Contents

Generalized linear spatial modeling of HIV in Kenya

*Thomas N O Achia

April 16, 2010

Abstract

Using geoRglm, Generalized Linear spatial models are used to study HIV prevalence in Kenya using data from the 2003 Demographic and Health survey. The relationship between HIV prevalence computed at 400 clusters and behavioral, socio-demographic and biological determinants was determined and significant covariates identified. We included a stationary Gaussian process S with a powered exponential spatial correlation function to account for the spatial variation in the model and used Kriging to estimate the missing data. The study finds large regional variations in the prevalence of HIV in Kenya and identifies key socio-cultural practices, among them male circumcision and societal acceptance of concurrent/multiple partnership, as principle determinants of the HIV transmission in Kenya.

Key Words: Generalized linear spatial model; Kriging; spatial clustering; Kenya

*Email: achia@uonbi.ac.ke, School of Public Health, University of Western Cape, SA

Using *R* for Data Mining in Vaccine Manufacturing:
Finding Needles in Biological Haystacks

Nelson Lee Afanador

Center for Mathematical Sciences
Merck Manufacturing Division
Merck & Co., Inc.
WP28-100
P.O. Box 4
West Point, PA 19486
215-652-0067
215-993-1177
nelson.afanador@merck.com

Presenter: Nelson Lee Afanador

Key Words: Data Mining, Random Forest, Recursive Partitioning, Vaccine Manufacturing

Purpose: To illustrate the application of data mining tools available in *R* in helping drive at root causes for changes in either cellular growth or viral propagation. Two vaccine manufacturing case studies will be presented.

Abstract

Vaccine manufacturing is the latest field in which advanced data mining methods are beginning to gain widespread acceptance among biologists, engineers, and data analysts. Vaccine manufacturing itself is a complex biological process composed of hundreds of steps carried out over several months. During the manufacture of a single vaccine lot hundreds of processing variables and raw materials are monitored. The challenging aspects with respect to mining biological manufacturing process data begins with obtaining a suitable dataframe which to analyze, proceeds to inherent variation in raw material composition and processing conditions, and ends with high inherent variation in the measurement systems. As such, identifying the root cause candidates for changes in cellular growth or viral propagation is extremely challenging due to the high number of candidate variables and variable processing conditions.

Given the large numbers of available candidate variables the traditional methods of univariate statistical process control charting, analysis of variance, and least squares regression leave many questions unanswered. Random Forest (**randomForest**) and single-tree recursive partitioning (**rpart**), coupled with cumulative sum charts (**qcc**), have proven to be important methods in helping drive at potential root causes for observed changes. Leading candidate variables identified via the overall analysis can then be further investigated using more traditional statistical methods, including designed experiments.

These data mining methods are setting a new standard for vaccine root cause investigations and have proven valuable at helping solve complex biological problems. Their effectiveness, and implementation using *R*, will be illustrated with two case studies.

Teaching statistics to biologists: the R-library asbio

Ken A. Aho

Idaho State University

An understanding of statistical concepts can be enhanced through computer programming and graphical applications. The R-library asbio "applied statistics for biologists" (Aho 2010) uses these capabilities to demonstrate introductory statistical concepts often poorly understood by biological scientists. In this presentation I describe how the teaching of four important statistical ideas can be enhanced with asbio. These topics are: 1) probability density functions, 2) parameter estimation, 3) likelihood, and 4) recognition of sampling and experimental designs. First, the theory underlying classical probability can be addressed with function Venn, derivation of probability for continuous probability density functions is addressed with shade, and demonstrations of sampling distributions for descriptive and test statistics are provided by the function samp.dist. Second, least square and maximum likelihood parameter estimation can be visually demonstrated using a number of asbio functions including ls.plot and loglik.plot. Third, concepts underlying likelihood (e.g. the influence of sample size the shape of the likelihood function) and REML are explained with the functions loglik.plot and reml.plot. Fourth, a large number of sampling and experimental designs are graphically depicted with the functions runSampDesign and runExpDesign. Randomization in these designs can be demonstrated by repeatedly creating plots using these functions. I will also address the practical applications of asbio in biological research including survivorship models for species using transition matrices, relevé table summaries, models for environmental engineering in vegetation reclamation, pruning analysis of multivariate community classifications, and the tracking animal trajectories in the context of habitat patches.

BaSyLiCA★ : A web interface for automatic process of Live Cell Array data using R.

Leslie Aïchaoui-Denève[1*], Vincent Fromion [1], Stéphane Aymerich[2], Matthieu Jules[2], Ludovic Le-Chat[2], Anne Goelzer[1].

1. INRA- Mathématique, Informatique et Génome – Jouy-en-Josas, FRANCE
2. INRA- Microbiologie et Génétique Moléculaire – Grignon, FRANCE

* Contact author: laichaoui@inra.jouy.fr

Keywords: Web Interface, Polynomial Regression, Live Cell Array.

As new technologies in biology produces large amounts of data, the necessity of developing dedicated tools for management and processing increases.

In this study, we present our recent efforts in developing a user-friendly tool allowing biologists to handle time series data associated to Live Cell Array (LCA; Zaslaver A. and al., 2006). It consists of a triple tool combining a database for storage, a WEB interface to manage database and specific *R* functions for processing data.

The LCA is a library of fluorescent transcriptional reporters, which is cultivated in microtiter plates (96 wells, 12 x 8, named A1 to H12). It allows to monitor live cell (for example Bacillus subtilis) transcriptional changes over time (e.g. every 1 to 10 minutes for hours). Two quantities are measured over time by the robot (a multi-detection microplate reader), the optical density (which is proportional to the number of individuals) and the fluorescence (which informs on the expression of a gene thanks to the fluorescent reporter).

The first problem was to store the data automatically acquired by the robot, as well as all information related to experimental conditions (temperature, injection ...) and the description of strains used. To do so, we chose a *MySQL* database. Automatic insertion and formatting of data were performed with the *R* package **RMySQL** (David A. and al., 2009). Manual insertions of experimental conditions were made through the interface.

