These kinds of behaviors are frequently observed in the product category of 'low involvement' (Dick and Basu (1994), Peter and Olson (1999)). Consumers tend to purchase a 'low involvement' product such as beverage or cake based solely on experience, inertia, or atmosphere. In addition to 'inertia' or 'variety seeking', Bawa (1990) proposed a model for segmentation purposes. It has an additional segment of 'hybrid' customer, of which purchasing tendency changes from 'inertia' to 'variety seeking' or vice versa. Moreover, it is getting increasingly important to understand the heterogeneity of customers in recent years, particularly from the view point of the category attribute.

A comparison was made between a hierarchical Bayesian model and a finite mixture model on the product category of Japanese tea and Chinese tea. The result shows that the hierarchical Bayesian model is superior to the finite mixture model in terms of 'hit rate'. Further, the model with the variables of 'inertia' or 'variety seeking' was superior to the one without them in terms of Deviance Information Criterion, DIC. In addition, we extended the Bawa's formula on 'inertia', 'variety seeking' or 'hybrid' behavior by considering the influence of purchasing intervals. Our proposed model that considers a timing of customer's brand switching was superior to the Bawa's formula. We obtained the results that each customer has a tendency of 'inertia' or 'variety seeking' or 'hybrid' in product choice, which is different between the category of Japanese tea and that of Chinese tea.

Finally, we proposed a CRM related strategy that calculates a necessary discount rate for individual brand switching and offering the brand according to the brand switching timing of each consumer.

Profiling the parameters of models with linear predictors

by Ioannis Kosmidis
Department of Statistics, University of Warwick
Coventry, CV4 7AL, UK

March 27, 2008

Profiles of the likelihood can be used for the construction of confidence intervals for parameters, as well as to assess features of the likelihood surface such as local maxima, asymptotes, etc., which can affect the performance of asymptotic procedures. The `profile` methods of the **R** language (**stats** and **MASS** packages) can be used for profiling the likelihood function for several classes of fitted objects, such as `glm` and `polr`. However, the methods are limited to cases where the profiles are almost quadratic in shape and can fail, for example, in cases where the profiles have an asymptote.

Furthermore, often the likelihood is replaced by an alternative objective for either the improvement of the properties of the estimator, or for computational efficiency when the likelihood has a complicated form (see, for example, Firth (1993) for a maximum penalized likelihood approach to bias-reduction, and Lindsay (1988) for composite likelihood methods, respectively). Alternatively, estimation might be performed by using a set of estimating equations which do not necessarily correspond to a unique objective to be optimized, as in quasi-likelihood estimation (Wedderburn, 1974; McCullagh, 1983) and in generalized estimating equations for models for clustered and longitudinal data (Liang & Zeger, 1986). In all of the above cases, the construction of confidence intervals can be done using the profiles of appropriate objective functions in the same way as the likelihood profiles. For example, in the case of bias-reduction in logistic regression via maximum penalized likelihood, Heinze & Schemper (2002) suggest to use the profiles of the penalized likelihood, and when estimation is via a set of estimating equations Lindsay & Qu (2003) suggest the use of profiles of appropriate quadratic score functions.

In this presentation we introduce the **profileModel** package, which generalizes the capabilities of the current `profile` methods to arbitrary, user-specified objectives and, also, covers a variety of current and potentially future implementations of fitting procedures that relate to models with *linear predictors*. We give examples of how the package can be used to calculate, evaluate and plot the profiles of the objectives, as well as to construct profile-based confidence intervals. The presentation focuses on the following:

- **Generality of application:** The **profileModel** package has been designed to support classes of fitted objects with linear predictors that are constructed according to the specifications given by Chambers & Hastie (1991, Chapter 2). Such generality of application stems from the appropriate use of generic **R** methods such as `model.frame`, `model.matrix` and `formula`.

- **Embedding:** The developers of current and new fitting procedures as well as the end-users can have direct access to profiling capabilities. The only requirement is authoring a simple function that calculates the value of the appropriate — for their specific application — objective to be profiled.

- **Computational stability:** All the facilities have been developed with computational stability in mind, in order to provide an alternative which improves and extends the capabilities of already available `profile` methods.

References

Chambers, J. M. & Hastie, T. (1991). *Statistical Models in S.* Chapman & Hall.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Heinze, G. & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lindsay, B. G. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes*, Ed. N. U. Prabhu, pp. 221–239. American Mathematical Society.

Lindsay, B. G. & Qu, A. (2003). Inference functions and quadratic score tests. *Statistical Science* **18**, 394–410.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics* **11**, 59–67.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Graphical Functions for Prior Selection

Stephanie Kovalchik
University of California, Los Angeles
MS Biostatistics, BS Biology
Department of Biostatistics
UCLA School of Public Health Box 951772
Los Angeles, CA 90095-1772
skoval@ucla.edu

March 11, 2008

In a Bayesian analysis the choice of prior distributions for model parameters reflects the analyst's *a priori* belief. Most discussion and presentation of prior densities are in terms of the model parameter values using the BUGS squiggle notation. With the exception of the Uniform and Normal distributions, using this notation alone can make it difficult to immediately assess the beliefs represented by the priors. Thus, the squiggle notation presentation creates interpretational difficulty in that it does not reflect the process by which analysts will choose priors. In most cases, particularly when experts outside of the statistical field are asked to give information to elicit priors, the construction will be in terms of the moments, mode and/or coverage probabilities of the parameters. It would be useful then to have a set of functions that will take these quantities as arguments and translate them into the corresponding prior density, returning the parameter values and providing a density plot.

This presentation will demonstrate a set of graphical functions written in R which allow the user flexibility in specifying the desired moments, mode or coverage probabilities when deciding on the appropriate prior. Examples from the literature are given showing how these functions can facilitate prior determination when eliciting priors from experts as well as reveal misspecification of *a priori* beliefs. The graphical functions are based on the base graphics system which enables the user to easily annotate and customize the display. The tools are available for commonly used densities of the stats package including the Normal, Student's t, Beta and Gamma. Current work is being done to expand these plotting functions so as to allow the specification of mixture priors. It is the goal of this work to provide prior selection tools in the R language comparable to those of Tony O'Hagan's First Bayes. With these simple extensions of R's standard statistical and graphical facilities Bayesian statisticians working in R will be able to more efficiently select and present prior distributions.

The BayHaz package for Bayesian estimation of smooth hazard rates in R

Luca La Rocca
University of Modena and Reggio Emilia, Italy
luca.larocca@unimore.it

March 29, 2008

Abstract

Package BayHaz (La Rocca, 2007) for R (R Development Core Team, 2008) consists of a suite of functions for Bayesian estimation of smooth hazard rates using compound Poisson process priors, introduced by La Rocca (in press), and first order autoregressive Bayesian penalized spline priors, based on Hennerfeind *et al.* (2006). Prior elicitation, posterior computation, and visualization are dealt with. For illustrative purposes, a data set in the field of earthquake statistics is supplied. An interface to package coda (Plummer *et al.*, 2007) facilitates output diagnostics. Future plans are to implement other Bayesian methods for hazard rate estimation, and to make available an extension to the proportional hazards model.

References

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive survival models. *J. Amer. Statist. Assoc.* **101**, 1065–1075.

La Rocca, L. (2007). *BayHaz: R Functions for Bayesian Hazard Rate Estimation.* R package version 0.1-3. URL http://www-dimat.unipv.it/luca/bayhaz.htm.

La Rocca (in press). Bayesian Non-parametric Estimation of Smooth Hazard Rates for Seismic Hazard Assessment. *Scand. J. Statist.* to appear.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2007). *coda: Output analysis and diagnostics for MCMC.* R package version 0.13-1.

R Development Core Team (2008). *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.

New possibilities for interactive specification and validation of models for Fluorescence Lifetime Imaging Microscopy (FLIM) data with the TIMP package.

Laptenok P. Sergey[1,3], Katharine M. Mullen[2], Jan Willem Borst[1], Herbert van Amerongen[1,3], Antonie J. Visser[1]

[1] MicroSpectroscopy Centre, Wageningen University and Research Center, The Netherlands
[2] Department of Physics and Astronomy, Faculty of Sciences Vrije Universiteit Amsterdam, The Netherlands
[3] Laboratory of Biophysics, Wageningen University and Research Center, The Netherlands

The detection of protein-protein interactions in a biological cell is required to enhance our knowledge about mechanisms that regulate intracellular processes. Förster Resonance Energy Transfer (FRET) between donor and acceptor molecules is a widely used technique to monitor protein-protein interactions. As FRET is a fluorescence quenching process, it can be detected by the shortening of the fluorescence lifetime of the donor molecule. Fluorescence Lifetime Imaging Microscopy (FLIM) allows the mapping of fluorescence lifetimes with (sub-) nanosecond time resolution and a spatial resolution of 250 nm. FRET phenomena measured with the FLIM technique provides temporal and spatial information about molecular interaction in living cells.

For accurate and quantitative FLIM data analysis well-designed analysis protocols are required. The dynamical features of a FRET system are often well described by a small number of kinetic processes, in which the associated fluorescence lifetimes in all pixels have similar values, but the relative amplitudes may vary from pixel to pixel. In this case significant advantages and accuracy in analysis can be achieved by global analysis of the image. Global analysis uses fluorescence decay traces from all pixels to estimate both kinetic parameters (lifetimes) and relative amplitudes of components in each pixel. The TIMP package has been shown to be effective at performing global analysis of FLIM images [1].

A typical FLIM image represents in the order of 10^3 pixels with 10^3 time points per pixel. Presentation of the data analysis results needs to be well-organized and interactive, allowing the user to obtain a detailed graphical presentation of the fit at any pixel selected. Here we present new options for the TIMP package allowing interactive presentation of global analysis results, as well as importing and preprocessing FLIM data. The analysis of FLIM images of transcription factors fused with either cyan fluorescence protein (CFP) or yellow fluorescence protein (YFP) in plant cells will be given as an example. The novel data analysis methodology could reveal molecular interactions among different transcription factors in the nucleus of a plant cell.

References:

[1] Laptenok S, Mullen KM, Borst JW, van Stokkum IHM, Apanasovich VV, Visser AJWG (2007). "Fluorescence Lifetime Imaging Microscopy (FLIM) Data Analysis with TIMP." Journal of Statistical Software, 18 (4). URL http://www.jstatsoft.org/v18/i08/.

ivivc - A Tool for *in vitro-in vivo* Correlation Exploration with *R*

Hsin-ya Lee, Pao-chu Wu, Yung-jin Lee

College of Pharmacy, Kaohsiung Medical University,
Kaohsiung, Taiwan

Introduction *In vitro-in vivo* correlation (IVIVC) is defined as the correlation between *in vitro* drug dissolution and *in vivo* drug absorption. The main purpose of an IVIVC model is to utilize *in vitro* dissolution profiles as a surrogate for *in vivo* bioequivalence and to support biowaivers. In order to prove the validity of a new formulation, which is bioequivalent with a target formulation, a considerable amount of efforts is required to study bioequivalence/bioavailability. Thus, data analysis of IVIVC attracts attention from the pharmaceutical industry. The purpose of this study is to develop an IVIVC tool (*ivivc*) in R. **Methods** Development and validation are 2 critical stages in the evaluation of an IVIVC model. In the first stage, the development of level A IVIVC model is usually estimated by a two-stage process. (1) Deconvolution: the observed fraction of the drug absorbed is based on the Wagner-Nelson method. IV, IR or oral solution was attempted as the reference. Then, the pharmacokinetic parameters will be estimated using a nonlinear regression tool or be attempted from literatures reported previously. The IVIVC model is developed using the observed fraction of the drug absorbed and that of the drug dissolved. Based on the IVIVC model, the predicted fraction of the drug absorbed is calculated from the observed fraction of the drug dissolved. (2) Convolution: the predicted fraction of the drug absorbed is then convolved to the predicted plasma concentrations by using the convolution method. In the second stage, evaluating the predictability of a level A correlation focuses on estimating the percent prediction error (%PE) between the observed and predicted plasma concentration profiles, such as the difference in pharmacokinetic parameters (C_{max}, and the area under the curve from time zero to infinity, AUC_∞). **Results and Discussion** We call this tool as *ivivc*. It can be used to calculate the observed fraction of the drug absorbed in different pH media and formulations with multiple subjects at the same time. Based on the linear regression, the predicted fraction of the drug absorbed is calculated from the observed fraction of the drug dissolved. Furthermore, the percent prediction error (%PE) between the observed and predicted plasma concentration profiles, such as C_{max} and AUC_∞ are also calculated. **Conclusion and Future Work** In this study, we have successfully created the package, *ivivc*. *ivivc* will be released to public soon. In the future, we will include more methods that have been published and frequently used to develop IVIVC.

PKfit - A Pharmacokinetic Data Analysis Tool on *R*

Chun-Ying Lee[1], Yung-Jin Lee[2]

[1]Pharmacy Department, Changhua Christian Hospital, Changhua, Taiwan

[2]Graduate Institute of Clinical Pharmacy, College of Pharmacy, Kaohsiung Medical University, Kaohsiung, Taiwan

Introduction: Pharmacokinetic (PK) data analysis heavily depends on computer calculation power. In this study, we tried to create a nonlinear regressions tool on R using its available packages and functions. **Methods and Materials**: Design goal of this tool was aimed to be easy-to-use, so a menu-driven interface on RGui was developed. We used *lsoda* function (in *odesolve* package) to solve all differential equations used to define PK models. As for data fitting algorithms, Gauss-Newton algorithm (*nls* function in *stats* package) for non-linear regression, and the Nelder-Mead simplex method (*optim* function in *stats* package) for minimization of weighted sum of squares, as well as the genetic algorithm (*genoud* function in *rgenoud* package) were applied. Users just follow the menu step by step, and then will get the job done. Fourteen pharmacokinetic models: intravenous drug administrations with i.v. bolus or i.v. infusion, extravascular drug administrations, linear with 1st-ordered absorption/elimination or nonlinear (Michaelis-Menten models were built. Two weighting schemes, 1/Cp(obs), and 1/Cp^2(obs) were also included. The output information included a summarized table (consisting of time, observed and calculated drug plasma/serum concentrations, weighted residuals, area under plasma concentration curve (AUC), and area under the first moment (AUMC), goodness-of-fit, final PK parameter values, and plots such as linear plots, semi-log plots, and residual plots. In the part of simulation, *runif* and *rnorm* functions from *stats* package provide the generation of random uniform distribution derivates and normal distribution derivates for PK parameters, respectively. Further, we also provide the function of Monte-Carlo Simulation. **Results and Discussion**: We called this tool as *PKFit*. It has been announced publicly, and can be downloaded from mirror sites of CRAN (package name: *pkfit*). With only a few examples, most results obtained from in *PKfit* were comparable to those obtained from other two pharmacokinetic programs, *WinNonlin* and *Boomer*. **Conclusion and Future Work**: *PKfit* running on *R* has been built and has been proved that it can provide efficiency and accuracy in data fitting functions. Multiple dosing models or algorithms may be required for further development of *PKfit*.

Keywords: *R*, Pharmacokinetics, Nonlinear Regression, Data Fitting, Simulation

Acknowledgement: We like to thank Pharsight, USA (www.pharsight.com) to freely provide us *WinNonlin pro* (through Pharsight Academic License program), and also Dr. David Bourne for his Boomer (www.boomer.org) for this study. Also we really appreciate the assistance from Dr. Woodrow Setzer (*odesolve*), Dr. Jasjeet Sekhon (*rgenoud*) and Dr. Anthony Rossini (scripting). Without these nice people, we definitely cannot finish this project.

SMALL GROUPS and QUESTIONNAIRES

L.F. Lemmens
Universiteit Antwerpen Departement Fysica & StatUa
CGBU415
Groenenborgerlaan 171 B 2020 Antwerpen, Belgium
(lucien.lemmens at ua.ac.be)

March 31, 2008

Most administrations want to have surveys on the quality of services provided by their officials. The questionnaire technique is often used. For a number of items one asks the respondents to indicate how strongly they agree or disagree with a given statement. Usually several items form a dimension – a name given to an essential part of the service – and the survey of the dimensions reports a summary of the attitude of the respondents. This summary is used to evaluate the performance of the official and can have consequences for promotion. Because these surveys can have consequences, they are contested when the groups are small: a classical analysis rarely avoids the use of the central limit theorem or the law of large numbers. In bayesian statistics, however, the inverse probability problem is readily solved given the likelihood function of the problem and prior density and the evidence follows as usual from normalization.

Assume that a respondent can take 6 attitudes for an item, the information we want to obtain is then $\{N_i \quad i \in [1, \cdots, 6]\}$ with $\sum_i N_i = N$ where N is the number of possible respondents in the complete group. The information we obtain in the questionnaire is $\{n_i \quad i \in [1, \cdots, 6]\}$ with $\sum_i n_i = n$ where n is the number of respondents for that item. The knowledge about $\{n_i\}$ will be used to guess the $\{N_i\}$ This model has a multivariate hypergeometric density with the N_i as parameters. The prior starts from an educated guess that predicts the $\{N_i\}$ without using results from the questionnaire on that item. The most convenient density is a multinomial with given $\{p_i\}$ where the p_i indicate the plausibility that a respondent takes the attitude i. Combining this setting for an item to a model for a dimension we can use the posterior of the first item as a prior for the second item and so on. This leads, for the dimension, to a Dirichlet-model, that belongs to the exponential family. Hence the results are obtained by upgrading, avoiding numerical integration for the calculation of the evidence.

Although the statistical analysis is computationally simple, there are a lot of surveys to be analyzed and communicated to decision makers. A relatively simple R–code was written to automize the analysis and decision theoretical arguments are used to implement a representation of the uncertainties on the data graphically.

Comparison of spatial interpolation methods using a simulation experiment based on Australian seabed sediment data

Jin Li* and Andrew D. Heap
Marine & Coastal Environment
Petroleum and Marine Division
Geoscience Australia
GPO Box 378
Canberra ACT 2601

* Corresponding author:
Jin Li
GPO Box 378
Canberra ACT 2601
Australia
Email: jin.li@ga.gov.au

Spatial distribution data of environmental variables are increasingly required as geographic information systems (GIS) and modelling techniques become powerful tools in natural resource management and biological conservation. However, the spatial distribution data are usually not available and the data available are often collected from point sources. This is particularly true of seabed data for the world's oceans, especially the deep ocean. A typical example is Australia's marine region. Here, Geoscience Australia has to derive spatial distribution data of seabed sediment texture and composition for 8.9 million km^2 of Australia's marine region from about 14,000 sparsely and unevenly distributed samples. The need for these data comes from seabed habitat classifications and predictions of marine biodiversity as key information sources supporting ecosystem-based management. Spatial interpolation techniques provide essential tools to generate such spatial distribution data by estimating the values for the unknown locations using the point samples, but they are often data- or even variable- specific. The estimation of a spatial interpolator is usually affected by many factors including the nature of data and sample density. There are no consistent findings about how these factors affect the performance of spatial interpolators. Therefore, it is difficult to select an appropriate interpolator for a given input dataset. In this study, we aim to select appropriate spatial interpolation methods by comparing their respective performance using a simulation experiment based on Australian seabed sediment data in R. Three factors affecting the accuracy and precision of the interpolations are considered: the spatial interpolation method, spatial variation in data, and sample density. Stratification based on geomorphic features is also used to improve estimation. Bathymetry data are considered as secondary information in the experiment. Cross-validation is used to assess the performance of spatial interpolation methods. Results of this experiment provide suggestions and guidelines for improving the spatial interpolations of marine environmental data, which have application for using seabed mapping and habitat characterisations in achieving management and conservation goals.