The goal of this experiment is to evaluate the activity of the promoter (Zaslaver A. and al., 2004). This activity is defined as the derivative of fluorescence with respect to time, divided by the optical density. We have to take into account that data are noisy, bacteria have a natural auto-fluorescence, and shifted with respect to time because bacteria do not start growth at the same time in all wells (so-called “lag-phase”). To obtain this derivative we proposed several methods of treatment of curves. One of them is to smooth the curves then adjust the offset curves and remove automatically problematic ones. Another method uses the Kalman filter to estimate the fluorescence derivative.

A *PHP* interface is available to enable biologist to upload files and to enter information in the database in a few clicks. Furthermore, different results and graphs can be produced through *R*, according to different criteria, and downloaded as CSV file (Comma-Separated Values) and PDF format. Users can also simply click and view the contents of a plate (strains, culture medium, injected products ...). The interface has an admin part that can manage access rights to the data according to their property (public or private) and the level of the user (administrator or simple user).

The tool has been applied on a data set of 80 plates in the European project BaSysBio (Bacillus Systems Biology). The aim of this project is to study the global regulation of gene transcription in a model bacterium: Bacillus subtilis.

★BaSyLiCA : BaSysBio Live Cell Array.

References

David A. James and Saikat DebRoy (2009). *RMySQL: R interface to the MySQL database*, http://biostat.mc.vanderbilt.edu/wiki/Main/RMySQL.

Zaslaver A, Bren A, Ronen M, Itzkovitz S, Kikoin I, Shavit S, Liebermeister W, Surette MG, Alon U. (2006), A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli. Nat Methods*. 3, 623-8.

Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG, Alon U. (2004). Just-in-time transcription program in metabolic pathways. *Nat Genet*. 36, 486-91.

The R package simFrame: An object-oriented approach towards simulation studies in statistics

Andreas Alfons[1,⋆], **Matthias Templ**[1,2], Peter Filzmoser[1]

1. Department of Statistics and Probability Theory, Vienna University of Technology
2. Department of Methodology, Statistics Austria

⋆Contact author: alfons@statistik.tuwien.ac.at

Keywords: R, statistical simulation, object-oriented programming

Due to the complexity of modern statistical methods, researchers frequently use simulation studies to gain insight into the quality of developed procedures. Two types of simulation studies are often distinguished in the literature: *model-based* and *design-based* simulation. In model-based simulation, data are generated repeatedly based on a certain distribution or mixture of distributions. Design-based simulation is popular in survey statistics, as samples are drawn repeatedly from a finite population. The R package **simFrame** (Alfons 2009, Alfons et al. 2009) is an object-oriented framework for statistical simulation, which allows researchers to make use of a wide range of simulation designs with a minimal effort of programming.

Control objects are used to handle the different steps of the simulation study, such as drawing samples from a finite population or inserting outliers or missing values. Along with the function to be applied in every iteration, these control objects are simply passed to a generic function that carries out the simulation study. Loop-like structures for the different simulation runs (including, e.g., iterating over various contamination levels or performing the simulations independently on different subsets) are hidden from the user, as well as collecting the results in a suitable data structure. Moving from simple to complex simulation designs is therefore possible with only minor modifications of the code. In native R, on the other hand, such modifications would require a considerable amount of programming. In addition, the object-oriented implementation provides clear interfaces for extensions by the user. Developers can easily implement, e.g., specialized contamination or missing data models.

Since statistical simulation is an *embarrassingly parallel* process, **simFrame** supports parallel computing in order to increase computational performance. The package **snow** (Rossini et al. 2007; Tierney et al. 2009) is thereby used to distribute the workload among multiple machines or processor cores. Furthermore, an appropriate plot method for the simulation results is selected automatically depending on their structure.

References

Alfons A. (2009). **simFrame**: Simulation Framework. R package version 0.1.2.
`http://CRAN.R-project.org/package=simFrame`

Alfons A., Templ M. and Filzmoser P. (2009). **simFrame**: An object-oriented framework for statistical simulation. *Research Report CS-2009-1*, Department of Statistics and Probability Theory, Vienna University of Technology.
`http://www.statistik.tuwien.ac.at/forschung/CS/CS-2009-1complete.pdf`

Rossini A.J., Tierney L. and Li N. (2007). Simple parallel statistical computing in R. *Journal of Computational and Graphical Statistics*, 16(2), 399–420.

Tierney L., Rossini A.J., Li N. and Sevcikova H. (2008). **snow**: Simple network of workstations. R package version 0.3.3.
`http://CRAN.R-project.org/package=snow`

US Census Spatial and Demographic Data in R: the UScensus2000-suite of packages

Zack Almquist[1*]

Department of Sociology, University of California, Irvine
*Contact author: almquist@uci.edu

Keywords: Spatial Analysis, Demography, Data, Census

The US Decennial Census is arguably the most important data set for social science research in the United States. The **UScensus2000**-suite of packages allows for convenient handling of the 2000 US Census spatial and demographic data. The goal of this presentation is to showcase the **UScensus2000**-suite of packages for R, to describe the data contained within these packages, and to demonstrate the helper-functions provided for handling this data. The **UScensus2000**-suite is comprised of spatial and demographic data for the 50 states and Washington DC at four different geographic levels (Block, Block Group, Tract, and Census Designated Place (CDP)). The **UScensus2000**-suite also contains a number of functions for selecting and aggregating specific geographies or demographic information such as Metropolitan Statistical Areas (MSA), Counties, etc. These packages rely heavily on the spatial tools developed by Bivand et al. (2008) (i.e., the **sp** and **maptools** packages). This presentation will provide the necessary background for working with this data set, helper-functions, and finish with an applied spatial statistics example.

References

Bivand, Roger S., Edzer J. Pebesma, and Virgilio Gómez-Rubio. 2008. *Applied Spatial Data Analysis with R*. New York, NY: Springer.

Making R accessible to Business Analysts with TIBCO Spotfire

Lou Bajuk-Yorgan[1], **Stephen Kaluzny**[1]

1. TIBCO Spotfire

While R provides a huge breadth of deep analytical capabilities which can improve the decision-making of a wide range of users, many business analysts are intimidated by a command-line driven statistical application, and unlikely to invest the time required to become proficient with R. Putting the power of R into the hands of this wider community of users is critical if an organization wants to leverage its investments in developing algorithms and models, and help these users make more analytically-driven decisions.