A Closer Examination of Extreme Value Theory Modeling in Value-at-Risk Estimation

Wei-han Liu
Department of Banking and Finance
Tamkang University
Taipei, Taiwan [*]

Extreme value theory has been widely used for modeling the tails of return distribution. Generalized Pareto distribution (GPD) is popularly acknowledged as one of the major tools in Value-at-Risk (VaR) estimation. As Basel II stipulates the significance level for VaR estimation from previous 5% quantile level to more extremal quantile levels at 1%, it demands a more accurate estimation approach. It is imperative to take a closer examination at GPD modeling performance at more extremal quantile levels. Empirical analysis outcomes show the acknowledged outperformance of GPD is sustained at 5% quantile level but not at 1% level. Alternative methods are introduced and the empirical outcomes indicate that both the penalized spline smoothing in semiparametric regression and maximum entropy density based dependent data bootstrap outperform GPD in modeling extremal quantile levels lower than 1%.

Keywords: Value-at-Risk, Extreme Value Theory, Generalized Pareto Distribution, Semiparametric Regression, Penalized Spline Smoothing, Maximum Entropy Bootstrap

JEL: classification: C13, C14, G32

[*] Correspondence Information: Tamsui Campus, 151 Ying-chuan Road Tamsui, Taipei County Taiwan 25137, Republic of China, Fax: 886-2-26214755, Email: weihanliu2002@yahoo.com

R and Stata for Building Regression Models

Andras Low

Stata system provides many routines for data manipulation and data analysis. Stata has excellent capabilities for developing models on count data. At a specific task a user interface can facilitate the researcher to focus on the topic of his/her research instead of syntax. To create a web browser based user interface for special purpose is easier in R with Tcl/Tk, R2HTML and Rpad. With this interface the user can

- specify different models,
- compare the models with each others,
- diagnose the residuals,
- update the models,
- verify the assumptions derived from models,
- document them.

Surface and Sprinkle Irrigation Analysis with R

Marcio Antonio Vilas Boas: vilasboas@unioeste.br
Miguel Angel Uribe-Opazo : mopazo@unioeste.br
Edson Antonio Alves da Silva: edsonsilva@unioeste.br

Universidade Estadual do Oeste do Paraná- Unioeste – Brasil
Centro de Ciências Exatas e Tecnológicas/Engenharia Agrícola/Cascavel-PR: Rua Universitária
2069-Jardim Universitario- CEP85807460: tel: 45 3220 3155

The application of water to agricultural lands for the purpose of irrigation is one of the alternate uses of this natural resource in many areas. It is essential that water be used effectively and efficiently, whether the supply is limited or excessive. Irrigation efficiency is a concept used extensively in system design and management. It can be divided into two components, uniformity of application and losses. If either uniformity is poor or losses are large, efficiency will be low. This paper will present Irrigation-R for basic irrigation analysis. The functions uniformity and efficiency for irrigation surface e sprinkle cried. The program is not intended for experts and gives direct access only to a very limited set of R-functionality.Strengths and weaknesses of the approsch 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. The visualization is implemented using R under GLIB over Windows environment.

Keywords: uniformity, efficiency, Coefficient, evaluating, overlapping, furrow, field test.

References

Jack K. and Ron D. B. Sprinkle and Trickle irrigation. Van Nostrand Reinhold, 1990, 652p.

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.

Vilas Boas, M. A. (2002). Hidráulica da irrigação por superficie: desenvolvimento computacional do modelo matemático Zero-Inércia , Cascavel: Edunioeste, 2002. 120p.

Walker, W. R., and Skogerboe, G. V. Surface irrigation: theory and practice. Englewood Clifs: Prentice Hall, 1987. 386p.

The statistical evaluation of DNA crime stains in R

Miriam Marušiaková

Department of Statistics, Charles University in Prague
Centre of Biomedical Informatics, Institute of Computer Science AS CR
E-mail: maruskay@gmail.com

Suppose a crime has been committed and a blood stain was collected from the crime scene. It is believed the stain was left by an offender. A suspect is arrested and it is found out that his DNA profile matches the DNA profile of the crime stain. In forensic science, it is common to consider DNA profile match probabilities under the hypothesis that the offender was someone else than the suspect. The problem was investigated, e.g., in [1], under general assumptions allowing for population substructure and relatedness.

In case of DNA mixtures (from more than one person), the weight of the DNA evidence is assigned in terms of likelihood ratio of match probabilities, comparing two hypotheses about origin of the mixture. Authors in [2] obtained a general formula for calculation of match probabilities under assumption of independent alleles in DNA profiles. The result was further extended by [3] and [4] to allow for population substructure and dependence. The DNA mixture problems with presence of relatives were discussed, e.g., in [5].

The aim of this talk is to introduce an R package called *forensic* where the calculations of match probabilities mentioned above are implemented. The functionality of the package will be demonstrated using data from real situations.

The work was supported by GAČR 201/05/H007 and the project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] Balding, D. and Nichols, R. (1994). *Forensic Science International* **64**, 125-140.

[2] Weir, B., Triggs, C., Starling, L., Stowell, L., Walsh, K., and Buckleton, J. (1997). *Journal of Forensic Sciences* **42**, 213-222.

[3] Fung, W. and Hu, Y. (2000). *Journal of the Royal Statistical Society A* **163**, 241-254.

[4] Zoubková, K. and Zvárová, J. (2004). Master's thesis, Charles University, Prague.

[5] Hu, Y. and Fung, W. (2003). *International Journal od Legal Medicine* **117**, 39-45.

R packages from a Fedora perspective

José Abílio Matos*
Porto University - Portugal

Abstract

Fedora is a Linux distribution that showcases the latest in free and open source software. It serves as the common root where other Linux distributions branch, with the best known being Red Hat Enterprise Linux, CentOS and Scientific Linux.

Both Fedora, R and most R packages are free software as defined by the Free Software Foundation. Although Fedora and R share such an important feature technically they have different goals. The purpose of R is to work in the largest possible set of platforms (assuming that there are interested developers). The purpose of Fedora is to package the largest possible set of free software packages and have them smoothly integrated into a single set.

When packaging R packages in Fedora the chalenge becomes then how to bring together these different goals. This talk deals with some of these issues. It should be noted that these challenges are common to other free software projects like Perl, Python, TEX (and others languages) on one side and Linux distributions on the other.

*jamatos@fep.up.pt

Desirabilitiy functions in multicriteria optimization Observations made while implementing desiRe

Olaf Mersmann[1] Heike Trautmann[1] Detlef Steuer[2] Claus Weihs[1]
Uwe Ligges[1]

Desirability functions and desirability indices are powerful tools for multicriteria optimization und multicriteria quality control purposes. The package desiRe not only provides functions for computing desirability functions of Harrington- (Harrington, 1965) and Derringer/Suich-type (Derringer and Suich, 1980) but also allows the specification of functions in an interactive manner. Density and distribution functions of the desirability functions and the desirability index are integrated including the possibility of random number generation (Steuer, 2005), (Trautmann and Weihs, 2006). Optimization procedures for the desirability index and a method for determining the uncertainty of the optimum influence factor levels (Trautmann and Weihs, 2004) as wells as a control chart for the desirability index with analysis of out-of control-signals are implemented (Trautmann, 2004). The Desirability Pareto-Concept allows focussing on relevant parts of the Pareto-front by integrating a-priori-expert-knowledge in the multicriteria optimization process (Mehnen et al., 2007).

We will focus on the implementation of the Desirability Pareto-Concept in R. First we will give a short review of the traditional optimization strategy using desirability indices. Then, after showcasing NSGA-II (Deb et al., 2002), we will briefly talk about how desirability functions can be integrated into optimization procedures that estimate the pareto front. Finally some of the problems faced during the development will be discussed. These include interfacing R and C code and using functions as first class objects.

In addition a short overview of the package will be given.

References

K. Deb, A. Pratap, and S. Agarwal. A Fast and Elitist Multiobjectiv Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(8):182–197, 2002.

G. C. Derringer and R. Suich. Simultaneous Optimization of Several Response Variables. *Journal of Quality Technology*, 12:214 – 219, 1980.

J. Harrington. The desirability function. *Industrial Quality Control*, 21:494 – 498, 1965.

J. Mehnen, H. Trautmann, and A. Tiwari. Introducing User Preference Using Desirability Functions in Multi-Objective Evolutionary Optimisation of Noisy Processes. In Kay Chen Tan and Jian-Xin Xu, editors, *CEC 2007, IEEE Congress on Evolutionary Computation*, pages 2687–2694, Swissotel The Stamford, Singapore, 2007.

D. Steuer. *Statistische Eigenschaften der Multikriteriellen Optimierung mittels Wünschbarkeiten*. PhD thesis, Technische Universität Dortmund, 2005. URL http://hdl.handle.net/2003/20171.

H. Trautmann. *Qualitätskontrolle in der Industrie anhand von Kontrollkarten für Wünschbarkeitsindizes – Anwendungsfeld Lagerverwaltung*. PhD thesis, Technische Universität Dortmund, 2004. URL http://hdl.handle.net/2003/2794.

H. Trautmann and C. Weihs. Uncertainty of the Optimum Influence Factor Levels in Multicriteria Optimization Using the Concept of Desirability. Technical Report 23/2004, SFB 475, Statistics Faculty, Technische Universität Dortmund, Germany, 2004.

H. Trautmann and C. Weihs. On the distribution of the desirability index using Harrington's desirability function. *Metrika*, 63(2):207–213, 2006.

[1] Fakultät Statistik, Technische Universität Dortmund
[2] Fakultät WiSo, Helmut-Schmidt Universität Hamburg

An automated R tool for identifying individuals with difficulties in a large pool of raters

Pete Meyer and Shaun Lysen
Google, Inc.
Santa Monica, California, USA

R is used extensively by the analysts at Google for analyzing everything from very small to very large datasets, from one-off analyses to regular production runs. In this talk we describe the use of R in flagging raters involved in the assessment of ad quality, who appear to be having difficulty performing their rating tasks. The use of this R script has resulted in an increase in system efficiency, improved timeliness of responding to rater needs, and decreased burden on those managing the raters.

The package RMySQL allows R to seamlessly integrate with MySQL databases, enabling data access directly to the production databases containing rater scores. Likewise, the R2HTML package provides output in a browser supported format, enabling report generation that can display web content and which enables movement between summary tables and supporting documentation using hyperlinks. Leveraging these features of R, we describe generating flags for three warning signs of rater difficulty:

1. excessive run lengths of repeated values,
2. the repetitive use of identical values for two distinct measures, and
3. identifying sequences of scores that appear to be assigned randomly rather than specific to the ads involved.

These tests could not be done by eye, either because of the large number of tasks involved or because they depend upon comparisons to reference distributions that are not visually apparent. However, those managing the raters easily grasp the conceptual basis for the tests and the summary tables contain hyperlinks to documentation that enables them to quickly find, cut and paste constructive feedback to the raters into emails in a simple and efficient manner. While these flags would be difficult to program in SQL, they are straightforward in R.

The strucplot framework for Visualizing Categorical Data

David Meyer Achim Zeileis Kurt Hornik

The **vcd** package (‘Visualizing Categorical Data’) has been around for quite a while now. This talk demonstrates the capabilities of a major part in this package: the strucplot framework.

We give an overview on how state of the art displays like mosaic and association plots can be produced, both for exploratory visualization and model-based analysis. Exploratory techniques will include specialized displays for the bivariate case, as well as pairs plot-like displays for higher-dimensional tables. As for the model-based tools, particular emphasis will be given to methods suitable for the visualization of conditional independence tests (including permutation tests), as well as for the visualization of particular GLMs (such as log-linear models).

Random Forests for eQTL Analysis: A Performance Comparison

Jacob J. Michaelson and Andreas Beyer
Biotechnology Center, TU Dresden, Germany

In recent years quantitative trait locus (QTL) methods have been combined with microarrays, using gene expression as a quantitative trait for genetic linkage analysis. Finding genetic loci significantly linked to the expression of a gene can help to identify regulators of the expressed gene. Traditional QTL methods used to find expression quantitative trait loci (eQTL) typically apply a univariate model to each genotyped locus in order to assess linkage to the quantitative trait. This univariate approach makes it difficult to uncover the interacting genes in the upstream regulatory pathway of the target. As has been previously suggested[1], in this work we view the eQTL problem as one of multivariate model selection: finding the genotyped loci which together best explain the variability of target gene expression in a population. We performed regression with Random Forests using the genotyped loci as predictor variables and the gene expression as the response. Measures of variable importance returned by Random Forests were used in locating eQTL. To assess whether this was a valid approach to eQTL, we determined eQTL for transcriptional targets of several canonical regulatory pathways using both Random Forests and several conventional QTL methods provided by the R `qtl` package. Gene expressions derived from several tissues of recombinant inbred mouse strains were used, and each eQTL method was evaluated for its ability to recapitulate known members of the canonical regulatory pathways. The results of our work demonstrate the biological validity and performance advantages of using Random Forests as a tool for finding eQTL.

References

[1] K. W. Broman and T. P. Speed, "A model selection approach for the identification of quantitative trait loci in experimental crosses," *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 641–656, 2002.

Cross-sectional and spatial dependence in panels

Giovanni Millo

March 31, 2008

Abstract

Econometricians have recently turned towards the problems posed by cross-sectional dependence across individuals, which may range from inefficiency of the standard estimators and invalid inference to inconsistency. Panel data are especially useful in this respect, as their double dimensionality allows robust approaches to general cross-sectional dependence.

A general object oriented approach to robust inference is available in the R system (Zeileis, 2004), for which all that's needed are coefficients $\hat{\beta}$ and robust estimators for $vcov(\hat{\beta})$. An useful implementation is, e.g., in linear hypotheses testing (see Fox, package `car`). The `plm` package for pael data econometrics already has features for heteroskedasticity– and serial correlation–robust inference (Croissant and Millo, forthcoming).

If cross-sectional dependence is detected, using a robust covariance estimator allows valid inference. I describe the implementation in the `plm` package for panel data econometrics of:

- tests for detecting cross-sectional dependence in the errors of a panel model (Friedman 1928, Frees 1995, Pesaran 2004)
- robust estimators of covariance matrices for doing valid inference in the presence of cross-sectional dependence (White 1980, Beck and Katz 1995, Driscoll and Kraay 1998)

If a particular spatial structure is assumed, this allows a parsimonious characterization of spatial dependence but, on the converse, the resulting models are computationally expensive to estimate, all the more so in the panel case. Efficient ML estimators for spatial models on a cross-section (Anselin 1988) are implemented in the `spdep` package (Bivand et al.). I describe implementation in a forthcoming package of

- marginal and conditional LM tests for spatial correlation, serial correlation and random effects (Baltagi, Song, Jung and Koh 2007)
- ML estimators for panel models including spatial lags, spatial errors and possibly serial correlation (Anselin 1988, Elhorst 2003, Baltagi, Song, Jung and Koh 2007)

I illustrate the functionalities by application to Munnell's (1990) data on 48 USA states observed over 17 years. On an ordinary desktop machine, the estimators and tests all take under one minute (few seconds for the basic ones). The ML approach is nevertheless structurally limited to a few hundred cross-sectional observations, so further work is warranted to implement Kapoor, Kelejian and Prucha (2007)'s GM approach, which promises to handle problems with n in the thousands.

Resolving components in mass spectrometry data: parametric and non-parametric approaches

Katharine M. Mullen, Ivo H. M. van Stokkum
Department of Physics and Astronomy
Vrije Universiteit Amsterdam
E-mail: {kate|ivo}@nat.vu.nl

March 31, 2008

A fundamental problem in mass spectrometry data analysis is decomposition of a matrix of measurements D, the rows of which represent times and the columns of which represent mass-to-charge ratio, into two matrices C and S, so that $D = CS^T$ and column i of C represents a component contributing to the data with respect to time (called an *elution profile*), and column i of S represents the mass spectrum of that component. This decomposition allows the compounds in a complex sample to be identified by taking the maximum of the elution profile of a component (that is, its *retention time*) and its mass spectrum and matching these properties to those of a known compound stored in a database.

A popular nonparametric means of resolving C and S given D is multivariate curve resolution alternating least squares (MCR-ALS), which combines the alternating least squares algorithm with constraints to impose nonnegativity, unimodality, selectivity, etc. MCR-ALS also allows the resolution of components in many datasets $D_1, \ldots, D_K$ simultaneously. We present a package **ALS** to perform MCR-ALS in R. While the package can be applied to any kind of data, it includes functions to plot mass spectra in particular.

A new methodology for resolving C and S given D currently in development uses a parametric description for C (in which components are usually described by functions based on a exponentially modified Gaussian), and optimizes the resulting separable nonlinear least squares problem to improve estimates for nonlinear parameters, while treating the mass spectra S as conditionally linear parameters. Like MCR-ALS, the methodology is well-suited to resolving components in many datasets simultaneously. We present options for the package **TIMP** that implement this parametric model-based methodology, and address issues such as outliers, a baseline, and instrument saturation.

Package Development in Windows

Duncan Murdoch

Developing R packages in Windows is much like developing them on Unix/Linux systems, except that most Windows users don't have the necessary tools installed. In this talk I will describe how to get the tools (which is much easier now than it was even two years ago), and give an overview of how to use them.

I will follow this with a demonstration of how to put together a simple package including external C code. The issues here are common to all platforms: how to set up the package, how to install and test it, and how to package it for distribution to others.

Speeding up R by using ISM-like calls

Junji Nakano
The Institute of Statistical Mathematics

Ei-ji Nakama
COM-ONE Inc.

R sometimes analyzes huge amount of data and requires huge size of memory operation for them. Many operating system have calls to help handling such huge memory. For example, Solaris has ‘ISM (Intimate Shared Memory)’ mechanism, Linux has ‘Huge TLB (Translation Look aside Buffer)’ and AIX has ‘Large Page’. OS usually translates 4-8 KB logical addresses to physical addresses at a time. These ISM-like mechanisms can change this size to much larger, such as 2-256 MB to speed up handling large memory. However, the cost of translation between logical addresses and physical addresses is called ‘TLB miss’ and sometimes becomes a bottle-neck. We introduce the use of ISM-like mechanisms in R by adding a wrapper program on the memory allocation function of R and investigate the performance of them.

ccgarch: An R package for modelling multivariate GARCH models with conditional correlations

Tomoaki Nakatani

Department of Economic Statistics, Stockholm School of Economics,
P.O. Box 6501, SE-113 83 Stockholm, Sweden
E-mail: sttn@hhs.se

Abstract

The multivariate GARCH models with explicit modelling of conditional correlations (the CC-GARCH models) have been widely used in modelling high-frequency financial time series. Examples include the Constant Conditional Correlation GARCH, the Dynamic Conditional Correlation GARCH, and the Smooth Transition GARCH models and their extensions to allow for volatility spillovers.

The package ccgarch provides functionality for estimating the major variants of the CC-GARCH models in arbitral dimensions. Both normal and robust standard errors for the parameter estimates are calculated through analytical derivatives. Numerical optimisations are carried out in such a way that negative volatility spillovers are allowed. The package is capable of simulating data from the major family of the CC-GARCH models with multivariate normal or student's t innovations. Procedures for misspecification diagnostics such as a test for volatility interactions are also included in ccgarch.

In presentation, we will discuss usefulness, limitation and directions for modification of the package.

R meets the Workplace - Embedding R in Excel to make it more accessible

Erich Neuwirth, Faculty of Computer Science, University of Vienna

March 31, 2008

One of the problems withstanding a more widespread use of R by nonspecialists (i.e. users with neither a high proficiency in software controlled by a classical programming language paradigm nor with a deeper knowledge of statistical methods) is the difficulty of starting to use R.

Most statistical data become analyzable data by being entered into Excel. Therefore, being able to transfer data from Excel to R is a key issue for wider use of R. There are technical answers like the packages **RODBD** or **xlsReadWrite** which essentially allow transfer of data frames. **RExcel**, an add-in for Excel, offers very similar facilities, but also brings the sequential programming paradigm of R and the dependency base automatic recalculation model of Excel closer together. It allows to use R-expressions as Excel formulas, combining the power of R's computational engine with the dependency tracking mechanisms of Excel.