This talk will show how R users can combine R scripts and functions with Spotfire visualizations to provide these business analysts with analytically-rich, easy-to-use and relevant applications. Spotfires interactive, visual capabilities for data analysis empower individuals to easily see trends, patterns outliers and unanticipated relationships in data without extensive training.

The talk will include examples from Financial Services, Life Sciences, and Customer Analytics.

Building Segmented Models Using R and Hadoop

Collin Bennet[1], David Locke[1], Robert Grossman[1,2*], Steve Vecik[1]

1. Open Data Group
2. Laboratory for Advanced Computing, University of Illinois at Chicago
*Contact author: rlg1@opendatagroup.com

Keywords: segmented models, preprocessing data using Hadoop, scoring data using R and Hadoop

We introduce a simple framework for building and scoring models on datasets that span multiple disks in a cluster. We assume that through another analysis we know an appropriate means to segment the data. We use Hadoop to clean the data, preprocess the data, compute the derived features, and segment the data. We then invoke R individually on each of the segments and produce a model. By model here, we mean a model as formalized by Version 4.0 of the Predictive Model Markup Language (PMML) [1]; similarly, by segment, we mean a segment as formalized by the PMML Specification for Multiple Models, which includes both multiple models from ensembles and segments. We then gather the resulting models from each segment to produce a PMML multiple model file. Scoring is similar, except each node has access to the entire PMML multiple model file and not just the segment associated with the node. As in the first step that produces the multiple model file, preprocessing may use all the nodes, but each feature vector is sent to the appropriate node via segmentation for scoring. The framework is called Sawmill, depends upon Hadoop and R, and is open source. Sawmill also supports other parallel programming frameworks that generalize MapReduce, such as Sector/Sphere's User Defined Functions [2].

References

[1] PMML 4.0 - Multiple Models: Model Composition, Ensembles, and Segmentation, retrieved from www.dmg.org on February 10, 2010.

[2] Yunhong Gu and Robert L Grossman, Sector and Sphere: Towards Simplified Storage and Processing of Large Scale Distributed Data, Philosophical Transactions of the Royal Society A, Volume 367, Number 1897, pages 2429–2445, 2009. Sector/Sphere is open source and available from sector.sourceforge.net.

To Do or Not To Do Business with a Country: A Robust Classification Approach

Kuntal Bhattacharyya, Pratim Datta and David Booth

ABSTRACT

In the face of global uncertainty and a growing reliance on 3rd party indices to gain a snapshot of a country's operations, accurate decision making makes or breaks relationships in global trade. Under this aegis, we question the validity of the maximum likelihood regression model in classifying countries for doing business. This paper proposes that a weighted version of the Bianco and Yohai (BY) estimator, free of distributional assumptions and outlier-effects, is a superlative tool in the hands of practitioners to gauge the correct antecedents of a country's internal environment and decide whether to do or not do business with that country. In addition, the robust process is effective in differentiating between "problem" countries and "safe" countries for doing business. An existing "R" program for the BY estimation technique has been modified to fit our cause.

Massively parallel analytics for large datasets in R with nza package

Przemyslaw Biecek[2,1*], Pawel Chudzian[1*], Cezary Dendek[1], Justin Lindsey[1]

1. Nzlabs, Netezza
2. Warsaw University

* Both authors contributed equally to this work.

Corresponding author: przemyslaw.biecek@gmail.com

Keywords: large datasets, parallel processing, high performance processing, Netezza Performance Server, data aggregates

One of the bottlenecks towards processing large datasets in *R* is the need of storing all data in memory. Therefore, users are limited to datasets that fit in 2 or 4 GB memory limit. To avoid this, the natural approach is to split statistical algorithms in two steps. In the first step the data processing is performed outside *R*, e.g. in database or on flat text file resulting in precomputed data aggregates. In the second step these aggregates are imported into *R* where the rest of the analysis is performed. Such data aggregates are called sufficient statistics, because they contain all information necessary to compute parameter estimates, test statistics, confidence intervals and model summaries while are much smaller than the original dataset.

Such an approach is implemented in the **nza** package. In the first step the **nzr** package is used to connect with Netezza Performance Server (NPS) and stored procedures are used to compute data aggregates in a parallel fashion. Having data stored in a parallel/multi-nodes database has two main advantages: there is no limit on the size of accessible datasets and data aggregates are computed in a parallel fashion which significantly reduces computation time. In some cases achieved reduction is linear with the number of processors in the database server.

Following algorithms are implemented in the **nza** package using sufficient statistic approach: correspondence analysis, canonical analysis, principal component analysis, linear models and mixed models, ridge regression, principle component regression and others.

In our presentation we will show the performance study for algorithms implemented in the **nza** package.

An Experiment Data Analysis Framework: Evaluating Interactive Information Behaviour with R

Ralf Bierig[1,*], **Jacek Gwizdka**[1], **Michael Cole**[1], **Nicholas J. Belkin**[1]

1. Rutgers University, SC&I, 4 Huntington Street, New Brunswick, NJ 08901, USA
*Contact author: user2010@bierig.net

Keywords: user experimentation, modeling, data integration

Interactive information behavior is concerned with analyzing large volumes of user data and applying results to improve information search systems. This paper describes a system framework that 1) supports the collection of a wide variety of interactive user data through experimentation and 2) processes (integrates and segments) and analyses this data with the integrative application of R.

The *experiment system*, described in more detail in [1], combines an extensible set of tasks with progress and control management and allows researchers to collect data from a set of extensible logging tools collecting data on both server and client.

The *analysis system* unifies the data and allows researchers to segment data in semantic units (e.g. based on screen regions or user decisions) and develop models of information behaviour (e.g. for detecting reading activity and predicting usefulness of web content). R is closely integrated with the framework to enable models that can process millions of data points and visualize results for the researcher through a web-based user interface. Integration with Java is facilitated through JRI [1] at the modeling layer and through **RJDBC** at the database back-end.

Our work presents an application of R in information science and points to a promising open source project that allows for an integrative use of R for experiment data analysis in interactive information behavior.

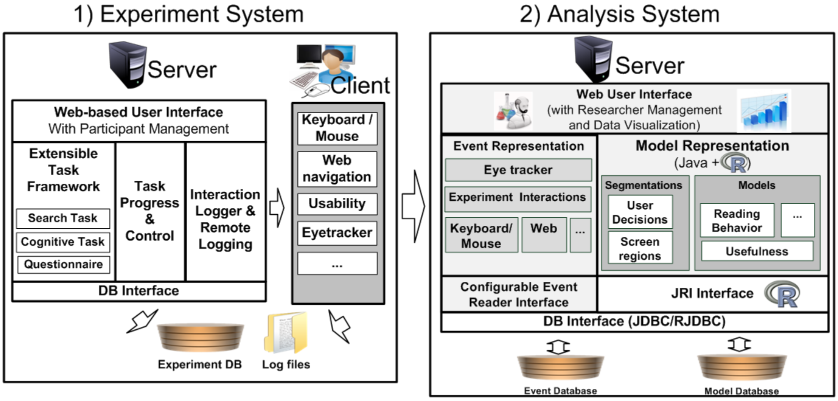


Figure 1: Components of the experiment and analysis system framework

References

[1] Ralf Bierig, Jacek Gwizdka, and Michael Cole, *A user-centered experiment and logging framework for interactive information retrieval*, SIGIR 2009 Workshop on Understanding the user - Logging and interpreting user interactions in IR (Boston, MA) (Nicolas J. Belkin, Ralf Bierig, Georg Buscher, Ludger van Elst, Jacek Gwizdka, Joemon Jose, and Jaime Teevan, eds.), 2009.

[1] `http://www.rforge.net/JRI`

binGroup: A Package for Group Testing

Christopher R. Bilder[1]*, Boan Zhang[1], Frank Schaarschmidt[2], Joshua M. Tebbs[3]

1. Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE
2. Institute of Biostatistics, Leibniz University Hannover, Germany
3. Department of Statistics, University of South Carolina, Columbia, SC

*Contact author: chris@chrisbilder.com

Keywords: Binary response, Generalized linear model, Group testing, Latent response, Pooled testing

When the prevalence of a disease or of some other binary characteristic is small, group testing is frequently used to estimate the prevalence and/or to identify individuals as positive or negative. Examples of its use include human infectious disease detection, veterinary screening, drug discovery, and insect vector pathogen transmission rate estimation (Pilcher et al., 2005; Peck, 2006; Remlinger et al., 2006; Tebbs and Bilder, 2004). We have developed the **binGroup** package as the first package designed to address the estimation problem in group testing. We present functions to estimate an overall prevalence for a homogeneous population. Also, for this setting, we have functions to aid in the very important choice of the group size. When individuals come from a heterogeneous population, our group testing regression functions can be used to estimate an individual probability of disease positivity by using the group observations only. We illustrate our functions with data from a multiple vector transfer design experiment and a human infectious disease prevalence study.

References

C. Peck. (2006). Going after BVD. *Beef*, 42, 34–44.

C. Pilcher, S. Fiscus, T. Nguyen, E. Foust, L. Wolf, D. Williams, R. Ashby, J. O'Dowd, J. McPherson, B. Stalzer, L. Hightow, W. Miller, J. Eron, M. Cohen, and P. Leone. (2005). Detection of acute infections during HIV testing in North Carolina. *New England Journal of Medicine*, 352, 1873–1883.

K. Remlinger, J. Hughes-Oliver, S. Young, and R. Lam. (2006). Statistical design of pools using optimal coverage and minimal collision. *Technometrics*, 48, 133–143.

J. Tebbs and C. Bilder. (2004). Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. *Journal of Agricultural, Biological, and Environmental Statistics*, 9, 75–90.

Analysis of Data from Method Comparison Studies Using *R*, **merror**, and **OpenMx**

Richard A. Bilonick[1,2,*]

1. University of Pittsburgh School of Medicine, Dept. of Ophthalmology
2. University of Pittsburgh Graduate School of Public Health, Dept. of Biostatistics
*Contact author: rab45@pitt.edu

Keywords: measurement error, bias, imprecision, structural equation model, latent variable

Method comparison studies are used to compare the measurements made of the same quantity using different instruments, monitors, devices or methods. Researchers often try to compare the measurements made by just two devices, in some cases using regression analysis. As is well known, when both devices are subject to measurement error (ME), regression analysis presents a distorted view of the actual bias (or shows bias when none exists). Analogously, correlation is often used to describe the agreement even though correlation only measures linear association and cannot provide any insight into agreement of the two instruments. Other researchers resort to the use of Bland-Altman plots (paired differences versus paired averages). This approach is useful if there is good agreement but provides no way to determine the cause of poor agreement.

A more fruitful and general approach is provided by using the classic ME model (Jaech, 1985):

$$X_{ij}=\alpha_i+\beta_i\mu_j+\epsilon_{ij}$$

where μ denotes the “true” but unknown value of the *j*th item being measured (often assumed to be Normally distributed with mean $\bar{\mu}$ and standard deviation σ), X_{ij} denotes the observed measurement from instrument *i* for item *j*, α_i and β_i describe the bias introduced by the instrument (assumed to be a linear function of μ), and ϵ_{ij} denotes a Normally distributed random error of instrument *i* and item *j* with mean of 0 and standard deviation of σ_i . The instrument imprecision adjusted for differences in scale is given by σ_i/β_i . The ME model can be described using a path diagram with μ as a latent variable and X_i as manifest variables, and represented as a structural equation model (SEM). SEM path analysis readily explains the deficiencies of using only two devices and the necessity of including repeats and/or 3 or more devices. SEMs can easily include repeated measurements, or, for example as in ophthalmology, having measurements from both eyes. Parameter estimates can be made using the method of moments (MOM) or the method of maximum likelihood (ML). Using these estimates, calibration equations relating measurements from different instruments can be easily derived.

The **merror** package (Bilonick, 2003) provides the function `ncb.od` for computing ML estimates of the ME imprecision standard deviations for unclustered data and least squares estimates of β_i . The function `lrt` tests whether $\beta_i=\beta_{i'}$. Other functions are available for Grubbs (MOM) estimators. For clustered data, the more flexible **OpenMx** (SEM) package (Boker, et al., 2010) can be used for ML estimates of any desired (functions of) model parameters. When using **OpenMx** for ME models, **merror** can be used to provide good starting values for the parameter estimates. These methods will be illustrated using data from ophthalmic studies and from air pollution studies.