An additional problem is the syntactic complexity of R formulas. A very nice tool for "guided discovery learning" how to build R expressions is the **R Commander** which gives the user a menu driven interface to statistical methods comparable to, say, SPSS, but at the same time displays the the R expressions needed to produce the requests results. The user then has the option of modifying these expression to adapt the result (data or graphs) to his needs.

The latest incarnation of **RExcel** embeds **R Commander** within Excel. The **R Commander** becomes an Excel menu, and in this way the naive user is presented with an already well established interface to R as an extension to Excel.

R Commander allows developers to write plugins, i.e. their own extensions to the methods offered by the menu interface, and thereby becomes a hub for making any R method available through a menu driven interface. **RExcel is compatible** with this plugin mechanism, any extension to **R Commander** also becomes an extension of Excel.

The embedding mechanism of **R Commander** into Excel does not directly use the autmatic recalculation engine, but **R Commander** can be used to support "production" of R formulas which then can be turned into Excel formulas.

Combining the power of Excel's dependency tracking mechanism and spatial paradigm for establishing relationships, R's powerful programming paradigm and computational engine, and **R Commander**'s menu system ease the creation of R formulas seems to offer a very powerful combination of methodologies to make R more accessible to a much wider class of users than R alone.

Proposal for useR!: ‘READ.ISI’

Rense Nieuwenhuis*

March 27, 2008

1 Background

Due to technological and software development, it sometimes is no longer possible to automatically read older data-files into statistical software. Especially data-files that originate from the times magnetic tapes were used to store data are often distributed as raw (ASCII) data, without proper means to read those data into statistical packages.

However, for those interested in using data to perform longitudinal analyses, these older sets of data are very valuable.

In the Netherlands, the national archive for data storage (DANS) is currently organizing conferences on a unified and time-proof manner of storing data-files. But what to do with those data that already have become difficult to access?

2 The Problem

In a research project on fertility issues, it was found that the ‘World Fertility Surveys’[1] are stored in a format that is no longer (directly) accessible to commonly used statistical software. Only data converted to ASCII directly from magnetic tape and a code-book are provided. The code-books are in a format specific by the ‘International Statistical Institute’ (ISI) and provides for each variable information on starting and ending positions in the data-file, value- and variable labels and information on missing values. However, no statistical software package presently used is known to be able to automatically read data based on this type of code-book.

It was required to read all variables into the statistical software manually. Variable names and value labels have to be assigned manually as well. This is not an inviting process and a highly laborious when many variables are needed.

*Author can be contacted at:
Email: contact@rensenieuwenhuis.nl
Telephone: +31 6 481 05 683

[1] http://opr.princeton.edu/archive/wfs/

3 The Solution

This problem may however be solved – in select cases – by using R-Project. Applying the flexible data-structure provided by R-Project, it was possible to read and interpret the code-books (meant for the human eye) and to use this to automatically read the data, add value and variable labels, assign missing values, and to do this for whole data-sets at once. The resulting syntax was transformed to the function called 'READ.ISI'.

```
V106      141  2   0   1   88      Remariee
                                      0    Non
                                      1    Oui
                                     88    Non rompue
V107      143  2   1   3   88   99 Etat actuel                    V104
```

Above, a small fragment of one of the code-books is shown. The function READ.ISI reads these fixed-width ASCII file twice. Once to read the variable names, labels, starting- and ending positions, and missing values (on the first and last row of the example above). The second time to read the value labels (in the middle rows of this example). As is illustrated on the last row of the fragment above, the value labels of variable 'V107' are identical to that of 'V104'. This is taken into account as well. Based on this automatic interpretation of this code-book, either the ASCII data-file is read, or a SPSS-syntax is created illustrating that people using other statistical packages can benefit from this function as well.

4 Proposal

Applicable to a select number of R users, but highly valuable for those who want to use (some) old data, this approach will help and inspire those who are interested in longitudinal analysis. Possibly, this approach can be transferred to the code-books of other collections of data. Therefore, I feel that this would make an excellent poster presentation on the userR! conference. On this poster the problem could be clearly illustrated and the steps needed to read this type of data automatically will be identified.

Invariant coordinate selection for multivariate data analysis - the package ICS

Klaus Nordhausen[1], **Hannu Oja**[1] **and David E. Tyler**[2]

[1] *University of Tampere, Finland*
[2] *Rutgers, The State University of New Jersey, USA*

Abstract

Invariant coordinate selection (*ICS*) has recently been introduced by Tyler et al. (2008) as a method for exploring multivariate data. It includes as shown in Oja et al. (2006) as a special case a method for recovering the unmixing matrix in independent components analysis (*ICA*). It also serves as a basis for classes of multivariate nonparametric tests. The aim of this paper is to briefly explain the *ICS* method and to illustrate how various applications can be implemented using the R-package ICS. Several examples are used to show how the *ICS* method and ICS package can be used in analyzing a multivariate data set.

Keywords

Independent components analysis, invariant coordinate selection, transformation-retransformation method.

References

H. Oja, S. Sirkiä, and J. Eriksson (2006). Scatter matrices and independent component analysis. *Austrian Journal of Statistics*, 35, 175-189.

D. E. Tyler, F. Critchley, L. Dümbgen, and H. Oja (2008). Exploring multivariate data via multiple scatter matrices. Submitted.

Automating Business Modeling with the `AutoModelR` Package

Derek McCrae Norton*

March 31, 2008

Abstract

Many issues arise in the business environment, like other fields, which must be addressed. Many of these issues have been addressed in other fields separately, but need to be jointly addressed in a business environment. The objective of `AutoModelR` is to attempt to address these issues in an automatic manner.

The issues addressed are: Exploratory Data Analysis, Dimension Reduction, and Automatic Initial Modeling. To address these issues, a data set is passed to `AutoModelR` consisting of a dependant variable and one or more (often many more) independent variables. The first step is to remove variables which have zero variation, missing value percentages above a threshold, and variables with one unique value accounting for more than some threshold percentage. A report is then generated using `Sweave` which gives tables of descriptive statistics for numeric and factor data separately as well as graphical displays of the data. The next step is an application of a filter type dimension reduction to arrive a smaller data subset for initial modeling. The last step is to automatically fit various simple models determined by the type of dependent variable and report on those fits.

`AutoModelR` is an attempt to automate some of the repetitive steps in modeling, so that more time can be spent on advanced modeling.

*InterContinental Hotels Group, USA

rPorta - An R Package for Analyzing Polytopes and Polyhedra

Robin Nunkesser[1], Silke Straatmann[2], and Simone Wenzel[2]

[1] Department of Computer Science, TU Dortmund, 44221 Dortmund
[2] Department of Statistics, TU Dortmund, 44221 Dortmund

Summary. In application fields like mechanics, economics and operations research the optimization of linear inequalities is of interest. There are algorithms handling such problems by utilizing the theory of polyhedral convex cones (PCCs) in particular that PCCs can be defined by a span form or a face form. These two representations are called *double description pair* and represent the link between PCCs and linear inequalities. In practice, the transformation from one form into the other is often useful. Here, we present an R package called rPorta providing a set of functions for polytopes and polyhedra mainly intended for the double description pair.

The underlying algorithms used are part of a program named PORTA (Polyhedron Representation Transformation Algorithm) that comprises a collection of routines for analyzing polytopes and polyhedra in general. In particular, it supports both representations of PCCs, i.e. the representation as a set of vectors and as a system of linear equations and inequalities. The main functionality of PORTA is the transformation from one representation to the other, but PORTA also provides other handy routines, e.g. to check whether points are contained in a PCC or not. All functions of PORTA read and write data from text files containing one of the two representations, i.e. the user interface of PORTA communicates through text files. Our package rPorta provides an interface to use the routines of PORTA in R by enwrapping the text file information in S4 Objects. This ensures an easy-to-use and R friendly way to run the functions of PORTA.

In addition, an application of rPorta in design of experiments is presented. In engineering processes the parameter space often contains parameter settings that produce missing values in the design since the produced workpiece fails. The goal is to create a design where the design points concentrate in the feasible area, although the boundaries where missing values are liable to occur are not known. The design is created sequentially and emerging missing values are used to update the excluded failure regions, which is done with the help of PCCs determined with rPorta.

A first glimpse into 'R.ff', a package that virtually removes R's memory limit

Jens Oehlschlägel Daniel Adler Oleg Nenadic Walter Zucchini

The availability of large atomic objects through package 'ff' can be used to create packages implementing statistical methods specifically addressing large data sets (like subbagging or package biglm). However, wouldn't it be great if we could apply all of R's functionality to large atomic data? Package 'R.ff' is an experiment to provide as much as possible of R's basic functionality as 'ff-methods'. We report first experiences with porting standard R functions to versions operating on ff objects and we discuss implications for package authors (and maybe also R core). Instead of a summary, here we just quicken your appetite through the list of functions and operators where we have first experimental ports:

! != \%\% \%*\% \% /\% \& | * + - / < <= == > >= ^ abs acos acosh asin asinh atan atanh besselI besselJ besselK besselY beta ceiling choose colMeans colSums cos cosh crossprod cummax cummin cumprod cumsum dbeta dbinom dcauchy dchisq dexp df dgamma dgeom dhyper digamma dlnorm dlogis dnbinom dnorm dpois dsignrank dt dunif dweibull dwilcox exp expm1 factorial fivenum floor gamma gammaCody IQR is.na is.nan jitter lbeta lchoose lfactorial lgamma log log10 log1p log2 logb mad order pbeta pbinom pcauchy pchisq pexp pf pgamma pgeom phyper plnorm plogis pnbinom pnorm ppois psigamma psignrank pt punif pweibull pwilcox qbeta qbinom qcauchy qchisq qexp qf qgamma qgeom qhyper qlnorm qlogis qnbinom qnorm qpois qsignrank qt quantile qunif qweibull qwilcox range range rbeta rbinom rcauchy rchisq rexp rf rgamma rgeom rhyper rlnorm rlogis rnbinom rnorm round rowMeans rowSums rpois rsignrank rt runif rweibull rwilcox sample sd sign signif sin sinh sort sqrt summary t tabulate tan tanh trigamma trunc var.

BMDS: A Collection of R Functions for Bayesian Multidimensional Scaling

Kensuke Okada & Kazuo Shigemasu, The University of Tokyo

Abstract

Bayesian MDS has recently attracted a great deal of researchers' attention because (1) it provides a better fit than classical MDS and ALSCAL, (2) it provides estimation errors of the distances, and (3) the Bayesian dimension selection criterion, MDSIC, provides a direct indication of optimal dimensionality; see the original paper by Oh & Raftery (2001). However, Bayesian MDS is not yet widely applied in practice. One of the reasons can be attributed to the apparent lack of software: there is none except for the original Oh & Raftery's code, which requires good experience in Fortran programming and the IMSL library, which is a commercial library for numerical calculation. It may be difficult to require such environment for many researchers.

Considering this situation, we propose a set of R functions, BMDS, to perform Bayesian MDS and to evaluate the results. Using BMDS, researchers can (1) perform Bayesian estimation in MDS, (2) check the convergence of Markov chain Monte Carlo (MCMC) estimation, (3) evaluate the optimal number of dimensions, (4) evaluate the estimation errors and (5) plot the resultant configurations. Also, using BMDS users can comparatively evaluate the result of Bayesian and classical MDS in terms of the value of stress and the plot of observed and estimated distances.

In our functions, we made use of WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007) via R2WinBUGS package (Sturtz, Ligges, & Gelman, 2005) for MCMC estimation. Because the Bayesian MDS model is rather complex and it is impossible to use single WinBUGS script for any model, our bmds() function automatically produces a BUGS script that is adequate for the current data every time we run the R function. By using WinBUGS in this way we can speed-up the MCMC estimation while maintaining the readability of the code, which tends to be complex in Bayesian estimation.

References

Oh, M-S. & Raftery, A.E. (2001) Bayesian multidimensional scaling and choice of dimension. *Journal of American Statistical Association*, 96(455), 1031-1044.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2007) *WinBUGS Version 1.4.3 User Manual.* MRC Biostatstics Unit.

Sturtz, S., Ligges, U. & Gelman, A. (2005) R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12(3), 1-17.

Forecasting species range shifts: a Hierarchical Bayesian framework for estimating process-based models of range dynamics

J. Pagel & F. Schurr

Shifts of species ranges have been widely observed as ‘fingerprints’ of climate change and more drastic shifts are expected in the coming decades. Current studies projecting range shifts in response to climate change are predominantly based on phenomenological models of potential climate space (climate envelope models). These models assume that species distributions are at equilibrium with climate, both at present and in the future. A more reliable projection of range dynamics under environmental change requires process-based models that can be fitted to distribution data and permit a more comprehensive assessment of forecast uncertainties [1]. To achieve this goal, we develop a Hierarchical Bayesian framework [2] that utilizes models of local population dynamics and regional dispersal to link data on species distribution and abundance to explanatory environmental variables.

In a simulation study we investigate the performance of this approach in relation to the biological characteristics of the target species and the quantity and quality of biological information available. We use R to implement an integrated routine that combines a grid-based ecological simulation model and a ‘virtual ecologist’ with efficient MCMC algorithms (e.g. DRAM [3]) for sampling from the full posterior distribution of model parameters and derived predictions of spatially distributed abundances under prescribed climatic changes. This enables us to run a range of virtual scenarios differing in both ecological assumptions and sampling design in order to examine how forecast uncertainty depends on a species' ecology as well as on data quality and quantity.

References

[1] Araujo, M. B., and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography **33**:1677-1688.

[2] Clark, J. S., and A. E. Gelfand. 2006. A future for models and data in environmental science. Trends in Ecology & Evolution **21**:375-380.

[3] Haario, H., M. Laine, A. Mira, and E. Saksman. 2006. DRAM: Efficient adaptive MCMC. Statistics and Computing **16**:339-354.

Joern Pagel, Institute of Biochemistry and Biologie, University Potsdam, Germany.
joern.pagel@uni-potsdam.de

Random Forests and Nearest Shrunken Centroids for the Classification of eNose data

Matteo Pardo*, Giorgio Sberveglieri
Sensor Lab, CNR-INFM & University of Brescia, Brescia, Italy
*pardo@ing.unibs.it

Artificial Olfactory Systems or eNoses are instruments that analyze gaseous mixtures for discriminating between different (but similar) mixtures and, in the case of simple mixtures, quantify the concentration of the constituents. eNoses consist of a gas sampling system (for a reproducible collection of the mixture), an array of chemical sensors, electronic circuitry and data analysis software (Pearce, 2003). Random Forests (RF) and (NSC) are state of the art classification and feature selection methodologies and have never been applied to eNose data.

RFs are ensembles of trees, where each tree is constructed using a different bootstrap sample of the data and each node is split using the best among a subset of features randomly chosen at that node. RF has only two hyper parameters (the number of variables in the random subset at each node and the number of trees in the forest) (Breiman, 2001). NSC classification makes one important modification to standard nearest centroid classification. It "shrinks" each of the class centroids toward the overall centroid (for all classes) by an amount called the threshold (Tibshirani, et al., 2003).

In this paper we compare the classification rate of RF, NSC and Support Vector Machines (SVM) -which we consider as a top level reference method- on three eNose datasets for food quality control applications. Classifiers' parameters are optimized in an inner cross-validation cycle and the error is calculated by outer cross-validation in order to avoid any bias. To carry out computations we used the R package MCRestimate (Ruschhaupt, et al., 2004). MCRestimate is built on top of a number of R packages, e.g. the *randomForest* package (Liaw and Wiener, 2002).

We were interested in three computational aspects:

1. Relative performance of the three classifiers.
2. Since nested cross-validation is computationally expensive we also investigate the dependence of the error on the number of inner and outer folds. We considered a grid of 25 outer folds/ inner folds numbers: outer CV folds: 2, 4, 6, 8, 10; inner CV folds: 2, 4, 6, 8, 10. Altogether this means training e.g. 45050 SVM.
3. Feature rankings produced by RF and NSC.

We find that:

1. SVM and RF perform similarly (each classifier does better on one problem), while NSC consistently performs worse. NSC is by far the simpler (and faster) classifier.
2. There is a slight dependence on the number of external CV folds (particularly in the *fungi* dataset, where four external CV folds produce a consistently –over internal CV folds number- higher classification rate), while the number of inner CV folds seems to be immaterial.

   2x2 nested CV is often enough for a good result. With respect to 10x10 CV, 2x2 CV requires 4% of the training time, so this result may spare quite some time in future computational studies.
3. Of the 30 original features, RF and NSC have the same two top positions. Further, they share other four features in the top ten. Other four features have quite, or even very different rankings. In fact, NSC – differently from RF- ranks features individually and independently in the classifier construction process. In this way, on the one hand it cannot consider the joint discrimination capabilities of features groups and on the other hand it does not exclude correlated features.

Breiman, L. (2001) Random forests, *Machine Learning,* **45**, 5 - 32.

Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *The Newsletter of the R Project.*

Pearce, T.C. (2003) *Handbook of machine olfaction : electronic nose technology*. Wiley-VCH, Weinheim [Germany].

Ruschhaupt, M., Huber, W., Poustka, A. and Mansmann, U. (2004) A Compendium to Ensure Computational Reproducibility in High-Dimensional Classification Tasks, *Statistical Applications in Genetics and Molecular Biology,* **3**, 37.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003) Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays, *Statist. Sci.*, **18**, 104-117.

TCL Expect: Yet another way to ~~do the wrong thing~~ develop GUI for R

Ivailo Partchev
University of Jena, Germany

Point-and-click interfaces, as opposed to a sensible command language, will always have their supporters and critics. We present an approach to developing a point-and-click interface based on the TCL/TK extension, Expect.

Compared to the internal support of TCL/TK through packages like tcltk, rpanel, tkrplot, and others, the Expect approach is slower and less appropriate for animated displays. On the positive side, applications are very easy to develop, the very existence of R is hidden from the user, and the interactive nature of R is exploited fully.

As an example, we present a toy application (a t-test or a regression), and a larger application for the analysis of treatment effects in experimental and quasi-experimental research.

References

Bowman, A.W., Crawford, E., Alexander, G. & Bowman, R.W. (2006). rpanel: simple interactive controls for R functions using the tcltk package. *Journal of Statistical Software 17,* issue 9

Libes, Don (1995). *Exploring Expect: A Tcl-based Toolkit for Automating Interactive Programs.* O'Reilly.

Steyer, Rolf et al. (2008). *Causal Effects in Experiments and Quasi-Experiments* (in print)

Dynamic Linear Models in R

Giovanni Petris
University of Arkansas, USA

Dynamic Linear Models (DLMs) are a very flexible tool for time series analysis. In this talk we introduce an R package for the analysis of DLMs. The design goal was to give the user maximum flexibility in the specification of the model. The package allows to create standard DLMs, such as seasonal components, stochastic polynomial trends, regression models, autoregressive moving average processes and more, and it also provides functions to combine in different ways elementary DLMs models as building blocks of more complex univariate or multivariate models. For added flexibility, completely general constant or time-varying DLMs can be defined as well.

The drawback of allowing so general models to be used is that for many DLMs the standard algorithms for Kalman filtering and smoothing are not numerically stable. The issue has been addressed in the package by using filtering and smoothing algorithms that are based on the recursive calculation of the relevant variance matrices in terms of their singular value decomposition (SVD). The same SVD-based algorithm employed for Kalman filter is also used to find maximum likelihood estimates of unknown model parameters.

In addition to filtering, smoothing and maximum likelihood estimation, the package provides some functionality for simulation-based Bayesian analysis of DLMs. A function that generates the unobservable states from their posterior distribution is available, as well as a multivariate version of adaptive rejection Metropolis sampling, which can be used to generate random vectors having an essentially arbitrary continuous distribution. Both generators can be fruitfully employed within a Gibbs sampler or other Markov chain Monte Carlo algorithm.

In the talk I will give an overview of the most important features of the package, illustrating them with practical examples.

Objects, clones and collections: ecological models and scenario analysis with simecol

Thomas Petzoldt[a]

R is increasingly accepted as one of the standard environments for ecological data analysis and ecological modeling. An inreasing collection of packages explicitly developed for ecological applications (Kneib and Petzoldt, 2007) and a number of textbooks that use R to teach ecological modeling (Ellner and Guckenheimer, 2006; Bolker, 2007) are just an indicator for this trend. In this context, the package **simecol** (simulation of ecological models) was developed in order to facilitate implementation, analysis and share of simulation models by means of object oriented programming (OOP) with S4 classes.

The idea behind **simecol** is an object model of ecological models, i.e. to put everything needed (state variables, parameters, inputs, equations) to define an ecological model together in one code object (an instance of a subclass of `simObj`), that can be handled by appropriate generic functions.

Because all essentials of a particular model are encapsulated in one code object, individual instances can simply be copied with the assignment operator `<-`. These clones can be modified with accessing functions to derive variants and scenarios without copying and pasting source code. This way, it is also possible to interactively enable or disable online-visualisation (using observer-slots), to adapt numerical accuracy or to compare scenarios with different structure, e.g. ecological models with different types of functional response.

After a short overview the presentation will concentrate on examples how to clone, modify and extend **simecol** objects and how to organize scenario analyses. From the user's perspective these are:

1. Implement a model-prototype by filling out a pre-defined structure or by modifying existing examples,
2. Simulate and test your model with existing solvers or develop your own algorithms,
3. Clone your prototype object and modify data and/or code of individual clones to generate scenarios,
4. Simulate, analyse, compare scenarios, fit parameters, supply observer functions for run-time visualisation.
5. Save your model object and share it with your colleagues, students and readers of your papers.

Individual **simecol**-objects can be stored persistently as binaries or human-readable list representation which can be distributed in reproducible and fully functional form. In addition, it is also possible to assemble collections of models as separate R-packages, together with necessary documentation and examples, e.g. to reproduce the figures of a paper, or with additional classes and functions extending **simecol**.

References

Bolker, B. (2007). *Ecological Models and Data in R*. Princeton University Press, Princeton. in press.

Ellner, S. P. and Guckenheimer, J. (2006). *Dynamic Models in Biology*. Princeton University Press.

Kneib, T. and Petzoldt, T. (2007). Introduction to the special volume on ecology and ecological modeling in R. *Journal of Statistical Software*, 22(1):1–7.

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

Bayesian Modelling in R with rjags

Martyn Plummer

March 31, 2008

JAGS (Just Another Gibbs Sampler) is a portable engine for the BUGS language, which allows the user to build complex Bayesian probability models and generates random samples from the posterior distribution of the model parameters using Markov Chain Monte Carlo (MCMC) simulation.

The **rjags** package currently provides a small library that permits a direct interface from R to the main JAGS library. Future versions of **rjags** should provide additional Bayesian modelling tools. However, there are still outstanding problems, such as the choice of R class for representing MCMC output, that still need to be resolved.

This talk will discuss some of the issues involved in creating a portable interface package for R. I will illustrate the way that R and JAGS can be combined to provide tools for Bayesian modelling, such as the deviance information criterion (DIC) and related penalized loss functions for model comparison.

Direct Marketing Analytics with R

Jim Porzak
Senior Director of Analytics,
Responsys, San Francisco, CA
JPorzak@responsys.com

Abstract

Direct marketing has traditionally claimed to be a quantified discipline. Marketing campaigns are measured by actual results. Testing is routinely done to improve performance. The concept of a control group is ingrained in the direct marketing culture. Modern marketing is based on *relevant* messaging. Segmentation is a practical way to deliver relevant communication to individuals.

That said, in practice there is a lot of confusion as to exact definition and implementation of specific metrics, methods, and statistical procedures. And, unfortunately, rigor is often more hype than actuality.

The goal of the dma package is to provide the direct marketing community with a well defined set of procedures to easily do direct marketing analytics. Since the source code is available, there is total visibility into the methods. No black box. No proprietary secrets.

At the end of the day, all needed analytics could be done manually in R, or most any other computing platform. The dma package wraps and bundles appropriate R procedures in a way direct marketers will, hopefully, find intuitive and natural to use. Having a single code base should eliminate errors and make results reproducible. We also leverage the graphics power of R to add visualizations that make the findings more intuitive than just presenting numerical results.

Direct marketing has certain characteristics which influence the design. These include:

- a huge N.
- very small proportions.
- users typically work in a Windows/Office environment.
- the paradox of the need for rigor and client comprehensibility.

There are three main modules in the dma package.

Basic Metrics: Straightforward definitions of key direct marketing performance indicators. These include response, profitability, and LTV metrics. The point is to implement these metrics based on accepted best practices in an open and transparent way.

Testing: In addition to single and multiple tests against a control (A/B and A/BCD... tests) visualizations are generated that make marketing campaign test results obvious to non-statisticians. Methods are wrapped to make them accessible to direct marketing staff without requiring any understanding of R.

Segmentation: The classic direct marketing segmentation method is RFM (Recency, Frequency, and Monetary). Using customer order history data, loaded into an orders object, recency and frequency are used to create actionable customer life-stage segments. Adding product purchase details allows creation of prospect relevant messaging tactics.

R for the Masses: Lessons learnt from delivering R training courses

Richard Pugh, Matt Aldridge – *Mango Solutions*

useR! 2008. Dortmund.

Abstract

Over the last 5 years, Mango have delivered R and S training courses to approximately 1,100 people at organizations around the world. To ensure the standard of training material is of the highest quality, regular reviews are held based on feedback and lessons learnt from the courses delivered.

This presentation describes the manner in which our training package has evolved based on the experience of Mango trainers and feedback from course attendees. We will also present our view on how a complex software language such as R is best taught, including the use of Mango *tip sheets* and continued support beyond the training period. We will also discuss the challenges of training non-statistical end users in languages like R.

Equilibrium Model Selection

Tomas Radivoyevitch, Department of Epidemiology and Biostatistics,
Case Western Reserve University, Cleveland, OH 44106 USA; txr24@case.edu

Ribonucleotide reductase (RNR) is precisely controlled to meet the dNTP demands of scheduled (replication driven) and unscheduled (repair driven) DNA synthesis. It has a small subunit R2 that exists almost exclusively as a dimer, and a large subunit R1 (R) that dimerizes when dTTP (t), dGTP, dATP, or ATP binds to its specificity site, and hexamerizes when dATP or ATP binds to its activity site. In general, RNR is modeled as a pre-equilibrium of proteins, ligands, and substrates whose parameters of interest are dissociation constants *K*, and a set of turnover rate parameters *k* that map distributions of active enzyme complexes into expected *k* measurements of mixtures. Because the masses of R1 and R2 are known, it is logical to focus first on *K* estimation from protein oligomer mass measurements, and later on *k* estimation from enzyme activity measurements. Further, it is also logical to begin with the simplicity of ligand-induced R1 dimerization.

The total concentration constraint full model for dTTP-induced R1 dimerization is

$$0 = p[R_T] - [R] - \frac{[R][t]}{K_{Rt}} - 2\frac{[R]^2}{K_{RR}} - 2\frac{[R]^2[t]}{K_{RRt}} - 2\frac{[R]^2[t]^2}{K_{RRtt}}$$

$$0 = [t_T] - [t] - \frac{[R][t]}{K_{Rt}} - \frac{[R]^2[t]}{K_{RRt}} - 2\frac{[R]^2[t]^2}{K_{RRtt}}$$

where the subscript T denotes totals (note that $[R]^2[t]^2/K_{RRtt} = [RRtt]$) and the probability that an R molecule is undamaged and capable of dimerizing is *p*. This full model generates 58 *a priori* plausible equilibrium models/hypotheses as follows. Firstly, *K*=∞ assumptions are used to remove specific terms one at time, two at a time, and so on, to yield $2^4 = 16$ models, each hypothesizing that the deleted complexes are not detectable above noise. Secondly, of these models, the 4 single *K* models yield 4 additional models via *K*=0 assumptions, each alleging that the free concentration of the reactant that is not in excess is indistinguishable from zero. Thirdly, after expanding *K* into products of strictly binary *K*, nine additional models that allege that some *K*s equal others also arise; these nine models correspond to hypotheses of independence between the R and t binding sites on R. Finally, for each model it can be hypothesized that the data are not rich enough to discriminate *p* close to one from *p* = 1, and this expands the model space by an additional factor of two to 58.

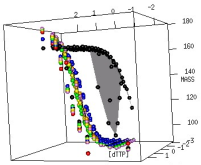
Using the following average mass output model

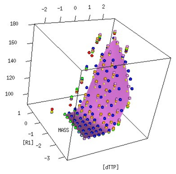
$$90\frac{[R]+[R_T](1-p)}{[R_T]} + 180\frac{2[RR]+2[RRt]+2[RRtt]}{[R_T]}$$

the 58 models were fitted to available data. The top 6 models based on AIC$_c$ are

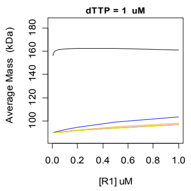
Model	Parameter	Initial value	Final value	CI
3M	RRtt	1.000	18.697	(4.807,72.966)
violet	p	1.000	1.000	fixed
	SSE	0.064	0.034	
	AIC$_c$	-48.066	-54.448	
3Mp	RRtt	1.000	5.558	(0.370,84)
blue	p	1.000	0.907	(0.787,1.044)
	SSE	0.064	0.027	
	AIC$_c$	-44.852	-53.308	
3Rp	p	1.000	0.822	(0.736,0.918)
black	RRtt	0.000	0.000	fixed
	SSE	0.106	0.041	
	AIC$_c$	-42.954	-52.590	
3I	RRt	1.000	49.568	(5.755,428)
green	RRtt	1.000	37.930	(5.003,290)
	p	1.000	1.000	fixed
	SSE	0.165	0.030	
	AIC$_c$	-35.303	-52.218	
2M	R_R	75.000	685.986	(2.801,162755)
yellow	RR_t	0.550	0.142	(0.005,3.975)
	RRt_t	0.550	0.142	constrained
	p	1.000	1.000	fixed
	SSE	0.041	0.032	
	AIC$_c$	-49.222	-51.815	
3F	Rt	1.000	91.059	(1.557,5324)
orange	RRtt	1.000	14.612	(2.545,84)
	p	1.000	1.000	fixed
	SSE	0.221	0.032	
	AIC$_c$	-32.422	-51.627	

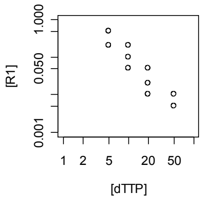
Of these, model 3Rp differs substantially from the other five in its predictions over physiological values of $[t_T]$=.1 to 50 μM and $[R_T]$=.005 to 1 μM.

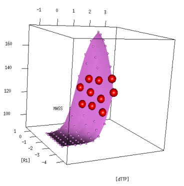




If 3Rp is rejected by a measurement of 95 kDa at $[t_T]$ = 1 μM and $[R_T]$ = 0.2 μM, the best next 10 measurements for discrimination between the remaining 5 models are the following points on the hill where their predictions differ most.







The methods, data, R functions and R scripts used to fit the model space can be found in

Radivoyevitch T: Equilibrium model selection: dTTP induced R1 dimerization. *BMC Systems Biology* 2008, **2** (1):15.

Abstract for UseR! 2008 Conference

RReportGenerator: Automatic reports from routine statistical analysis using R.

Wolfgang Raffelsberger, Luc Mouliner, David Kieffer, Yannick Krause and Olivier Poch

Laboratoire de BioInformatique et Génomique Intégratives (LBGI), IGBMC,
1 rue Laurent Fries, 67404 Illkirch-Strasbourg, France

With RReportGenerator we have developed a tool dedicated to performing automatic routine statistical analysis using R via a graphical user interface (GUI) in a highly user-friendly way that can be run on Windows and Linux platforms. The program is freely available under http://www-bio3d-igbmc.u-strasbg.fr/~wraff .

Since the command-line syntax of R is very powerful but difficult to access for non-statisticians, we have developed a simple graphical interface designed for routine execution of predefined "analysis scenarios" for a given problem (written as Sweave code). The key function of RReportGenerator consists in automatically generating a pdf-report combining results from statistical analysis, tables and figures. Depending on the analysis scenario chosen, reports can be accompanied by supplemental data-sets for exporting results to other programs, too.

At this point several applications (“analysis scenarios”) for quality control and low-level data analysis in the fields of transcription profiling (e.g. extensive QC for Affymetrix GeneChips or QC & data normalization of printed arrays), CGH-analysis (simultaneous comparison using multiple segmentation approaches) and transfected cell array (TCA) platforms have been developed and are getting further enhanced. For example, use of RReportGenerator may help technology platforms considerably as it produces automatically well documented analysis reports in a standardized format for transferring QC results and assay data to other research teams.

A special function of this GUI allows accessing directly the most recent versions of the distributed analysis scenarios and facilitates using automatically the most recent analysis scenarios. Finally, our automatic analysis platform is open to distribute contributed and documented analysis scenarios from the R community.

ESTIMATING EVOLUTIONARY PATHWAYS AND GENETIC PROGRESSION SCORES WITH RTREEMIX

Jörg Rahnenführer[1], Jasmina Bogojeska[2], Adrian Alexa[2], André Altmann[2], Thomas Lengauer[2]

[1]Fakultät Statistik, Technische Universität Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany

[2]Max Planck Institute for Informatics, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany

Email: rahnenfuehrer@statistik.uni-dortmund.de

In genetics, many evolutionary pathways can be modeled on the molecular level by the ordered accumulation of permanent changes. We have developed the class of mixture models of mutagenetic trees (Beerenwinkel et al., 2005a) that provides a suitable statistical framework for describing these processes. These models have been successfully applied to describe disease progression in cancer and in HIV. In cancer, progression is modeled by the accumulation of lesions in tumor cells such as chromosomal losses or gains (Ketter et al., 2007). In HIV, the accumulation of drug resistance-associated mutations in the genome is known to be associated with disease progression. Mutations in the genome of the dominant strain in the infecting virus population arise when a patient receives a specific medication.

From such evolutionary models, genetic progression scores can be derived that assign measures for the disease state to single patients (Rahnenführer et al., 2005). Progression of a single patient along such a model is typically correlated with increasingly poor prognosis. In the cancer application, we showed that higher genetic progression scores are significantly associated with shorter expected survival times in glioblastoma patients (Rahnenführer et al., 2005) and times until recurrence in meningioma patients (Ketter et al., 2007).

We present applications in this framework as well as the easy-to-use and compute-efficient R package *Rtreemix* for estimating such mixtures of evolutionary models from cross-sectional data. *Rtreemix* builds up on efficient C/C++ code provided in the Mtreemix package (Beerenwinkel et al., 2005b) for estimating mixture models. It contains additional new functions for estimating genetic progression scores with corresponding bootstrap confidence intervals for estimated model parameters. Furthermore, the stability of the estimated evolutionary mixture models can be analyzed (Bogojeska et al., 2008).

Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J., Lengauer, T. (2005a) Learning multiple evolutionary pathways from cross-sectional data, Journal of Computational Biology, 12(6), 584-598.

Beerenwinkel, N., Rahnenführer, J., Kaiser, R., Hoffmann, D., Selbig, J., Lengauer, T. (2005b) Mtreemix: a software package for learning and using mixture models of mutagenetic trees, Bioinformatics, 21(9), 2106-2107.

Bogojeska, J., Rahnenführer, J., Lengauer, T. (2008) Stability analysis of mixtures of mutagenetic trees, BMC Bioinformatics, 9(1): 165.

Rahnenführer, J., Beerenwinkel, N., Schulz, W.A., Hartmann, C., Deimling, A.V., Wullich, B., Lengauer, T. (2005) Estimating cancer survival and clinical outcome based on genetic tumor progression scores, Bioinformatics, 21(10), 2438-2446.

Ketter, R., Urbschat, S., Henn, W., Feiden, W., Beerenwinkel, N., Lengauer, T., Steudel, W.-I., Zang, K.D., Rahnenführer, J. (2007) Application of oncogenetic trees mixtures as a biostatistical model of the clonal cytogenetic evolution of meningiomas, International Journal of Cancer, 121(7), 1473-1480.

ESTIMATION OF STANDARD ERRORS IN NON-LINEAR REGRESSION MODELS: SPATIAL VARIATION IN RISK AROUND PUTATIVE SOURCES.

Ramis, Rebeca[1,2,3]; Diggle, Peter[1]; López-Abente, Gonzalo[2,3]

[1] *Department of Medicine, Lancaster University, UK.*
[2] *Cancer and Environmental Epidemiology Area, National Centre for Epidemiology. Carlos III Institute of Health, Madrid, Spain.*
[3] *CIBERESP*

Background

We consider the problem of investigating spatial variation in the risk of non-infectious diseases in populations exposed to pollution from one or more point sources.

The data most commonly available to study this question include case-counts (O_i) in each of a set of areas that partition the geographical region of interest, suitable denominators, Ei, proportional to the expected number of cases in each area, and the locations of the relevant point sources, from which we can compute distances d_{ij} between the *j*th focus and a reference location, typically the centroid, within the *I*th area. Also available in most applications are covariates relating to socio-economic status or other risk-factors associated with each area, which we denote by Z_k.

The standard approach to the analysis of data of this kind is a log-linear regression of the case-counts on the covariates, with log-transformed denominators as an offset variable. To model distance-related point source effects, a log-linear formulation is unrealistic because of the need to combine an elevated risk close to the source with a neutral long-distance effect. We therefore extend the model by including a non-linear distance function, $f(d_{ij})$, hence [1,2]:

$$O_i \sim Po(E_i \mu_i)$$

$$\mu_i = \rho \sum_k (\vartheta_k Z_{ik}) \prod_j f(d_{ij}); \qquad f(d_{ij}) = 1 + \alpha_j \exp\left[-(d_{ij} / \beta_j)^2\right]$$

Parameter estimation and standard error calculations

Generic functions available in R to fit non-lineal regression models include the *"gnlm"* library by J. K. Lindsey [3], which in turn uses the *nlm* function of Bates and Pinheiro to estimate the parameters. These functions use a numerical estimate of the Hessian matrix evaluated at the parameter estimate to calculate standard errors.

We have found that, for point source models like the one described above, even when numerically accurate values are returned for the maximum likelihood parameter estimates, the associated standard errors derived by inverting the estimated Hessian can be unreliable. As an alternative strategy, we obtain standard errors by combining an R function for direct maximisation of the likelihood with replicated Monte Carlo simulations of the fitted model.

Results

We have carried out a simulation study to compare the estimators yielded by the two methodologies and to asses the performance of the Hessian and Monte Carlo methods for calculating approximate standard errors. As expected, parameter estimates obtained from the two methods are almost identical. However, standard errors for the non-linear parameters (α_j, β_j) are estimated more reliably by Monte Carlo than by inversion of the estimated Hessian.

REFERECES:

1. Diggle, P., Elliott, P., Morris, S., Shaddick, G., 1997. Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society, Series A* 160, 491:505.
2. Diggle, P., Rowlingson, B., 1994. A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A* 157 (3), 433:440.
3. Lindsey, J. K. *Nonlinear Models in Medical Statistics.* Oxford University Press (2001)

Indicators of Least Absolute Deviation's sensibility

* Soumaya REKAIA

Abstract

This paper aims to propose indicators which make it possible to control the sensitivity of the robust method least absolute deviation (LAD), in the presence of single observation. Indeed, recent studies noted that this estimator has gained a relatively little favour in the robustness since data comprise outlier raised in X. Relaying on the sensitivity curve, we focus our study on the measure the sensitivity of LAD by the bias of estimate and by the importance of the leverage in this skew, expressed by its contribution to the total inertia of the sample. Simulations of Monte Carlo enabled us to retain two models: a sigmoid model and a model with threshold. The results of the estimates show that Least Absolute Deviation is sensitive to the points whose contribution is higher than 80%. We also verify this résulte for real data sets.