References

Richard A. Bilonick (2003). **merror**: Accuracy and Precision of Measurements. R package version 1.0.

Steven Boker, Michael Neale, Hermine Maes, Paras Metah, Sarah Kenny, Timothy Bates, Ryne Estabrook, Jeffrey Spies, Timothy Brick and Michael Spiegel (2010). **OpenMx**: The OpenMx Statistical Modeling Package. R package version 0.2.3-1006. http://openmx.psyc.virginia.edu.

John L. Jaech (1985). **Statistical Analysis of Measurement Errors**, John Wiley & Sons, New York.

Graphics Device Tabular Output

Carlin Brickner[1*] , Iordan Slavov[1] PhD, Rocco Napoli[1]

1. Visiting Nurse Service of New York

* Contact author: carlin.brickner@vnsny.org

Keywords: Tabular Output, Tables, Report

R has provided users with powerful graphic capabilities to produce sophisticated, aesthetically pleasing plots that meet the high standards in today's scientific reporting. However, *R* has lacked the ability to create quality tabular output within the *R* environment. Most users who produce quality tabular output rely on the typesetting system *LaTeX*. This may deter some new users from further exploring the dynamic language supported by *R*'s environment.

The gap between *R*'s graphical capabilities and its inability to produce tabular output is the underlying motivation to create a function, utilizing the **gridBase** library, to produce high-level tabular output completely within the *R* environment. The proposed tabular function provides a granular level of control by looping through a data frame and printing every element one-by-one to the graphics device. In addition, the user is able to add additional formatting through parameter declaration and defined escape characters.

Some highlights of its functionality are:

- Column, Row, Title, Subtitle Labeling
- Apply additional formatting to grouped row and column label hierarchies
- Add vertical and horizontal dividers (lines)
- Row highlighting
- Footer
- Foot Notes
- Page overflow management as well as page number (designed for long PDF reports)

The proposed tabular function can also be utilized to create wrappers to *R* functions that produce a high volume of text to the *R* console, such as the `lm` function. This wrapper captures the summary statistics, organizes them into a presentable format, and displays the tables adjacent to the model diagnostic plots.

This function should be easy to implement for any user who is familiar with calling an *R* function, while also providing the expert with additional flexibility to present high quality tabular output in any format supported by the *R* graphics device. There is a desire to further develop the logic used in this function so that its application may span different needs to present tabular output in *R*.

Venn: Powerful and High-Level and Set Manipulating ... a whole new way of working with data

Christopher Brown[1]

[1] Open Data (http://www.opendatagroup.com)

Native R does not provide great tools for working with set. Some funcitons, like subset and split exist, but these can become tedious, overly verbose and error-prone especiallly with high-dimensional data. The venn package alleviates the frustration of working sets. The venn pacakges provides: a simple R-like mechanism for defining sets, all the usual set functions (union, interesection, complement, etc.), high-level syntax and semantics, simple partioning, straight-forward segmentaion, the ability to work abstractly and apply the results to new data, and the ability to apply functions to regions of data that is very difficult using the *apply alone. Depending on how you look at it, venn is either a tool to make working with sets possible and easy OR a whole new way of working with data.

Monte Carlo Simulation for Pricing European and American Basket option

Giuseppe Bruno*

1. Bank of Italy, Economic Research and International Relations
*Contact author: giuseppe.bruno@bancaditalia.it

Keywords: Basket Options, Monte Carlo Simulation, Option pricing.

Accurate and simple pricing of basket options of European and American style can be a daunting task. Several approaches have been proposed in the literature to price path-dependent derivatives of European or American kind. They can be categorized as follows:

1) analytical approximations;
2) tree based methods;
3) Monte Carlo simulations;

Examples of the analytical approximations are provided in Milevsky and Posner 1998 [3] and [4] who compare the relative accuracy of the lognormal or the inverse gamma distribution for approximating the sum of lognormal distributions. Tree based methods were originally proposed by Cox et al 1979 [2] and adopted in Wan 2002 [5]. Monte Carlo methods were first proposed by Boyle 1977 [1] as an alternative to the closed form solution of a partial differential equation or the use of tree based methods. Monte Carlo methods can be fruitfully used to price derivatives lacking an analytical closed-form. In the more general setup, assuming a deterministic risk-free interest rate, the value of an option of European kind is given by the following expectation

$$P_t = e^{\int_t^T(-r(T-\tau)d\tau)}E_Q[g(T,S)] \tag{1}$$

In the paper we focus on the problem of pricing options on the following portfolio composed by n correlated stocks:

$$V_t = \Sigma_{i=1}^n a_i \cdot S_i \tag{2}$$

The pool of stocks follow a system of geometric brownian motion (GBM):

$$\begin{aligned} \frac{dS_1}{S_1} &= rdt + \sigma_1 dw_1 \\ \frac{dS_2}{S_2} &= rdt + \sigma_2 dw_2 \\ \frac{dS_n}{S_n} &= rdt + \sigma_n dw_n \end{aligned} \tag{3}$$

where r is the continuosly compounding risk-less interest rate, σ_i and dw_i are respectively the return volatility and the standard brownian motion driving the asset S_i. The different brownian motions are generally correlated with a given correlation matrix $\Omega = E[dw' \cdot dw]$. The Cholesky decomposition of the correlation matrix is adopted to build a recursive system for the equations of each stock composing the portfolio. Once we have computed M sample paths for our stocks we can obtain M sample path for the value of the whole portfolio. At this point we are able to evaluate the payoff function $g^s(T,V)$ for each replication s . Finally the option price will be approximated by the following mean and standard deviation:

$$\begin{aligned} \hat{C} &= \frac{1}{M}\Sigma_{s=1}^M e^{(-r(T-t))}g^s(T,V) \\ \sigma_{\hat{C}} &= \sqrt{\frac{1}{M}\Sigma_{s=1}^M (C_s - \hat{C})^2} \end{aligned} \tag{4}$$

This article shows the numerical features of a suite of R functions aimed at pricing European and American style basket options. The performances of these R functions are compared with a simulation engine equipped with an automatic FORTRAN translator for the equations describing the portfolio composition. The R functions are tested against three different sizes correlated multiasset options. Two different variance reduction techniques are also explored. In the models we have examined, the comparison among the different methods has shown a definite superiority of the control variates techniques in reducing the estimating variance. Code vectorization is going to be carried out for performance improvements

References

[1] P. Boyle. Options: a monte carlo approach. *Journal of Financial Economics*, 4:323–338, 1977.