Mots clés : Outliers, estimation robuste, singularité, sensibilité, LAD.

MCS Code : 62G32, 62G35, 62F35, 93B51.

JEL Code : B23, C13, C15.

* Rekaia Soumaya :
Université de Panthéon-Assas Paris 2
ERMES (UMR 7181-CNRS)
Mail: Rekaia.Soumaya@u-paris2.fr

Functional regression analysis using R

Christian Ritz

Department of Natural Sciences, Faculty of Life Sciences, University of Copenhagen, Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark
ritz@life.ku.dk

Functional data consist of observations which can be treated as functions rather than just numeric vectors. One example is fluorescence curves commonly used in photosynthesis research: The curve reflects the biological processes taking place in a plant in the first second of exposure to sunlight. As several (mostly unexplained) processes are involved the resulting curves can have several local minima and maxima and is not easily described using parametric models. Another example is repeated measurements over time on the same subject, frequently encountered in dietary and growth studies. The resulting curves may fluctuate as a consequence of daily patterns or seasonal trends.

Any variety of models for describing functional data exists. Functional regression models are particularly appealing as they strike a balance between flexible non-parametric modelling of the unknown average curve and semi-parametric modelling of effects due to explanatory variables in much the same way as for ordinary ANOVA models.

This presentation shows how to use **R** for estimation, hypothesis testing and graphical model checking of functional regression models of the form:

$$y(t) = \phi(z)\mu(t) + \epsilon(t)$$

with $y(t), \mu(t)$ and $\epsilon(t)$ denoting the functional observation, the average curve and the error process, respectively. The term $\phi(z)$ is a multiplicative effect modifying the average curve according to the explanatory variable z. The functions μ and ϕ are estimated nonparametrically in a two-step procedure. A quasi-likelihood or glm approach can be used to estimate the effects of the explanatory variables.

Item Response Theory Using the ltm Package

Dimitris Rizopoulos

Item Response Theory has been steadily evolving in the past few decades and is starting to become one of the standard tools in modern psychometrics. The R package ltm has been developed to fit various latent trait models useful for Item Response Theory analyses. In particular, for dichotomous data the Rasch, the Two-Parameter Logistic, and Birnbaum's Three-Parameter models have been implemented, whereas for polytomous data Semejima's Graded Response model is available. In this talk the capabilities of ltm will be illustrated using real data examples.

Patient teenagers?: A comparison of the sexual behavior of virginity pledgers and matched non-pledgers[*]

Janet Elise Rosenbaum, Ph.D., A.M.[†]

February 13, 2008

Objective: The US government spends over $200 million annually on abstinence-promotion programs, including virginity pledges, and measures abstinence program effectiveness as the proportion of participants who take a virginity pledge. Past research used non-robust regression methods. This paper examines whether adolescents who take virginity pledges are less sexually active than matched non-pledgers.

Patients and Methods: National Longitudinal Study of Adolescent Health respondents, nationally representative sample of middle and high school who, when surveyed in 1995 never had sex or taken virginity pledge, and over age 15 (n=3440). Adolescents reporting virginity pledge on the 1996 survey (n=289) were matched with non-pledgers (n=645) using exact and nearest-neighbor matching within propensity score calipers on factors including pre-pledge religiosity and attitudes towards sex and birth control. Pledgers and matched non-pledgers were compared five years post-pledge on self-reported sexual behaviors and positive test result for *C. trachomatis*, *N. gonorrhoeae*, and *T. vaginalis*; and safe sex outside of marriage by use of birth control and condoms in past year and at last sex.

Results: Five years post-pledge, 84% of pledgers denied having ever pledged. Pledgers and matched non-pledgers did not differ in premarital sex, STDs, anal, and oral sex. Pledgers had 0.1 fewer past year partners, but the same number of lifetime sexual partners and age of first sex. Pledgers were 10 percentage-points less likely than matched non-pledgers to use condoms in the last year, and also less likely to use birth control in the past year and at last sex.

Conclusions: Virginity pledgers and closely-matched non-pledgers have virtually identical sexual behavior, but pledgers are less likely to protect themselves from pregnancy and disease before marriage than matched non-pledgers. Abstinence programs may not affect sexual behavior, but may increase unsafe sex. Federal abstinence education funds should be directed to programs which teach birth control, and do so accurately. Virginity pledges should not be used as a measure of abstinence program effectiveness.

[*] This work was supported by the Milton Fund of Harvard Medical School (Michael Ganz, Principal Investigator), the Harvard Graduate School of Arts and Sciences, the Harvard PhD program in Health Policy, the Department of Society, Human Development and Health at the Harvard School of Public Health, and the Institute for Health Research and Policy at the University of Illinois at Chicago.

[†] Harvard University; Institute for Health Research and Policy, University of Illinois at Chicago

distrMod — an S4-class based package for statistical models

P. Ruckdeschel[1] and M. Kohl[2]

[1] Fraunhofer Institut für Techno– und Wirtschaftsmathematik ITWM, Abteilung Finanzmathematik, Fraunhofer-Platz 1, D-67663 Kaiserslautern
[2] Universität Bayreuth, Mathematisches Intstitut, D-95440 Bayreuth

The S4 concept ([1]) is a strong tool for writing unified algorithms. As an example for this in R ([3]), we present a new package distrMod for a conceptual implementation of statistical models based on these S4-classes. It is part of the distrXXX-family of packages ([4]), which is available on CRAN for quite a while, and which is developed under the infrastructure of R-Forge ([6]) in project distr ([5]).

The infrastructure to package distrMod is laid in packages distr and distrEx.

In package distr, we introduce S4 classes for distributions with slots for a parameter and for functions r, d, p, and q corresponding to functions like rnorm, dnorm, pnorm and qnorm.
We have made available quite general arithmetical operations to our distribution objects, generating new image distributions automatically, including affine transformations, standard mathematical univariate transformations like sin, abs, and convolution.

Package distrEx provides additional features like evaluation of certain functionals on distributions like expectation, variance, median, and also distances between distributions like total variation-, Hellinger-, Kolmogorov-, and Cramér-von-Mises-distance. Also, (factorized) conditional distributions and expectations are implemented.

Package distrMod then implements parametric resp. L_2 differentiable models, introducing S4-classes ParamFamily and L2ParamFamily. Based on these, quite general "Minimum Criterium"-estimators such as Maximum-Likelihood- and Minimum-Distance-Estimators are implemented.

This implementation goes beyond the scope of fitdistr from MASS ([7]), as we may work with distribution objects themselves and have available quite general expectation operators...
In short, we are able to implement **one** static algorithm which by tt S4 method dispatch may take care dynamically about various models, thus avoiding redundancy and simplifying maintenance.

This approach is also taken up to implement optimally robust estimation in the infinitesimal setup of ([4]) and its refinements in ([2]); this will be the topic of a contribution to this conference by the second author.

References

[1]. J. Chambers (1998). Programming with data: a guide to the S language. Springer.

[2]. M. Kohl (2005). *Numerical contributions to the asymptotic theory of robustness.* Dissertation, Universität Bayreuth, Bayreuth.

[3]. R Development Core Team (2008). R: A language and environment for statistical computing.
http://www.r-project.org.

[4]. H. Rieder (1994). *Robust asymptotic statistics*. Springer.

[5]. P. Ruckdeschel, M. Kohl, T. Stabla, and F. Camphausen (2006). S4 Classes for Distributions. *R-News*, **6**(2): 10–13. http://CRAN.R-project.org/doc/Rnews/Rnews_2006-2.pdf.
Also available in extended and updated version as vignette package distrDoc on CRAN

[6]. P. Ruckdeschel, M. Kohl, T. Stabla, F. Camphausen, E. Feist, and K. Owzar (2008). Project distr. Project page on R-Forge, http://r-forge.r-project.org/projects/distr/

[7]. S. Theussl (2007) R-Forge User's Manual. http://r-forge.r-project.org/R-Forge_Manual.pdf

[8]. W. N. Venables and B. D. Ripley (2002). *Modern Applied Statistics with S.* Fourth edition. Springer.

Analysis of CGH arrays using MCMC with Reversible Jump: detecting gains and losses of DNA and common regions of alteration among subjects

Oscar M. Rueda[1], Ramon Diaz-Uriarte[1]

[1]Statistical Computing Team, Structural and Computational Biology Programme, Spanish National Cancer Center (CNIO), Melchor Fernández Almagro 3,
Madrid, 28029, Spain

Abstract

Copy number variation (CNV) in genomic DNA is linked to a variety of human diseases (including cancer, HIV acquisition and progression, autoimmune diseases, and neurodegenerative diseases), and array-based CGH (aCGH) is currently the main technology to locate CNVs. To be immediately useful in both clinical and basic research scenarios, aCGH data analysis requires accurate methods that do not impose unrealistic biological assumptions and that provide direct answers to the key question “What is the probability that this gene/region has CNAs?”. Current approaches fail, however, to meet these requirements.

We have developed RJaCGH, a method for identifying CNAs from aCGH. We use a non-homogeneous Hidden Markov Model fitted via Reversible Jump Markov Chain Monte Carlo, and we incorporate model uncertainty through Bayesian Model Averaging. RJaCGH provides an estimate of the probability that a gene/region has CNAs while incorporating inter-probe distance. Using Reversible Jump we do not need to fix in advance the number of hidden states, nor do we need to use AIC or BIC for model selection. We presented a first version of our model at UseR two years ago. Since then, we have explored different approaches to improve convergence and speed-up computations, including usage of Gibbs sampling vs. Metropolis-Hastings, delayed rejection, and coupled parallel chains.

Based on the output from RJaCGH, we have also developed two probabilistically-based methods for the indentification of regions of alteration that are common among samples. Our methods are unique and qualitatively different from existing approaches, not only because of the use of probabilities, but also because they incorporate both within- and among-array variability and can detect small subgroups of samples with respect to common alterations. The two methods emphasize different features of the recurrence (sample heterogeneity, minimal required evidence for calling a common region) and, thus, will be instrumental in the current efforts to standardize definitions of recurrent or common CNV regions, cluster samples with respect to patterns of CNV, and ultimately in the search for genomic regions harboring disease-critical genes.

We will discuss the statistical features of our models, as well as the implementation of RJaCGH, including the combined usage of R and C, and different approaches for improving speed and decrease memory consumption.

CXXR: Refactoring the R Interpreter into C++

Andrew R. Runnalls,
University of Kent, UK [*]

25 March 2008

CXXR (www.cs.kent.ac.uk/projects/cxxr) is a project to refactor (reengineer) the interpreter of the R language, currently written for the most part in C, into C++, whilst as far as possible retaining full functionality. It is hoped that by reorganising the code along object-oriented lines, by deploying the tighter code encapsulation that is possible in C++, and by improving the internal documentation, the project will make it easier for researchers to develop experimental versions of the R interpreter. The author's own medium-term objective is to create a variant of R with built-in facilities for provenance tracking, so that for any R data object it will be possible to determine exactly which original data files it was derived from, and exactly which sequence of operations was used to produce it. (In other words, an enhanced version of the old S AUDIT facility.)

At the time of this abstract:

- Memory allocation and garbage collection have now been decoupled from each other and from R-specific functionality, and encapsulated within C++ classes. Classes `CellPool`, `MemoryBank` and `Allocator` look after memory allocation; `GCManager`, `GCNode`, `GCRoot` and `WeakRef` look after garbage collection. (All CXXR classes are within the namespace `CXXR`.) Class `GCRoot` provides C++ programmers with a mechanism for protecting objects from the garbage collector, as a more user-friendly (and probably less error-prone) alternative to the PROTECT/UNPROTECT mechanism used in standard R

- The `SEXPREC` union of CR is being progressively converted into an extensible hierarchy of classes rooted at a class `RObject` (which inherits from `GCNode`). This has already happened for vector objects and CONS-cell type objects, and it is now straightforward to introduced new types of R object simply by inheriting from `RObject`.

The proposed paper will:

1. Describe the motivation behind CXXR;

2. Report on progress to date;

3. Illustrate some of the simplified coding practices that CXXR enables;

4. Describe the measures taken to keep CXXR in synch with successive releases of standard R;

5. Outline future plans.

The paper will assume some familiarity with C programming and with concepts of object-oriented programming (e.g. in R or in Java), but C++-specific concepts will be explained as required.

[*] Computing Laboratory, The University, Canterbury CT2 7NF. Email: `A.R.Runnalls@kent.ac.uk`

‘robande’: An R package for Robust ANOVA

Majid Sarmad* and Peter Craig†

The R package ‘robande’ will be described and demonstrated. The package is based on the methodology in Sarmad (2006)[1] which generalises the ideas introduced by Seheult and Tukey (2001)[2] for performing a robust analysis of variance for a factorial experimental design, those ideas being based on earlier work by Tukey and collaborators on median polish for two way tables. The method may be applied to any type of factorial design including full and fractional factorial designs with and without replication. A version of sequential ANOVA is proposed for non-orthogonal designs. The package includes functions to decompose the data using a specified sweep function, to present the resulting decomposition, to detect and highlight possible outliers and to compute the robust ANOVA table.

*sarmad@um.ac.ir

†p.s.craig@dur.ac.uk

[1] M. Sarmad. Robust data analysis for factorial experimental designs: Improved methods and software. PhD thesis, University of Durham, 2006.

[2] A. H. Seheult and J. W. Tukey. Toward robust analysis of variances. In *Data Analysis from Statistical Foundations: A Festschrift in Honour of the 75th Birthday of D.A.S.Fraser.* Ottawa, 2001, Nova Publishers.

Spatial Analysis and Visualization of Climate Data Using R

David Sathiaraj

NOAA Southern Regional Climate Center

Louisiana State University, USA

March 31, 2008

Abstract

R's spatial libraries and its efficient data handling abilities make it a very effective tool for spatial analysis and visualization of climate data. At the NOAA Southern Regional Climate Center, R is being used for the spatial analysis and visualization of climate data. This talk will demonstrate how climate maps are generated using the spatial and interpolation packages in R. The talk will also outline how R serves as an effective GIS tool in the development of climate data driven map layers. R-generated maps that visualize a number of climatological elements will be demonstrated. Techniques used in developing the maps will be discussed.

`RLRsim`: Testing for Random Effects or Nonparametric Regression Functions in Additive Mixed Models

Fabian Scheipl[1], Sonja Greven[2], and Helmut Küchenhoff[1]

[1] Institut für Statistik, Ludwig-Maximilians-Universität München, Germany
[2] Department of Biostatistics, Johns-Hopkins University, USA

Abstract. Testing for a zero random effects variance is an important and common testing problem. Special cases include testing for a random intercept, and testing for polynomial regression versus a general smooth alternative based on penalized splines. The problem is non-regular, however, due to the tested parameter on the boundary of the parameter space. Our package `RLRsim` uses the approximate null distribution for the Restricted Likelihood Ratio Test proposed in Greven et al. (2008) to provide a rapid, powerful and reliable test for this problem. This method extends the exact distribution derived for models with one random effect (Crainiceanu & Ruppert, 2004) to obtain a good approximation for models with several random effects. The test performed better than a number of competitors in an extensive simulation study covering a variety of typical settings (Scheipl et al. , 2008). `RLRsim` also proved to be an equivalent and fast alternative to computationally intensive parametric bootstrap procedures. Our package can be used in a variety of settings, providing convenient wrapper functions to test terms in models fitted using `nlme::lme`, `lme4::lmer`, `mgcv::gamm` or `SemiPar::spm`.

References

CRAINICEANU C, RUPPERT D (2004). Likelihood ratio tests in linear mixed models with one variance component., *JRSS-B, 66, 1, 165–185.*

GREVEN S, CRAINICEANU C, KÜCHENHOFF H, PETERS A (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *JCGS, to appear.*

SCHEIPL F, GREVEN S, KÜCHENHOFF H (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *CSDA, 52, 7, 3283–3299*

Keywords

Linear Mixed Model; Non-regular Problem; Penalized Splines; Restricted Likelihood Ratio Test; Variance Component

`robfilter` : An R Package for Robust Time Series Filters

Karen Schettlinger, Roland Fried and Ursula Gather

Fakultät Statistik, Technische Universität Dortmund
44221 Dortmund, Germany

`robfilter` is a package of R functions for robust extraction of an underlying signal from a time series. Assuming a standard signal plus noise model for the series, the general idea is to approximate the signal in a moving time window by a local parametric model like a locally constant level, i.e. *location-based methods* (Fried, Bernholt, Gather, 2006), or a local linear trend, so-called *regression-based methods* (Davies, Fried, Gather, 2004; Fried, Einbeck, Gather, 2007).

We present several filters which differ with respect to the signal characteristics and the outlier patterns they can deal with (Schettlinger, Fried, Gather, 2006). In particular, some filters are especially designed to preserve sudden shifts and local extremes (turning points) even if patches of subsequent outliers may occur (Fried, 2004). Furthermore, most of the filters are available both for retrospective filtering as well as online filtering without time delay. Estimation of the signal in the centre of a time window generally leads to better signal approximations. This approach is reasonable for retrospective data analysis, since the estimation always takes place with a time delay of half a window width. The proposed online filters estimate the signal value at the end of each time window without time delay, but the resulting signal estimates have a larger variability than their retrospective counterparts.

We present filters which are applicable to time series containing outliers, trends, trend changes or shifts in the signal level and give recommendations which filter is suitable for which data structure.

References

Davies, P.L., Fried, R., Gather, U. (2004) Robust Signal Extraction for On-line Monitoring Data. *J. Stat. Plann. Inference* **122**, 65-78.

Fried, R. (2004) Robust Filtering of Time Series with Trends. *J. Nonparametr. Stat.* **16**, 313-328.

Fried, R., Bernholt, T., Gather, U. (2006) Repeated Median and Hybrid Filters. *Comput. Stat. Data An.* **50**, 2313-2338.

Fried, R., Einbeck, J., Gather, U. (2007) Weighted Repeated Median Smoothing and Filtering. *J. Am. Stat. Assoc.*. **102**, 1300-1308.

Fried, R., Schettlinger, K. (2008) `robfilter` : Robust Time Series Filters. R package version 1.0, `http://cran.r-project.org/web/packages/robfilter/`.

Schettlinger, K., Fried, R., Gather, U. (2006) Robust Filters for Intensive Care Monitoring: Beyond the Running Median, *Biomedizinische Technik* **51** (2), 49-56.

Local Classification Methods for Heterogeneous Classes

Julia Schiffner and Claus Weihs

Technische Universität Dortmund, Germany

Abstract. Many classification methods, for example LDA, QDA or *Fisher discriminant analysis* (FDA), assume the classes to form homogeneous groups. But in practical applications heterogeneous classes that consist of multiple subclasses can often be observed. In such cases *local* classification methods that take the local class structure, i. e. the subclasses, into account can be beneficial. In package `klaR` (Weihs et al., 2005) the function `loclda` that performs *localized linear discriminant analysis* (Czogiel et al., 2007) is already available. Now, three more local classification methods are added.

The first two methods, the *common components classifier* and the *hierarchical mixture classifier* (Titsias and Likas, 2002), rely on modeling the class conditional densities by means of gaussian mixtures. The third method, *local Fisher discriminant analysis* (LFDA), was proposed by Sugiyama (2007). FDA seeks for a projection of the data into a subspace such that the between-class scatter is maximized and the within-class scatter is minimized. In LFDA the projection additionally has to fulfill the condition that nearby data points in the same class are kept close to each other and thus the local class structure is preserved.

The three local methods and their implementations in `R` are presented and their usefulness is demonstrated in several examples.

References

I. Czogiel, K. Luebke, M. Zentgraf, and C. Weihs. Localized linear discriminant analysis. In R. Decker and H.-J. Lenz, editors, *Advances in Data Analysis*, pages 133–140, Heidelberg, 2007. Springer.