[2] S. Ross J.C. Cox and M. Rubistein. Option pricing: a simplified approach. *Journal of Financial Economics*, 7:229–264, 1979.

[3] M.A. Milevsky and S.E. Posner. Asian options, the sum of lognormals and the reciprocal gamma distribution. *The Journal of Financial and Quantitative Analysis*, 33:409–422, 1998.

[4] M.A. Milevsky and S.E. Posner. A closed form approximation for valuing basket options. *The Journal of Derivatives*, 5:54–61, 1998.

[5] H. Wan. Pricing american style basket options by implied binomial tree. *Haas School of Business Working Paper*, 2002.

Estimating a multivariate normal covariance matrix subject to a Loewner ordering

Max Buot[1,*]

1. Xavier University
*Contact author: buotm@xavier.edu

Keywords: maximum likelihood estimation, Loewner ordering, Wishart distribution

Optimization of functions with matrix arguments appear in many statistical contexts. We consider the problem of determining the maximum likelihood estimate of the covariance matrix $\mathbf{\Sigma}$ of a multivariate normal distribution subject to simultaneous Loewner order constraints; for two positive semidefinite $\mathbf{L}$ and $\mathbf{U}$ matrices, we require that $\mathbf{\Sigma} - \mathbf{L}$ and $\mathbf{U} - \mathbf{\Sigma}$ are positive semidefinite, respectively. (The unconstrained maximum likelihood estimator is proportional to the usual sample covariance matrix.) The method described here is based on a reparametrization of the Wishart distribution. Our numerical approach compares favorably to the iterative method proposed by Calvin & Dykstra (Calvin & Dykstra, 1992). We present R implementations of both optimization algorithms, and overview further extensions to two multivariate normal samples.

References

J. A. Calvin & R. L. Dykstra (1992). A note on maximizing a special concave function subject to simultaneous Loewner order constraints. *Linear Algebra and Its Applications*, 176, 37–42.

Statistical Analysis Programs in *R* for FMRI Data

Gang Chen*, Ziad S. Saad, and Robert W. Cox
Scientific and Statistical Computing Core, National Institute of Mental Health
National Institutes of Health, Department of Health and Human Services, USA
* Contact author: gangchen@mail.nih.gov

Keywords: FMRI; group analysis; mixed-effects meta-analysis; LME; Granger causality.

1. Introduction

Open-source statistical tools are growing and evolving at a fast pace in the *R* community, presenting a valuable opportunity to incorporate frontier methodologies into data analysis in functional magnetic resonance imaging (FMRI). As FMRI data analysis usually involves 3D or 4D (3D+time) datasets in a massively univariate fashion, typical computation is quite intensive, with runtimes ranging from a few seconds to a day or two. With various state-of-art statistical packages in parallel computing using *R* [1] becoming widely available, they can be immediately beneficial to the brain imaging community. Here we present a few FMRI data analysis programs written in *R* during the past couple of years; these are available for download at http://afni.nimh.nih.gov/sscc/gangc.

2. Programs for FMRI data analysis

(1) Mixed-effects meta analysis (MEMA): **`3dMEMA`**

Data analysis in FMRI is typically carried out in two levels: effect size (linear combination of regression coefficients) and its (within-subject) variance are estimated at the 1st (individual subject) level, and group (population) inferences are performed at a 2nd level. The conventional approach only takes the effect size from the 1st level to the 2nd, assuming cross-subject variance is much larger than within-subject variance, or within-subject variance is uniform in the group [2]. In addition, the data are assumed to follow a normal distribution; thus outliers, when present, are not appropriately handled.

Here we present a computationally efficient frequentist approach [3,4] that incorporates both variance components at the group level. Specifically, we use a REML method to estimate cross-subject heterogeneity in a mixed-effects meta-analysis (MEMA) model and estimate the group effect through weighted least squares. The program handles one-, two-, and paired-sample test types, and covariates are also allowed to control for cross-subject variability. In addition to the group effect estimate and its statistic, the software also provides a cross-subject heterogeneity estimate and a chi-square test for its significance, the percentage of within-subject variability relative to the total variance in the data, and a *Z*-statistic indicating the significance level at which a subject is an outlier at a particular region.

Our approach is computationally economical, and generally more powerful and valid than the conventional method (summary statistics [2]), which ignores the effect size estimate's reliability from individual subjects. It is also relatively robust against outliers in the group data when the Gaussian assumption about the between-subject variability is relaxed by using a Laplace distribution instead [5], whose tails can better model outliers and reduce their impact on group effect and its significance.

(2) Linear mixed-effects (LME) modeling: **`3dLME`**

Using the **nlme** [6] and **contrast** [7] packages, **`3dLME`** takes effect size from each subject (at each voxel) as its only input, but allows more flexible designs than **`3dMEMA`**, such as missing data and unlimited number of categorical variables and covariates. **`3dLME`** can also handle conditions modeled with multiple time-dependent basis functions, and regression coefficients from all basis functions are analyzed in an LME model with no intercept/constant.

(3) Granger causality analysis: **`1dGC`** and **`3dGC`**

These multivariate Granger causality analysis tools use the **vars** package [8], and are suited for (but not limited to) FMRI data. The program identifies patterns of association among brain ROIs, and generates a graphic representation of the identified network. Unlike previous implementations of GC analysis for FMRI [9], this technique preserves information concerning the sign, in addition to the direction, of prior prediction (causality) between ROI time courses from individual subjects all the way to the group level analysis.

(4) Miscellaneous

`3dICA`: independent component (ICA) analysis using **fastICA** [10]
`3dICC`: intraclass correlation analysis for two- and three-way random-effects ANOVA
`3dKS`: Kolmogorov-Smirnov test using **`ks.test`**
`1dSEMr`: structural equation modeling or path analysis with covariance matrix of multiple ROIs using **sem** [11].