M. Sugiyama. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *Journal of Machine Learning Research*, (8):1027–1061, May 2007.

M. K. Titsias and A. C. Likas. Mixture of experts classification using a hierarchical mixture model. *Neural Computation*, 14:2221–2244, 2002.

C. Weihs, U. Ligges, K. Luebke, and N. Raabe. klaR analyzing german business cycles. In D. Baier, R. Becker, and L. Schmidt-Thieme, editors, *Data Analysis and Decision Support*, pages 335–343. Springer, Berlin, 2005.

Parallelized preprocessing algorithms for high-density oligonucleotide array data

M. Schmidberger and U. Mansmann

Chair of Biometrics and Bioinformatics, IBE, University of Munich, Germany
`schmidb@ibe.med.uni-muenchen.de`

Abstract. Studies of gene expression using high-density oligonucleotide microarrays have become standard in a variety of biological contexts. The data recorded using the microarray technique are characterized by high levels of noise and bias. These failures have to be removed, therefore preprocessing of raw-data has been a research topic of high priority over the past few years.

Actual research and computations are limited by the available computer hardware. For many researchers the available main memory limits the number of arrays that may be processed. Furthermore most of the existing preprocessing methods are very time consuming and therefore not useful for first and fast checks in laboratories. To solve these problems, the potential of parallel computing should be used. In microarray technologies and statistical computing parallel computing does not appear to have been used extensively. For parallelization on multicomputers, message passing (MPI) methods and the R language will be used.

Ideas for parallelization of VSN and FARMS as well as a large project in applied bioinformatics ($>$ 5000 microarrays) will be discussed. Furthermore this presentation proposes the new BioConductor package `affyPara` for parallelized preprocessing of high-density oligonucleotide microarray data. Partition of data could be done on arrays and therefore parallelization of algorithms gets intuitive possible. In view of machine accuracy, the same results as serialized methods will be achieved. The partition of data and distribution to several nodes solves the main memory problems and accelerates the methods by up to the factor ten.

References

R. Gentleman, et all. (2005): Bioinformatics and Computational Biology Solutions Using R and Bioconductor. *Springer, Statistics for Biology and Health*

A. Rossini (2003): Simple Parallel Statistical Computing in R. *UW Biostatistics Working Paper Series, 193*

H. Sevcikova (2003): Statistical Simulations on Parallel Computers. *Journal of Computational and Graphical Statistics, 13, pp. 886-906*

R. A. Irizarry, et all. (2004): Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4, Apr, Nr. 2, 249264*

M. Schmidberger, U. Mansmann (2008): Parallelized preprocessing algorithms for high-density oligonucleotide array data. *22th International Parallel and Distributed Processing Symposium (IPDPS 2008), Proceedings, 14-18 April 2008, Miami, Florida, USA. IEEE 2008 (in press)*

MORET - A Software For Model Management

Ralf Seger and Antony Unwin

Institut für Mathematik, Rechnerorientierte Statistik und Datenanalyse

Universität Augsburg

RalfSeger@googlemail.com,Antony.Unwin@Math.Uni-Augsburg.DE

ABSTRACT

Administrating sets of models is a cumbersome task, since the number of models which can be fit can be very large. The project MORET[1] is designed to facilitate this task. The first version introduced at useR!2006 was capable of handling *lm*, *glm*, *gam* and *rpart* models. There are many other model commands and new ones are continually being developed, so a more dynamic concept is needed. MORET now provides the user with a graphical interface that allows the management of previously unknown models.

The purpose of this presentation is to demonstrate how to incorporate new model commands and to describe other additions to MORET.

References

[1] Seger R. *MORET*, http://www.rosuda.org/Moret/main.html

MSToolkit: Distributed R for the creation and analysis of simulated clinical trial data

Mike K Smith – *Pfizer*
Richard Pugh, Romain Francois – *Mango Solutions*

useR! 2008. Dortmund.

Abstract

A key tool in the area of pharmaceutical research is simulation-based modeling. This requires the generation and analysis of clinical trial datasets, which can be both complex and computationally intensive.

Pfizer and Mango Solutions collaborated on the development of an R package which allows the generation and analysis of simulated clinical trial data. In order to leverage existing IT infrastructure and programming skillsets, the package was integrated closely with the internal Pfizer Linux Grid cluster. The package was designed to allow SAS to be called as well as R in order to perform the required analyses.

This presentation will discuss the design and implementation of the `MSToolkit` package, before giving a demonstration of its use.

TIMPGUI: A graphical user interface for the package **TIMP**

Joris J. Snellenburg, Katharine M. Mullen, Ivo H. M. van Stokkum
Department of Physics and Astronomy
Vrije Universiteit Amsterdam
E-mail: {jsnel|kate|ivo}@few.vu.nl

March 31, 2008

The package **TIMP** is in use by biophysicists who seek to discover models for (photo)-physical processes in complex systems. The measurements under consideration most often represent some spectroscopic property resolved with respect to time, and the goal is typically to discover a nonlinear model for the kinetics. This problem is approached by postulating an initial model, in which the spectra associated with the system are obtained as conditionally linear parameters, then optimizing the nonlinear parameters and finally validating the resulting model for physical interpretability.

We have been motivated to use Java to develop an interface to **TIMP** for several reasons. One reason is that many of the scientists using **TIMP** prefer a graphical user interface (GUI) to a command line interface. Another reason is that Java, and the JFreeChart plotting library we are using, along with the JRI library (part of the **rJava** package), allows for more possibilities for interacting with plots than is currently possible in R alone. This facilitates interactive data exploration, which can greatly improve the rate at which models can be formulated and tested. A third reason to use Java is that it allows the GUI to be programmed with a GUI builder (we use the Netbeans Integrated Development Environment (IDE)) as opposed to manually specifying the parameters of widgets in R code. We feel this allows for a flexible modular design which is easily extended by other developers. Finally, we require a fully crossplatform interface, for which Java is well-suited.

Here we showcase the current capabilities of the interface and demonstrate its usability by demonstrating several case studies, fitting kinetic models to time-resolved fluorescence and absorption data.

SQLiteMap: package to manage vector graphical maps using SQLite

Norbert Solymosi[1], Andrea Harnos[1,2], Jenő Reiczigel[1,2]

[1]Adaptation to Climate Change Research Group, Hungarian Academy of Science, Budapest, Hungary

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

Some server based database management systems implemented the OpenGIS ”Simple Features Specification for SQL”.[1] The OpenGIS specification defines two standard ways of expressing spatial features: the Well-Known Text (WKT) form and the Well-Known Binary (WKB) form. Both WKT and WKB include information about the type of the feature and the coordinates which form the feature.[2] These systems (e.g. PostgreSQL-PostGIS, MySQL, ORACLE, MSSQL) allow to store the topological features and the descriptive data in the same database. This makes it possible to connect the spatial and descriptive tables without any interface and to access the spatial data by a large number of users in a secure way.

But these systems assume that the user needs permission to a running service or to install a server to use the spatial data. In some cases, it is useful if the user can use the database stored maps on different computers and platforms. The SQLite is a good choice for a portable database, it is platform-idendependent and there are some R packages to manage SQLite databases. Unfortunately, it has no spatial extension, but there is an SQLite extension for the SharpMap library.[3]

Following the idea of this solution we developed a package that may help the user read and write spatial features from and to an SQLite database. Each table with geometry field is treated as a layer. The tables contain the topological features (polygon, linestring, point etc.) in one geometry field in WKT form.

[1] http://www.opengeospatial.org/standards

[2] http://postgis.refractions.net/

[3] http://www.codeplex.com/SharpMap

Approximate Conditional-mean Type Filtering for State-space Models

B. Spangl[1], P. Ruckdeschel[2] and R. Dutter[3]

[1] University of Natural Resources and Applied Life Sciences,
Institute of Applied Statistics and Computing, A – 1180 Vienna
[2] Fraunhofer ITWM, Department for Financial Mathematics, D – 67663 Kaiserslautern
[3] Vienna University of Technology, Deparment of Statistics and Probability Theory, A – 1040 Vienna

Keywords: Robust Kalman filtering, ACM type filter, rLS filter, robKalman.

Abstract

We consider in the following the problem of recursive filtering in linear state-space models. The classically optimal Kalman filter (Kalman, 1960; Kalman and Bucy, 1961) is well known to be prone to outliers, so robustness is an issue.

For an implementation in R (R Development Core Team, 2005), the first two authors have been working on an R package robKalman (Ruckdeschel and Spangl, 2007), where a general infrastructure is provided for robust recursive filters. In this framework the rLS (Ruckdeschel, 2001) and the ACM (Martin, 1979) filter have already been implemented, the latter as an equivalent realization of the filter implemented in Splus.

While this ACM filter is bound to the univariate setting, based on Masreliez's result (Masreliez, 1975) the first and the third author propose a generalized ACM type filter for multivariate observations (Spangl and Dutter, 2008).

This new filter is implemented in R within the robKalman package and has been compared to the rLS filter by extensive simulations.

References

R.E. Kalman (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering—Transactions of the ASME*, 82, p. 35–45.

R.E. Kalman and R. Bucy (1961). New results in filtering and prediction theory. *Journal of Basic Engineering—Transactions of the ASME*, 83, p. 95–108.

R.D. Martin (1979). Approximate conditional-mean type smoothers and interpolators. In *Smoothing Techniques for Curve Estimation.* Lect. Notes Math. 757, p. 117–143, Springer, Berlin.

C.J. Masreliez (1975). Approximate non-Gaussian filtering with linear state and observation relations. *IEEE Transactions on Automatic Control*, 20, p. 107–110.

R Development Core Team (2005). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna.

P. Ruckdeschel (2001). *Ansätze zur Robustifizierung des Kalman-Filters.* Bayreuther Mathematische Schriften, Vol. 64.

P. Ruckdeschel and B. Spangl (2007). robKalman: *An R package for robust Kalman filtering.* Web: http://r-forge.r-project.org/projects/robkalman/.

B. Spangl and R. Dutter (2008). *Approximate Conditional-mean Type Filtering for Vector-valued Observations.* Technical Report TR-AS-08-1, Universität für Bodenkultur, Vienna.

Regression Model Development and Yet Another Regression Function

Dr. Werner Stahel
Seminar für Statistik, ETH Zürich

31st March 2008

A strategy to develop a regression model involves many steps and decisions which are based on pertinent numeric tables and thorough analysis of residual plots. With the standard regression functions available in R, such an assessment consists of several function calls and informed settings of their arguments, depending also on the type of target variable (continuous, count, binary, multinomial, multivariate, ...). The examination of a logistic regression fit, e.g., involves calling glm, summary, drop1, influence, plot, and termplot, and selecting the useful information from what is obtained from them.

This contribution presents a user oriented function that sets the sensible choices for the different models. It produces an object which gives the useful information for judging the model fit when printed and plotted.

More specifically, the function accepts the same arguments as `lm` or `glm`, and some more. It also accepts ordered, multinomial, and multivariate responses. Of course, calculations are done by calling the available fitting functions.

The function stores results that are produced by the fitting function and by calling `summary` on the object, as well as some additional ones, like the leverage values. If printed, it gives a table that contains, for continuous or binary explanatory variables, the coefficients, their P-values, the collinearity measure R_j^2 and a new measure of significance that additionally characterizes the confidence interval. For factors, the P-value is given, since individual coefficients and their P-values are of limited information content. The coefficients of factor levels are reproduced separately. – The last part of the print output is very similar to the usual summary part of printing the `summary`, but includes, in the case of a `glm`, an overdispersion test if applicable.

The strength of the new function lies in its plotting method. All residual plots use a plotting scale that is not affected by outliers in the residuals, but outliers are still shown in a marginal region of the plot. Most plots are complemented by a smooth by default. In order to judge the significance of any curvature shown by this line, 19 such lines are simulated from random data corresponding to the model. Reference lines indicate contours of equal response values and help to identify suitable transformations of the explanatory variables.

In summary, the function `regr` and its printing and plotting methods have made my life much easier when developing regression models and have lead to higher quality of analyses obtained by students.

A pipeline based on multivariate correspondence analysis with supplementary variables for cancer genomics

Christine Steinhoff[1,*], Matteo Pardo[2] and Martin Vingron[1]

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestr 63-73, 14195 Berlin, Germany

[2] SENSOR Laboratory, CNR-INFM, Via Valotti 9, 25133 Brescia, Italy

***** christine.steinhoff@molgen.mpg.de

The development of several high throughput gene profiling methods, such as comparative genomic hybridization (CGH) and gene expression microarrays enables for studying specific disease patterns in parallel. The underlying assumption for studying both genomic aberrations and gene expression is that genomic aberration might effect gene expression either directly or indirectly. In cancer research, in particular, there have been a number of attempts to improve cancer subtype classification or study the relationship between chromosomal region and expression aberrations.

The intuitive way to analyze different data sources is separately and consecutively, e.g. first determine regions with copy number aberrations (possibly tissue or patients -specific) and then look for differentially expressed (onco)genes inside these regions [1]. There is a natural reason for integrating results rather than data: strong heterogeneity does not allow sensible alignments of the source data. Still, integrative approaches –where data are fused before their analysis- are preferable. Only recently, few integrative methods have been published [2]. Nevertheless, these approaches do not integrate covariate data like tumor grading, mutation status and other disease features. These features are frequently available and of interest for an integrative analysis.

We address these two problems, namely jointly analyzing different data sources and integrating supplementary categorical data. Furthermore, our approach can easily be applied to diverse data sources, even more than two, with and without supplementary patients' information.

We established a new data analysis pipeline for the joint visualization of microarray expression and arrayCGH data (aCGH), and the corresponding categorical patients' information. All computational analysis steps are programmed using R and Bioconductor. The pipeline comprises four parts: (a) data discretization, (b) binary mapping, (c) gene filtering, (d) multiple correspondence analysis. The first two steps transform data to a common binary format, a necessary step for jointly analyzing them. Filtering removes noise and redundancy by reducing the number of features (genes). We considered variance filtering, expression-aCGH correlation filtering and PCA loading on the first two principal components. In the last pipeline step, we apply a method based on correspondence analysis, namely multivariate correspondence analysis with supplementary variables (MCASV) [3]. MCASV has been applied in the context of social sciences but to our knowledge has not been used in the context of biological high throughput data analysis. Features (expression and aCGH) and covariates (patients' information) are transformed into a common space. Vicinity between features and covariates can then be visualized and quantified. We e.g. determine genes that are correlated with covariates, possibly for interesting subsets of patients. In MCASV vicinity is measured by the angle intercurring between covariate and feature.

We applied our approach to a published dataset on breast cancer. Pollack et al. [4] studied genomic DNA copy number alterations and mRNA levels in primary human breast tumors. We were able to retrieve candidate genes that show strong association with grade 3 tumors and p53 mutant status. Candidate genes display significant enrichment of cancer related GO terms. Moreover there are interesting differences between genes selected starting from aCGH and expression data alone and genes selected by integrating the datasets.

1. Jacobs, S. et al. Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays. *Cancer Res* **67**, 2544-2551 (2007).
2. Berger, J.A., Hautaniemi, S., Mitra, S.K. & Astola, J. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Trans Comput Biol Bioinform* **3**, 2-16 (2006).
3. Nenadic, O. & Greenacre, M. Multiple Correspondence Analysis and Related Methods. (Chapman & Hall/CRC, London; 2006).
4. Pollack, J.R. et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* **99**, 12963-12968 (2002).

Why and how to use random forest variable importance measures (and how you shouldn't)

Carolin Strobl
Ludwig-Maximilians-Universität
München

Achim Zeileis
Wirtschaftsuniversität
Wien

Abstract

Random forests are becoming increasingly popular in many scientific fields, especially in genetics and bioinformatics, for assessing the importance of predictor variables in high dimensional settings. Advantages of random forests in these areas are that they can cope with "small n large p" problems, complex interactions and even highly correlated predictor variables. The talk gives a short introduction to the rationale of random forests and the their variable importance measures as well as the two random forest implementations offered in the R system for statistical computing: randomForest in the package of the same name by Breiman *et al.* (2006) and cforest in the package party by Hothorn *et al.* (2008). Moreover, recent research issues are addressed:

- Solutions are presented for bias in random forest variable importance measures towards, e.g., predictor variables with many categories (Strobl, Boulesteix, Zeileis, and Hothorn 2007) and correlated predictor variables (Archer and Kimes 2008).
- Currently suggested tests for random forest variable importance measures (Breiman and Cutler 2008; Rodenburg *et al.* 2008) are critically discussed in an outlook.

Keywords: Feature selection, screening, variable importance, Gini index, CART, bagging.

References

Archer KJ, Kimes RV (2008). "Empirical characterization of random forest variable importance measures." *Computational Statistics & Data Analysis*, **52**(4), 2249–2260.

Breiman L, Cutler A (2008). "Random Forests – Classification Manual (website accessed in 1/2008)." http://www.math.usu.edu/~adele/forests/.

Breiman L, Cutler A, Liaw A, Wiener M (2006). *Breiman and Cutler's Random Forests for Classification and Regression.* R package version 4.5-16, URL http://CRAN.R-project.org/package=randomForest.

Hothorn T, Hornik K, Zeileis A (2008). "party: A Laboratory for Recursive Part(y)itioning." R package version 0.9-96, URL http://CRAN.R-project.org/package=party.

Rodenburg W, Heidema AG, Boer JM, Bovee-Oudenhoven IM, Feskens EJ, Mariman EC, Keijer J (2008). "A Framework to Identify Physiological Responses in Microarray Based Gene Expression Studies: Selection and Interpretation of Biologically Relevant Genes." *Physiological Genomics*, **33**(1), 78–90.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007). "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics*, **8:25**.

R AnalyticFlow: A flowchart-style GUI for R

Ryota Suzuki [*]

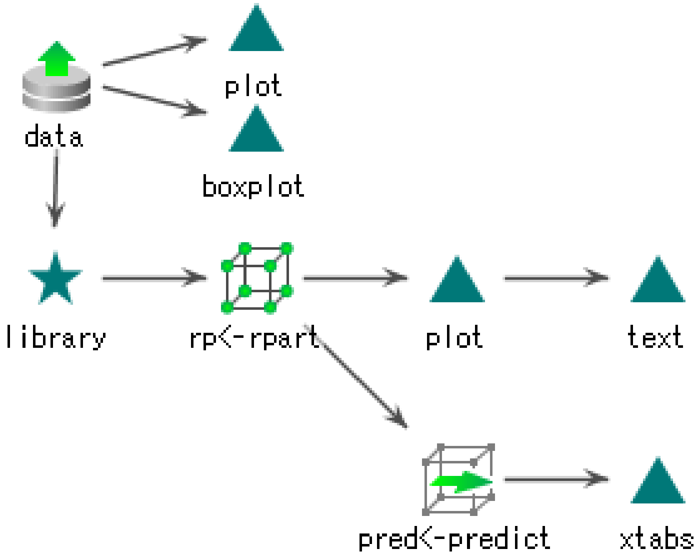
March 28, 2008

Abstract

R AnalyticFlow is a new flowchart-style GUI for R. A user draw an "analysis flow", which contains R expression nodes connected by directed edges. By "running" a series of nodes, the corresponding R expressions are executed by R engine.

There are two main advantages: (1) flowchart-style visualization helps us to overview the processes of data analysis, and (2) "branching" the processes enables flexible analysis strategies.

R AnalyticFlow is written in Java, with the help of several open-source Java libraries. Our source code is also open-sourced and available under the BSD license. It depends on Java ($\geq$ 5), R ($\geq$ 2.5.0), rJava and JavaGD. It currently runs on Windows, Linux and Mac OS X.



[*] Ef-prime, Inc. URL: `http://www.ef-prime.com/`

Some Aspects on Classification, Variable Selection and Categorical Clustering

Gero Szepannek, Uwe Ligges, and Claus Weihs

Department of Statistics, Technische Universität Dortmund, Germany

Abstract. The package `klaR` contains several utilities to handle classification problems, e.g. Friedman's RDA, an interface to `svmlight` (Joachims, 1999) as well as variable selection procedures like the `stepclass` algorithm or Wilk's Λ, a visualization tool for SOMs or several classification performance measures (see Weihs et al., 2006).

This poster presents recent extensions towards classification on minimal variable subspaces for multi class problems by performing class pair wise variable selection (see Szepannek and Weihs, 2006). Examples of situations are presented where this approach may be highly beneficial in terms of misclassification rates.

Furthermore, the k-modes algorithm (Huang, 1998) is implemented allowing to perform a k-means like clustering for categorical data.

Keywords

CLASSIFICATION, VARIABLE SELECTION, CLUSTERING, CATEGORICAL DATA, DATA MINING

References

Huang, Z. (1998): Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3), 283-304.

Joachims, T., (1999): Making large-Scale SVM learning practical, In: B. Schlkopf and C. Burges and A. Smola (ed.) (eds.): *Advances in Kernel Methods - Support Vector Learning*, MIT-Press.

Szepannek, G., Weihs, C. (2006): Variable selection for more than two classes where data are sparse. In: M.Spiliopolou, R.Kruse, C.Borgelt, A.Nürnberger and W.Gaul (eds): *From Data and Information Analysis to Knowledge Engineering*, Springer, 700-707.

Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005): klaR - analyzing German business cycles, In: D. Baier, R. Becker and L. Schmidt-Thieme (Eds.): *Data Analysis and Decision Support*, Springer, Berlin, 335-343.

rsm: an R package for Response Surface Methodology

Ewa M. Sztendur and Neil T. Diamond
Monash University
Melbourne, Australia
{ewa.sztendur,neil.diamond}@buseco.monash.edu.au

Introduction

rsm is an R package for Response Surface Methodology. For 1st order response surfaces rsm provides

- Calculation of the Path of Steepest Ascent
- Precision of the Path

For 2nd order response surfaces rsm provides

- Ridge Analysis
- Maximum or Minimum plots
- Canonical Analysis
- Precision of canonical analysis based on Double Linear Regression

1st Order Response Surfaces

The model is

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon_i$$

Path of Steepest Ascent

The path of steepest ascent is given by:

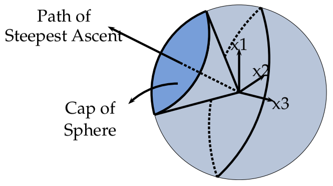
$$x_1 = \frac{rb_1}{\sqrt{\sum_{i=i}^{k} b_i^2}}, x_2 = \frac{rb_2}{\sqrt{\sum_{i=i}^{k} b_i^2}}, \ldots, x_k = \frac{rb_k}{\sqrt{\sum_{i=i}^{k} b_i^2}}$$

where $b_1, \ldots, b_k$ are the estimates of $\beta_1, \ldots, \beta_k$ and r is the Radius, the distance to the centre of the design region. The estimated response on the path is given by

$$\hat{y} = b_0 + r\sqrt{\sum_{i=i}^{k} b_i^2}$$

Precision of the Path of Steepest Ascent

Box (1955) and Box and Draper (1987, pp. 190-194) gave a method for computing a confidence cone for the direction of steepest ascent. The proportion of directions included in the confidence cone gives a measure of the precision of the path of steepest ascent, and is measured by taking the ratio of the surface area of the cap of the sphere within the confidence cone to the surface area of the sphere. See also Sztendur and Diamond (2002).



Implementation

rsm provides first order objects which include print, summary and plot methods:

`fit1 <- firstorder(X,y)` creates a firstorder fit object.

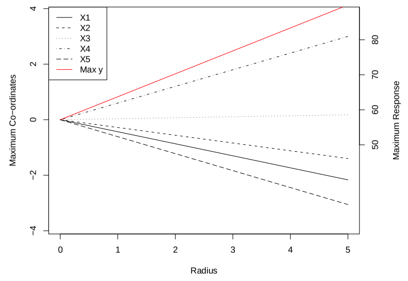
`print(fit1)` adds the path of steepest ascent:

```
      (Intercept)      X1      X2     X3     X4      X5
           57.175 -3.350 -2.162 0.275 4.638 -4.725
The path of steepest ascent is:

                         X1 = -0.433r
                         X2 = -0.280r
                         X3 =  0.036r
                         X4 =  0.598r
                         X5 = -0.611r

The estimated response on the path is ycap = 57.175 + 7.734r
```

`summary(fit1)` gives, in addition, the percentage of directions excluded from the 95% confidence cone.

```
The 95% confidence cone for the path of steepest ascent
excludes 99.03% of possible directions.
```

`plot(fit1)` gives the co-ordinates of the path of steepest ascent and the predicted response on the path:



Second Degree Response Surfaces

The equation is

$$\hat{y} = b_0 + \mathbf{x}^{\mathrm{T}}\mathbf{b} + \mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_{11} & \frac{1}{2}b_{12} & \cdots & \frac{1}{2}b_{1k} \\ \frac{1}{2}b_{12} & b_{22} & \cdots & \frac{1}{2}b_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}b_{1k} & \frac{1}{2}b_{2k} & \cdots & b_{kk} \end{bmatrix}$$

where $\mathbf{b}$ is the $(k \times 1)$ vector of the first-order regression coefficients and $\mathbf{B}$ is the $(k \times k)$ symmetric matrix whose diagonal elements are the pure quadratic coefficients and whose off-diagonal elements are one-half the mixed quadratic coefficients.

Ridge Analysis

Ridge analysis is equivalent to the path of steepest ascent applied to second order response surfaces and was developed by A.W. Hoerl (1959) and R.W. Hoerl(1985). In Ridge analysis, stationary points of the response surface subject to $\mathbf{x}^T\mathbf{x} = r^2$ are found, resulting in

$$\mathbf{x}_S = -\frac{1}{2}(\mathbf{B} - \mu\mathbf{I})^{-1}\mathbf{b}$$

for various values of μ. For maximisation of the response, only values of μ greater than the largest eigenvalue of $\mathbf{B}$ are used; while for minimisation of the response, only values of μ less then the smallest eigenvalue are used. A ridgeplot gives the dependence of the radius of the stationary values of the response against the value of the Lagrangian multipler, μ.

Canonical Analysis

Canonical analysis of the 2nd degree response surface allows the investigation of the underlying nature of the response surface and whether it is a maximum, minimum, saddle, rising ridge, or stationary ridge.

A Canonical Form

In the A Canonical Form, the axes are rotated so that the cross-product terms are removed, resulting in the model:

$$\hat{y} = b_0 + \mathbf{X}^{\mathrm{T}}\theta + \mathbf{X}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{X}$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$.

B Canonical Form

In the B Canonical Form, both cross-product and linear terms are removed by shifting the origin and rotating the axes, resulting in the model:

$$\hat{y} = \hat{y}_S + \tilde{\mathbf{X}}^T\mathbf{\Lambda}\tilde{\mathbf{X}}$$

The values of the λs show the nature of the surface. If all the λs are negative, the surface is a maximum; if all the λs are positive, the surface is a minimum; if the λs are of mixed sign, the surface is a saddle; while if some of the λs are zero, the surface is a stationary ridge. The latter is particularly important, as it indicates a linear or planar maximum or minimum, rather than a point maximum or minimum.

Double Linear Regression Method

In practice, because of experimental error and mild lack of fit, λs exactly equal to 0 will not occur. However, small λs indicate that the surface can be approximated by a ridge system. The standard errors of the λs are determined using the double linear regression method, due to Bisgaard and Ankenman (1996).

Implementation

rsm provides second order objects which include print, summary and plot methods:

`fit2 <- secondorder(X,y)` creates a secondorder fit object.

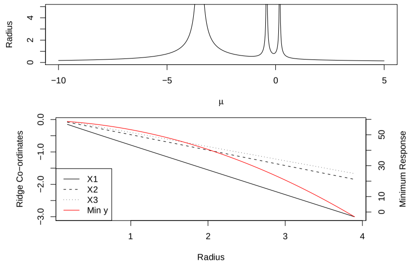
`print(fit2)` adds the A and B Canonical Forms:

```
     (Intercept)     x1     x2    x3   x1sq   x2sq   x3sq
          59.140  2.006  1.004 0.670 -1.999 -0.731 -0.998
            x1x2   x1x2   x2x3
          -2.801 -2.179 -1.154

A Canonical Form:
y=59.140+0.273X1-0.341X2+2.300X3+0.188X1sq-0.411X2sq-3.505X3sq

                X1 =  0.585x1 -0.797x2 -0.149x3
                X2 = -0.280x1 -0.371x2 +0.885x3
                X3 =  0.761x1 +0.476x2 +0.440x3

Location of Stationary Point: (-0.058, 0.888, -0.114)
Distance of Stationary Point from Origin:  0.898
B Canonical Form:
y=59.490+0.188XX1sq-0.411XX2sq-3.505XX3sq

   XX1 =  0.585(x1+0.058) -0.797(x2-0.888) -0.149(x3+0.114)
   XX2 = -0.280(x1+0.058) -0.371(x2-0.888) +0.885(x3+0.114)
   XX3 =  0.761(x1+0.058) +0.476(x2-0.888) +0.440(x3+0.114)
```

`summary(fit2)` adds the standard errors of the λs based on the double linear regression method:

```
                Estimate Std.  Error t value Pr(>|t|)
  (Intercept)  5.949e+01  8.307e-01  71.610 1.02e-13
        XX1sq  1.880e-01  1.188e-01   1.583    0.148
        XX2sq -4.114e-01  3.775e-01  -1.090    0.304
        XX3sq -3.505e+00  4.473e-01  -7.835 2.61e-05
```

`plot(fit2)` gives the ridge plot and the maximum or minimum plot.



References
Bisgaard, S. and Ankenman, B., (1996), *"Standard Errors for the Eigenvalues in Second-Order Response Surface Models,"* Technometrics, **38**, 238–246.
Box, G.E.P. (1955), *"Contribution to the discussion, Symposium on Interval Estimation,"* JRSS B, **16**, 211–212.
Box, G.E.P. and Draper, N.R., (1987), *Empirical Model-Building and Response Surfaces*, Wiley: New York.
Hoerl, A.E., (1959), *"Optimum Solution of Many Variables Equations,"* Chemical Engineering Progress, **55**, 69–78.
Hoerl, R.W., (1985), *"Ridge Analysis 25 years later,"* The American Statistician, **39**, 186–192.
Sztendur, E.M. and Diamond, N.T., (2002). *"Extension to confidence region calculations for the path of steepest ascent,"* Journal of Quality Technology, **34**, 288-295.

Collaborative Software Development Using R-Forge

Stefan Theußl and Achim Zeileis and Kurt Hornik

March 31, 2008

A key factor in open source software development is the rapid creation of solutions within an open, collaborative environment. The open source model had its major breakthrough with the increasing usage of the internet. Online communities successfully combined not only their programming effort but also their knowledge, work and even their social life.

The consequence was an increasing demand for centralized resources e.g., to manage projects or source code. The most famous of such platforms—the world's largest open source software development web site—is SourceForge.net.

For a decade, the R Development Core Team as well as many R package developeRs have been using development tools like Subversion (SVN) or Concurrent Versions System (CVS) for managing their source code. A central repository is hosted by ETH Zürich mainly for managing the development of the base R system. Now, the R-project wants to provide infrastructure for the entire R community.

R-Forge (`http://R-Forge.R-project.org`) is a set of tools based on the open source software GForge—a fork of the open source version of SourceForge.net. It aims to provide a platform for collaborative development of R packages, R related software or other projects which are somehow related to R. It offers source code management facilities through SVN and a wide variety of web-based services.

Furthermore, packages hosted on R-Forge are built daily for various operating systems, i.e., Linux, MacOSX and Windows. These package builds are downloadable from the project's website on R-Forge as well as installable directly in R via `install.packages()`.

In our talk we show how package developeRs can get started with R-Forge. In particular we show how people can register a project, use R-Forge's source code management facilities, provide their packages with R-Forge, host a project specific website, and finally submit a package to CRAN.

Multivariate Data Analysis in Microbial Ecology - New Skin for the old Ceremony

Jean Thioulouse

The molecular biology revolution has a particularly strong impact in microbial ecology, as molecular methods are now giving access to data that were previously impossible to obtain. Soil microbial ecology studies are a good example of this situation. Knowledge of soil bacterial diversity is of great interest, both from an applied agronomical perspective, and in the framework of theoretical ecological models like the species-area relationship. Previously, only culturable species could be studied, which represented an extremely low part of the total bacterial community. Today, tools like DNA fingerprints, DNA microarrays, and transcriptomic methods can be used directly on DNA extracts from bulk soil samples, providing new insights into the diversity and functioning of bacterial soil communities.

However, while molecular tools have been rapidly apropriated by microbiologists, the statistical methods needed to analyse the resulting huge amounts of numerical information still represent an obstacle for many microbial ecologists. The R environment already plays an important role in genomic data analysis thanks to the bioconductor project, but the statistical methods needed for multivariate ecological data analysis are part of standard R packages, like vegan and ade4. Although these packages are designed for plant or animal ecology, many of their functions can be used to analyse molecular biology data sets. Furthermore, other packages, like seqinr and made4 are very useful to bridge standard packages and genomic data structures. Lastly, graphical user interfaces are also needed to help biologists master the intricacies of some R functions.

Chipster: A graphical user interface to DNA microarray data analysis using R and Bioconductor

Jarno Tuimala, Aleksi Kallio, Janne Käki, Taavi Hupponen, Petri Klemelä, Mikko Koski, Mika Rissanen, Eija Korpelainen
Aleksi.Kallio@csc.fi, Jarno.Tuimala@csc.fi, Eija.Korpelainen@csc.fi
CSC, the Finnish IT Center for Science

In order to enable more researchers to benefit from the method development in the Bioconductor-project, we have created analysis software Chipster for microarray data. Chipster offers an intuitive graphical user interface to a comprehensive collection of up-to-date analysis methods.

Chipster supports all major DNA microarray platforms and, being a Java program, it is compatible with Windows, Linux and MacOS X. The basic analysis features such as preprocessing, statistical tests, clustering, and annotation are complemented with, e.g., linear (mixed) modeling, bootstrapping hierarchical clustering results, and finding periodically expressed genes from time series data. Analysis history is automatically recorded, and the analysis scripts can be viewed at the source code level.

Chipster can not only display images produced by R and Bioconductor, but also produce interactive visualizations for various clustering results, 2D and 3D scatter plots, histograms and time series plots. Users can freely choose different features of datasets to be plotted, such as log transformations of expression values.

Graphical client software runs on the user's computer, and connects to a remote server environment through a front-end server. Chipster can also connect to external Web Services. There is a possibility to set up a stand-alone version of the analysis environment on a Linux system, and an open source version will be available through SourceForge.

The technical implementation is designed to maximize flexibility and minimize memory usage and data transfer between components. New tools can be added using a simple annotation system, and no modifications or wrappers are needed. Analyzer instances are pooled so that analysis requests can be processed as fast as possible.

For more information about Chipster, please see:
http://www.csc.fi/molbio/microarrays/nami
http://chipster.csc.fi
http://www.sourceforge.org/projects/chipster

Custom Functions for Specifying Nonlinear Terms to `gnm`

Heather Turner, David Firth and Andy Batchelor

Department of Statistics, University of Warwick, UK

Contact: Heather.Turner@warwick.ac.uk

`gnm` is a function provided by the *gnm* package for fitting generalized nonlinear models. These models extend the class of generalized linear models by allowing nonlinear terms in the predictor. Nonlinear terms can be specified in the model formula passed to `gnm` by functions of class `nonlin`. A number of these functions are provided by the *gnm* package. Some specify basic mathematical functions, such as `Exp` for specifying an exponentiated term, whilst others are more specialized, such as the `Dref` function for specifying diagonal reference terms as proposed by Sobel (1981, 1985).

Users are able to nest the `nonlin` functions provided by *gnm* in order to specify more complex nonlinear terms. However this functionality is limited in the terms that can be specified and can result in rather long-winded model descriptions. The alternative is to write a custom `nonlin` function to fit the desired term. Turner and Firth (2007) explain how to write such a function using a standard example of a logistic model; whilst this provides a useful illustration, that particular model would be more simply handled in practice using `nls`. In this talk we demonstrate how to write a custom `nonlin` function in the context of a novel application of generalized nonlinear models.

Our application is modelling the hazard of entry into marriage for women in Ireland, based on data from the Living in Ireland Survey conducted in 1994-2001 by the Economic and Social Research Institute. We propose a nonlinear discrete-time hazard model, extending the approach of Blossfeld and Huinink (1991). This model may be fitted as a generalized nonlinear model, but requires a custom `nonlin` function to specify the terms. We show how to write such a function, exploring the different options available and considering the difficulties that can arise.

References

Turner, H. and Firth, D. (2007). gnm: A Package for Generalized Nonlinear Models. R News, 2007, 7/2, 8-12, R Foundation for Statistical Computing.

Blossfeld, H.-P. and J. Huinink (1991). Human capital investments or norms of role transition? How womens schooling and career affect the process of family formation. American Journal of Sociology, 97, 143-168.

M. E. Sobel (1981). Diagonal mobility models: A substantively motivated class of designs for the analysis of mobility effects. Amer. Soc. Rev., 46, 893-906.

M. E. Sobel (1985). Social mobility and fertility revisited: Some new models for the analysis of the mobility effects hypothesis. Amer. Soc. Rev., 50, 699-712.

Using R to test Bayesian adaptive discrete choice designs

Boris Vaillant

We present a proof of concept in R for the implementation of truly adaptive discrete choice designs.

These algorithms use MC methods to update the posterior probability after each new answer and generate new product comparisons based on a variety of possible target measures (A / D-criterion, minimal expected entropy of the posterior or maximal entropy of the next question).

We provide results comparing different adaptive strategies with fixed MNL- and linear designs based on a simulation study performed in R. Compared to well-known industrial solutions for adaptive question generation our methods are consistently based on discrete choice theory and should therefore lead to more reliable results.

Refactoring R Programs

Tobias Verbeke (Business & Decision)

Refactoring code has been daily bread for developers since the advent of programming languages and is given a central role in modern programming methodologies such as eXtreme programming. Automation of refactoring operations is therefore supported by many professional IDEs for common programming languages.

For the R language, there has not yet been an in-depth study of refactoring operations and the current IDEs have no or limited support for it. In this presentation we determine how the specificities of the R language (as a functional language with object orientation) impact R software change and refactoring.

In a first part, the common refactoring operations are reviewed and a typology of the operations is proposed. The typology is confronted with other refactoring categorizations and frameworks published in the software engineering literature. Special attention will be given to the possibilities the R package concept offers to keep R code and other software artifacts (documentation, tests, etc.) in sync.

In a second part, a reflection is offered on user interfaces for automated refactoring (refactoring browsers etc.). This reflection will be based on studying interfaces for other programming languages in comparative perspective. The resulting refactoring framework and interface are planned to be integrated into the StatET eclipse plugin for R, though it is hoped for that other IDEs will benefit as well from our results.

Segmented Poisson Models

Vidal E[1,2], Pastor-Barriuso R[1,2], Pollan M[1,2], Lopez-Abente G[1,2].

[1] Environmental and Cancer Epidemiology Unit, National Centre for Epidemiology, Carlos III Institute of Health. Madrid, Spain.

[2] CIBERESP, Spain.

Standard dose-response analyses (such as categorical, spline, or nonparametric regression) provide flexible tools to describe the overall shape of the dose-response relation across the entire exposure range, but the identification of trend changes with these methods is subjective. Specific methods are needed to formally test for the existence of change-points in risk trends.