References

[1] R Development Core Team (2009). *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.
[2] W D Penny, and A J Holmes (2007). *Random effects analysis*. In: *Statistical Parametric Mapping* (ed. Friston, K. *et al*.). Academic Press.
[3] W Viechtbauer (2005). *Bias and efficiency of meta-analytic variance estimators in the random-effects model*. J Educ Behav Stat, 30, 261-293.
[4] W Viechtbauer (2009). metafor: Meta-Analysis Package for R. R package version 0.5-7. http://CRAN.R-project.org/package=metafor
[5] E Demidenko (2004), *Mixed Models: Theory and Applications*. Wiley-Interscience.
[6] J Pinheiro, D Bates, S DebRoy, D Sarkar and the R Core team (2009). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-96.
[7] Max Kuhn, contributions from Steve Weston, Jed Wing and James Forester (2009). *contrast: A collection of contrast methods*. R package version 0.12.
[8] Bernhard Pfaff (2008). *VAR, SVAR and SVEC Models: Implementation Within R Package vars*. Journal of Statistical Software 27(4).
[9] A Roebroeck, E Formisano, and R Goebel (2005). *Mapping directed influence over the brain using granger causality mapping*. NeuroImage, 25:230–242.
[10] J L Marchini C Heaton, B D Ripley (2009). *fastICA: FastICA Algorithms to perform ICA and Projection Pursuit*. R package version 1.1-11.
[11] John Fox with contributions from Adam Kramer and Michael Friendly (2009). *sem: Structural Equation Models*. R package version 0.9-19.

PfarMineR: An User-Friendly Expandable Front-End For Biopharmaceutical Applications with R

Yauheniya Cherkas[1], Javier Cabrera[1], Birol Emir[2], Ha Nguyen[2] And Ed Whalen[2]

[1]Rutgers University
Department of Statistics
Hill Center
Piscataway, NJ 08854, USA
ycherkas@rutgers.edu

[2]Pfizer Inc,
235 East 42nd Street
New York, NY 10017,
USA

Abstract:

R is a well-established software for statistical computing and graphics, which among other things serves as a platform for implementation of new statistical research methodology. Many statisticians who work in regulated work environments would like to access these methods but lack the training in R to do so. In order to serve this purpose we have created a package called *PfarMineR*.

PfarMineR provides a menu-driven computing environment for data manipulation and analysis. It includes basic as well as novel statistical methods with emphasis in applications to clinical data.

The default package contains four main sections – “Data”, “Exploratory Data Analysis”, “Classifications” and “Clustering”. Each section has a submenu containing its corresponding set of methods. The sections include many the popular modern techniques such as SVM, LASSO, Boosting and some Bayesian methods among others.

PfarMineR is structured in a way that it allows for updates and modifications, in particular it allows the user to (i) modify the output form and destination, (ii) change the options of a method and (iii) implement new methods.

PfarMineR is also being used for instructional purposes because it is easy to adapt to the content of a specific course.

User friendly distributed computing with R

Karim Chine, Cloud Era Ltd

Abstract: To solve heavily computational problems, there is a need to use many engines in parallel. Several tools are available but they are difficult to install and beyond the technical skills of most scientists. Elastic-R solves this problem. From within a main R session and without installing any extra toolkits/packages, it becomes possible to create logical links to remote R/Scilab engines either by creating new processes or by connecting to existing ones on Grids/clouds. Logical links are variables that allow the R/Scilab user to interact with the remote engines. rlink.console, rlink.get, rlink.put allow the user to respectively submit R commands to the R/Scilab worker referenced by the rlink, retrieve a variable from the R/Scilab worker’s workspace into the main R workspace and push a variable from the main R workspace to the worker’s workspace. All the functions can be called in synchronous or asynchronous mode. Several rlinks referencing R/Scilab engines running at any locations can be used to create a logical cluster which enables to use several R/Scilab engines in a coordinated way. For example, the cluster.apply function uses the workers belonging to a logical cluster in parallel to apply a function to a large scale R data. When used in the cloud, the new functions enable scientists to leverage the elasticity of the Infrastructure-as-a-Service to control any number of R engines in parallel.

Karim Chine, "Open Science in the Cloud: Towards a Universal Platform for Scientific and Statistical Computing", Chapter 19 in “Handbook of Cloud Computing”, Springer, 2010 (in Press)

Elastic-R, a Google docs-like portal for data analysis in the Cloud

Karim Chine, Cloud Era Ltd

Abstract: Cloud computing represents a new way to deploy computing technology, where dynamically scalable and virtualized resources are provided as a service over the Internet. Amazon Elastic Cloud (EC2) is an example of ‘Infrastructure-as-a-Service’ that anyone can use today to access infinite computing capacity on demand. This new environment enables collaboration, resources sharing and provides the tools for traceable and reproducible computational research. This model of allocating processing power holds the promise of a revolution in scientific and statistical computing. However, bringing new era for research and education still requires new software that bridges the gap between the scientist’s everyday tools and the cloud. For instance, making R available as a service in the cloud and allowing its use without any memory or computing constraints would benefit the broad population of statisticians and research professionals. This is what Elastic–R (www.elasticr.net) delivers. It provides a Google docs-like portal and workbench for data analysis that makes using R on the cloud even simpler than using it locally. Elastic-R enables scientists, educators and students to use cloud resources seamlessly, work with R engines and use their full capabilities from within any standard web browser. For example, they can collaborate in real time, create, share and reuse machines, sessions, data, functions, spreadsheets, dashboards, etc. Compute-intensive algorithms can easily be run on any number of virtual machines that are controlled from within a standard R session. Elastic-R is also an applications platform that allows anyone to assemble statistical methods and data with interactive user interfaces for the end user. These interfaces and dashboards are created visually, and are automatically published and delivered as simple web applications.