We propose a log-linear model for aggregated data with Poisson variance and free dispersion parameter, in which the predictor function consists of two intersecting straight lines connected at an unknown change-point through a hyperbolic transition function, that allows for abrupt changes or more gradual transitions between the linear trends. The model, that was implemented as an R function, provides a p-value for the existence of a change-point, as well as point and interval estimates for its location and the slopes below and above it.

An application to two different scenarios is presented. First, relationship between Spanish renal cancer mortality (period 1994-2003) and distance to metallurgical facilities (provided by the EPER register) at municipal level was analysed, adjusting by age-group, sex and socio-economic indexes. Second, we look for changes in time trend of breast cancer incidence (adjusted by age) taken from Spanish registries covering 16 of the 50 Spanish provinces in the last 30 years.

The results are as follows: In the first scenario, we found a significant change point (at 5 Km, CI 95% 3 13 Km away from point source) for men. Below this point, relative risk decreased with distance and above it, the trend stabilizes. No change point was found for women. In the second, breast cancer incidence increased in Spain during the 70s, 80s and 90s (at a rate of 2.4 per year) and levelled in the XXI century (change point found at 1999 CI 95% 1996 2001).

As conclusion, it seems that change point models offer a good alternative for the linear dose-response relationships when using regression in a set of different epidemiological situations.

Acknowledgements: Staff from the Cancer Registries of Spain: Nerea Larrañaga, Miren Dorronsoro, Isabel Izarzugaza form the Basque Country, Carmen Martos from Zaragoza, Marcial Argüelles from Asturias, Mª Dolores Chirlaque from Murcia, Eva Ardanaz from Navarra, Mª Dolores Rojas from Canary Islands, Jaume Galcerán from Tarragona, Rafael Marcos-Gragera from Girona, Carmen Martínez from Granada, Isabel Garau from Majorca, Ana Torrella from Castellon, Antonio Mateos from Albacete, Josefina Perucha from La Rioja, Rosario Jiménez-Chillarón from Cuenca.

RGG: An XML-based GUI Generator for R Scripts

Visne Ilhami[1] *, Vierlinger Klemens[1], Leisch Friedrich[2], Kriegner Albert[1]
[1] Austrian Research Centers GmbH - ARC, Molecular Diagnostics, A-2444 Seibersdorf, Austria
[2] Institut für Statistik, Ludwig-Maximilians-Universität, Ludwigstraße 33, D-80539 München, Germany
* ilhami.visne@arcs.ac.at

R is the leading open source statistics software with many analysis packages developed by the user community. However, the use of R requires programming skills. We have developed a software tool, called R GUI Generator (RGG), which enables generation of Graphical User Interfaces (GUI) for R scripts by adding a few simple XML-tags. An RGG file (.rgg), which contains R code and GUI elements, serves as a template for the GUI engine. The GUI engine loads the RGG file and at runtime creates and arranges GUI elements from the XML tags. User-GUI interactions are converted into the corresponding R code, which replace the XML tags. As a result a new R script is generated from the template. The project's aim is to provide R developers with a tool to make R based statistical computing available to a wider audience less familiar with script based programming. The project further includes the development of a repository and documentation system for R-GUIs being developed by community. The project's website is at http://rgg.r-forge.r-project.org.

GridR: Distributed Data Analysis using R

Dennis Wegener, Stefan Rüping and Michael Mock

Fraunhofer Institute for Intelligent Analysis- and Information Systems
Schloss Birlinghoven
53754 St. Augustin, Germany
{ dennis.wegener, stefan.rueping, michael.mock }
@iais.fraunhofer.de

Abstract. In the last couple of years, the amount of data to be analyzed in different areas grows rapidly. Examples range from natural sciences (e.g. astronomy or particle physics), business data (e.g. a high increase use data volume is expected by the use of RFID technology), life sciences (such as high-throughput genomics and post-genomics technologies) or data generated by normal users on the internet (see Google, Youtube, etc.). The enormous growth of the amount of data is complemented by advances in distributed computing technology enabling the data analyst to handle this amount of data in reasonable time. Two main streams of current distributed technology development and research are particularly useful in this respect: the grid technology is aiming at making data stores and computing facilities which are geographically widely spread available for a common, global data analysis. The other stream of development is cluster-based computing which transforms large amounts of standard computers into high-performance computing bases.

However, even if the above mentioned advances in distributed computing technology make available the computing and storage resources for handling large amounts of data, they introduce another level of complexity in the system, such that the traditional data analyst, with a strong background in statistics and application domain knowledge, might be overwhelmed by the complexity of the underlying distributed technology. For instance, an application developer using R might not be interested in any details of how web services are built. Therefore, ongoing research aims at bridging the gap between advanced distributed computing technology and traditional statistical software.

The Advancing Clinico-Genomics Trials on Cancer project (ACGT) aims at providing a data analysis environment that allows the exploitation of an enormous pool of data collected in European cancer treatments. In the context of this project, the GridR package was developed, which was one of the first attempts to connect R to a grid environment - to grid-enable R.

Keywords: R statistical language, Grid, GridR, ACGT

Acknowledgements. The authors gratefully acknowledge the support of the ACGT project that is funded by the European Commission (FP6/2004/IST-026996).

References

1. Dennis Wegener, Thierry Sengstag, Stelios Sfakianakis, Stefan Rüping and Anthony Assi. GridR: An R-based grid-enabled tool for data analysis in ACGT clinico-genomic trials. In: Proceedings of the 3rd International Conference on e-Science and Grid Computing (eScience 2007), Bangalore, India.
2. Stefan Rüping, Stelios Sfakianakis and Manolis Tsiknakis. Extending Workflow Management for Knowledge Discovery in Clinico-Genomic Data. In: From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences, Proceedings of HealthGrid 2007, pp. 183-193, IOS Press, 2007.
3. Vlado Stankovski, Martin Swain, Valentin Kravtsov, Thomas Niessen, Dennis Wegener, Joerg Kindermann, and Werner Dubitzky. Grid-enabling data mining applications with DataMiningGrid: An architectural perspective. Future Generation Computer Systems Journal, 2007.
4. Vlado Stankovski, Martin Swain, Valentin Kravtsov, Thomas Niessen, Dennis Wegener, Matthias Röhm, Jerney Trnkoczy, Michael May, Jürgen Franke, Assaf Schuster and Werner Dubitzky. Digging Deep into the Data Mine with DataMiningGrid. IEEE Internet Computing, accepted for publishing in 2007.
5. Dennis Wegener and Michael May. Extensibility of Grid-Enabled Data Mining Platforms: A Case Study. In Proc. of the 5th International Workshop on Data Mining Standards, Services and Platforms, KDD 2007, pages 13--22, San Jose, USA, August 2007. ISBN 978-1-59593-838-1.

Commercial meets Open Source - Tuning STATISTICA with R

Christian H. Weiß*

March 10, 2008

Abstract

R is an extremely powerful environment for statistical computing: It provides packages designed for different areas such as data mining, econometrics, epidemiology, biostatistics, it offers methods from different statistical disciplines like time series analysis, statistical process control, bootstrapping, cluster analysis, and others. Besides its mere extent, R differs from competing statistics environments also in the fact that it reflects the state-of-the-art in statistical sciences. And not to forget: R is freely available.

On the other hand, R is not particularly user-friendly: It does not offer a graphical user-interface, where the repertoire of methods is fully integrated and available also for users, who have not learnt the R language. It does not offer a powerful spreadsheet environment, which enables an intuitive way of data manipulation. Therefore, (potential) users from applied sciences and industry often do not have the heart to work with R.

In this talk, I propose to combine the power of R with the comfort of a commercial package like STATISTICA. STATISTICA can be used as an easily operated interface with a respectable basic equipment of statistical procedures, see Weiß (2006). But if required, one can easily integrate specialised statistical procedures and sophisticated techniques offered by R into the user interface of STATISTICA. Besides the base version of STATISTICA with its Visual Basic development environment, and besides R together with the required packages, the user only needs to install the R DCOM Server of Baier & Neuwirth (2007).

The necessary procedure and essential commands to access R from STATISTICA are explained, also refer to StatSoft (2003). A number of examples highlight situations, where R can be used to extend the functionality of STATISTICA. Among others, we explain how an ARL calculator for computing average run lengths of EWMA and CUSUM control charts can be programmed, using the spc package of Knoth (2007). The ARL calculator supports the design of these control charts, which are themselves available through STATISTICA.

*Institute of Mathematics, Department of Statistics, University of Würzburg, Germany.
Email: christian.weiss@mathematik.uni-wuerzburg.de

References

BAIER, T., NEUWIRTH, E.: *R/Scilab (D)COM Server V 2.50*. March, 2007.
`http://cran.r-project.org/contrib/extra/dcom/`

KNOTH, S.: *The spc Package (Statistical Process Control), Version 0.21*. October, 2007.
`http://cran.r-project.org/src/contrib/Descriptions/spc.html`

STATSOFT: *STATISTICA Data Miner: Integrating R Programs into the Data Miner Environment*. StatSoft Business White Paper, June, 2003.
`http://www.statsoft.com/support/whitepapers/pdf/STATISTICA_Integrating_R.pdf`

WEISS, C.H.: *Datenanalyse und Modellierung mit STATISTICA*. Oldenbourg Wissenschaftsverlag, München, 2006.

A Compendium Platform for Reproducible, R-based Research with a focus on Statistics Education

Patrick Wessa

March 30, 2008

This paper discusses a new Compendium Platform (CP) that allows us to create Reproducible Research in R that is easily accessible for anyone who has access to the internet (freestatistics.org). The platform is based on the R Framework (wessa.net) and primarily focuses on ICT-based Statistics Education within a pedagogical paradigm of individual and social constructivism which received a great deal of interest in the academic community (Von Glasersfeld (1987), Erick Smith (1999), Eggen and Kauchak (2001), and Nyaradzo Mvududu (2003)).

The basic idea is to create an environment where students are allowed to interact with each other (and the tutor) about a series of research-related activities (such as assignments or workshops) based on the R language and the R Framework. The novelty about this approach lies in the fact that the newly developed CP empowers students to easily archive, exchange, reproduce, and reuse R computations. The underlying technology facilitates the creation of a learning environment that supports social constructivism which is very similar to the real world of applied statistical research. More importantly, the CP allows us to obtain physical measurements of the actual learning process of students based on detailed information about the use of the statistical software, and the socially constructivist learning activities (based on peer review of statistical analysis in R).

The CP was thoroughly tested in two undergraduate statistics courses with large student populations. During these courses a large number of physical and survey-based measurements were obtained and studied. The preliminary analysis of the relationships between learning attitudes, social interaction (through group work and peer review), learning experiences, software usability, usage of archived R computations, and exam scores (that are related to statistical competences rather than knowledge) is presented.

One of the most interesting results is that social interaction through peer review based on Reproducible Research (which is used as a "learning activity" rather than an "evaluation tool") is very beneficial for the learning experiences of students, and exam scores. Also, there is a strong, positive relationship between the use of the CP and exam performance - even if other important factors are taken into account. Another interesting result is that a large majority of students have a positive perception about the new system as a learning tool and prefer the constructivist approach based on Compendia above traditional learning methods.

In addition, it is (very) briefly illustrated how Compendia of Reproducible Research can be used to:

- write Compendium-based course materials
- detect plagiarism and free-riding
- quickly identify (and find solutions for) bugs and computing-related problems
- estimate the workload of an assignment
- support new forms of collaboration that lead to improved solutions in R

Finally, some important aspects about the near and distant future of the CP and the underlying R Framework (for the purpose of education, scientific research, and publishing) are illustrated and discussed.

Acknowledgements

This research is funded by the KULeuven Association (OOF 2007/13 Project)

Analyzing paired-comparison data in R using probabilistic choice models

Florian Wickelmaier, Department of Psychology, University of Tübingen

When human subjects are required to evaluate a set of options or stimuli with respect to some attribute, the simplest data that can be obtained are binary paired-comparison judgments. Such data might result from so-called sensory evaluation studies, where participants are asked to judge which of two audio samples sounds brighter or more natural, which of two coffee brands tastes better, or from surveys where subjects are to indicate which political party they would vote for or which insurance package they would rather buy. It is the goal of the analysis of the data to arrive at a scaling of the options involved.

A well known model for paired-comparison data is the Bradley-Terry-Luce (BTL) model that relates the pairwise choice probabilities to scale values representing the weight or strength of each option. Often in empirical studies, however, it is found that the data do not meet the restrictions imposed by the BTL model, one of them being that the choices are made independently of the context introduced by a given pair. In psychology, more general models have been developed, the most prominent one being the elimination-by-aspects (EBA) model (Tversky, 1972; Tversky & Sattath, 1979), which does not require context independence of the judgments.

Although these general models seem to be promising alternatives to the BTL model, they have not been frequently applied, presumably due to the lack of easy-to-use software for their fitting and testing. The presentation will illustrate the analysis of paired-comparison data using the eba package in R (Wickelmaier & Schmid, 2004). It will be demonstrated with examples from empirical research that, whenever similarity among the options of a choice set plays a role, the modeling is more successful when more complex choice models, such as EBA, are employed.

Tversky, A. (1972). Elimination by aspects: A theory of choice. Psychological Review, 79, 281-299.

Tversky, A., & Sattath, S. (1979). Preference trees. Psychological Review, 86, 542-573.

Wickelmaier, F. & Schmid, C. (2004). A Matlab function to estimate choice-model parameters from paired-comparison data. Behavior Research Methods, Instruments, and Computers, 36, 29-40.

Deploying Data Mining in Government – Experiences With R/Rattle

Graham J. Williams

Whilst R and its many packages provide an incredibly broad and comprehensive environment for data mining in practise, there are many challenges in bringing its power to the common data mining practitioner. It is a sad fact that many analysts today only feel comfortable with the usually limiting graphical user interfaces. Yet, we can unleash the full power of analytics only through languages like R. In this presentation I will reflect on how we are bringing the power of R to a large community of data analysts and new data miners through the development of the award winning open source Rattle package for R. I will present some case examples of using R from the Australian Taxation Office and discuss how we tackled various problems in using data mining tools in practise.

Building a Reuters Real-Time Market Data Interface in R

Rory Winston

Historically, R usage tends to centre more around offline, rather than real- time data analysis. However, there are some reasons why a real-time market data interface can be of benefit. In this talk, I will talk about an interface to the Reuters market data system that I built whilst working on a real- time foreign exchange algorithmic trading project. This proved to offer some surprising benefits, and the addition of a market data query interface into R, combined with its vast library of analysis functions and easily extensible native interface, makes it an incredibly powerful tool.

Computational Finance and Financial Engineering

The R/Rmetrics Software Environment

Diethelm Würtz, ETH Zürich
Yohan Chalabi, ETH Zürich and Finance Online GmbH Zürich

Abstract

R/Rmetrics has become the premier open source solution for teaching financial market analysis and valuation of financial instruments. With hundreds of functions build on modern methods R/Rmetrics combines explorative data analysis and statistical modeling. Rmetrics is embedded in R, both building an environment which creates for students a first class system for applications in statistics and finance.

In the heart of the software environment are powerful time/date and time series management tools, functions for analyzing financial time series, functions for forecasting, decision making and trading, functions for the valuation of financial instruments, and functions for portfolio design, optimization and risk management.

In this talk we give an overview on R/Rmetrics and present new directions and recent developments.

References:

R/Rmetrics Core Team, The R/Rmetrics Software Environment,
www.rmetrics.org and r-forge.r-project.org

Statistical Animations Using R

Yihui Xie [*]
School of Statistics, Renmin University of China

Animated graphs that demonstrate statistical ideas and methods can both attract interest and assist understanding. This paper describes approaches that may be used to create animations, and gives a brief overview to the R package animation. It gives examples of the use of animations in teaching statistics and in the presentation of statistical reports. Animations can add insight and interest to traditional static approaches to teaching statistics, making statistics a more interesting and appealing subject.

Keywords: animation, statistical demonstration, simulation, limiting distributions, resampling methods, R

References

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

Yihui Xie. *animation: Demonstrate Animations in Statistics*, 2008. URL http://R.yihui.name. R package version 0.1-9.

[*] Room 1037, Mingde Main Building, School of Statistics, Renmin University of China, Beijing, 100872 China; E-mail: xieyihui[at]gmail.com; URL: http://www.yihui.name/, http: //R.yihui.name

Using R for Spatial Shift-Share Analysis

Gian Pietro Zaccomer, Luca Grassetti

Department of Statistical Sciences - University of Udine, Italy

E-mail: zaccomer@dss.uniud.it, grassetti@dss.uniud.it

Abstract

During the second half of the 20^{th} century, the Shift-Share Analysis (SSA) have been largely applied in the economic growth studies. Starting from the formulation adopted by Dunn (1960), the literature proposed various decomposition procedures based on the identification of three or more components. The SSA has always been considered a spatial statistics tool but only with Nazara and Hewings (2004) the spatial dimension has been actually considered in the model specification. The authors, in fact, introduced the effect of interaction between territorial units by means of a spatial weights matrix. The proposed model is based on a generic row standardized weighting matrix. Consequently, the authors did not face the problem of weight construction. Zaccomer (2006) proposed a solution based on the variables deriving from the italian register of businesses. The information derived from this register can be used to define two important decomposition factors: the economic activity in NACE-ATECO classification and the firm legal status. In the cited article, instead of the well known spatial weighting systems based on contiguity or on generic distance functions, the author proposed an economic concept of neighborhood. In fact, the considered matrices are based on a given economic subdivision as, for example, the Local Labor Systems (LLS) or the Industrial Districts (ID). The neighborhood defined by the "economic contiguity" can be considered the best choice if the units' partition is based on supplementary information about the studied phenomenon. For example the ID are based on the observation of firms' productive network and can be used to study the labour growth rates.

In this work we aim to study the flexibility of spatial shift share model applied to analysis of labour growth rates obseved in the local system of Friuli Venezia Giulia and its LLS. All computational issues, plots and prints functions are developed using R (R Development Core Team, 2007).

Keywords: Shift-Share Decomposition, Growth Rates, Industrial Districts, Local Labor Systems, Statistical Register of Businesses.

References

Dunn, E., S. (1960) A statistical and analytical technique for regional analysis. *Paper and Proceedings of the Regional Science Association*, **6**, 97–112.

Nazara, S. and Hewings, G., J. D. (2004) Spatial structure and taxonomy of decomposition in shift-share anlysis. *Growth & Change*, **35**, 476–490.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Zaccomer, G., P. (2006) Shift-share analysis with spatial structure: an application to italian industrial districts. *Transition Studies Review*, **13**, 213–227.

Some Perspectives of Graphical Methods for Genetic Data

Zhao JH, Q Tan, S Li, J Luan, W Qian, R Loos, NJ Wareham, MRC Epidemiology Unit, Cambridge, UK, Odense University Hospital, Denmark, and MRC Clinical Trials Unit, London, UK

Abstract

Recent initiatives have made genetic data on single-nucleotide polymorphisms (SNPs) in humans widely available. The association study between these SNPs and a host of measures in humans and other species has led to a vigorous development of analytical tools as with a great understanding of the genetic basis of common diseases. Among many aspects of the data analysis, there is a need to synthesise the graphical methods involved. I give a brief account of the background, provide examples in recent analyses, and draw attention to further work.

Specifically, the examples provided are from a number of aspects: 1. Phenotypic data. While this includes the usual summary statistics it may also be specific to genetic context such as pedigree-drawing, 2. Genotypic data. This includes plotting missing data, Hardy-Weinberg equilibrium (HWE), and the correlation between neighbouring SNPs (LD). 3. Assessment of population substructure and genotype-phenotype association. This includes scree plot, Manhattan plot, Q-Q plot, SNP-based summary plot and regional association plot. 5. Representation of pathways. Other examples may arise from study of power, meta-analysis and interactions. In addition, a comparison will be made between graphics from CRAN packages with popular standalone programs such as LD plot.