Karim Chine, "Scientific Computing Environments in the age of virtualization toward a universal platform for the Cloud" pp. 44-48, 2009 IEEE International Workshop on Open-source Software for Scientific Computation (OSSC), 2009

Karim Chine, "Open Science in the Cloud: Towards a Universal Platform for Scientific and Statistical Computing", Chapter 19 in “Handbook of Cloud Computing”, Springer, 2010 (in Press)

Eat your hashes! Hash comes to R.

Christoper Brown[1]

[1]. Open Data (http://www.opendatagroup.com)

Perl has hashes. Python has dictionaries. Hash tables and associative arrays are one of the most common and indispensable tools. A very basic programming task is to ”look up” or ”map” a key to a value. In fact, there are projects whose sole raison dtre is making the hash as fast and as efficient as possible.

R actually has two equivalents to hashes, both lacking. The first is Rs named vectors and lists. These work well enough for small lists, but because they are not a hash table, they become intolerably slow for longer lists. The second equivalent is R environments. The structure of the environment is a hash table; look-ups do not appreciably degrade with size. But R environments are not full featured and suffer from a weird syntax that is not, well, R-like.

The hash packages uses R environments and alleviates these drawbacks. The feature set rivals that of Perl and Python and the interface and great effort has been made such that all behaviors mimic those of R::Core. This talk introduces that hashes, the implementation and features of the R hash package.

R role in Business Intelligence Software Architecture

Ettore Colombo[1,*], **Gloria Ronzoni**[1], **Matteo Fontana**[1]

1. CRISP (Interuniversity Research Center on Public Services), University of Milan Bicocca
*Contact author: ettore.colombo@crisp-org.it

Keywords: Business Intelligence, Data warehousing, Pentaho, **Rserve**

During the last years, design and development of Business Intelligence Systems and Data Warehouse have become more and more challenging. The reason why mainly stands in the continuos growing of both the amount of data to elaborate and the need of more effective models for data analysis and presentation.

The current situation has brougth to the definition of new software architectures aiming to reduce the computational burden and which, at the same time, could preserve qualities of reliability, accessibility, scalability, integrability, soundness and completeness of the final application.

The work hereby presented aims to show the solution adopted at CRISP (Interuniversity Reseach Center on Public Services) and how R has been adopted. R has been integrated in the suite of tools exploited to build a Decision Support System aiming at "knowledge workers"in the public sector. For example, end-users of this system are people like decision makers and data analysts working on service and policy design in Italian local government. The technological solutions adopted at CRISP are based on the integration of different tools coming from the Open Source community. This suite is composed by software tools that cover all the layers that are required in building a data warehouse system. *Data Transformation and Preparation* are perfomed using Talend Open Studio, an open source ETL (Extraction, Transformation and Loading) and data integration platform, *Data Storing* is obtained exploiting MySQL, the well-known open source DBMS (Data Base Management System), while Pentaho BI (the world leader open source Business Intelligence platform) is deputated to *OLAP* (On-Line Analysis Processing) and *Data Presentation* (e.g. reporting and dashboarding).

The role of R in this architecture is twofold. On the one hand, R can play a central role in data elaboration according to innovative or well-known analysis model when interfaced to Talend Open Studio during ETL processes. For instance, running R directly via command-line within Talend Open Studio, it is possible to apply specific models to analyze data stored in MySQL databases and directly modify them (e.g. to determine clusters for employee careers classification).

On the other hand, R has been integrated in the Data Presentation layer by means of **Rserve**, a package that makes R accessible via TCP/IP. A Pentaho BI component that extends the existing platform, called RComponent, has been directly developed at CRISP in order to manage communication issue between R and Pentaho BI.

In particular, the obtained integrated suite has been used in *Labor*, a project aiming to analyze the labour market in different Italian provinces of Regione Lombardia in order to classify employee careers and provide the system with advanced data presentation solutions based on complex statistical models (e.g. Markov chains for predictive models).

References

Golfarelli M. , Rizzi S. (2002) Data Warehouse. McGraw Hill.

Rserve project Home Page
`http://rosuda.org/Rserve/`.

Pentaho Home Page
`http://www.pentaho.com/`.

Real-time network analysis in R using Twitter

Drew Conway

April 7, 2010

Abstract

Social networking services generate a tremendous amount of relationship data on a near continuous basis. This network data can be exploited for any number of reasons, e.g., community detection, network evolution or key actors analysis, but as as analysts we require the tools to efficiently collect this data and separate the elements of interest. Twitter is particularly well suited for this type of analysis, as its open API provides ready access to this data, which is inherently social. In this talk, I will describe how to use a combination of tools in R; specifically, twitteR and igraph, to seamlessly extract, parse, and analyze social data from Twitter in near real-time. This discussion will be accompanied by a live demonstration of this methodology using members of the audience.

Evaluating Grant Applications with Generalized Chain Block Designs in R

Dedicated to the Memory of John Mandel

Giles Crane[*], Cynthia Collins, and Karin Mille[1,2,3]

1. NJ Department of Health and Senior Services[**],
   Research Scientist, retired.
2. NJ Department of Health and Senior Services,
   Family Health Services,
   PO Box 364, Trenton, NJ 08625-0364
3. NJ Department of Health and Senior Services,
   Family Health Services,
   PO Box 364, Trenton, NJ 08625-0364

* Contact author: gilescrane@aya.yale.edu

Keywords: Chain block designs, experimental design, grant application evaluation

Among the contributors to experimental design at the National Bureau of Standards, now the National Institute of Standards and Technology, were W.J. Youden , W.S. Connor, and John Mandel. W.J. Youden introduced chain block designs for physical measurements with high precision. John Mandel enlarged this series to chain block designs with two-way elimination of heterogeneity.

Reviews of grant applications can be organized into arrangements which permit the adjustment of evaluation scores for both the level of individual reviewers and the order of review. The generalized chain block designs permit the evaluation of large number of applications with minimal number of reviewers. This systematic approach to evaluations insures fairness and efficiency within the constraints of limited resources. R enables the analyses of these special incomplete block designs. R packages base, lattice, and ggplot2 may be utilzed to graph the results. Moreover, tables of available designs, (by number of applications, reviewers, and applications per reviewer,) as well as the designs themselves can be computed with R programs. Several real evaluations will be discussed which demonstrate that chain block designs do have application in the social sciences.

References

W. J. Youden and W. S. Connor (1953). The Chain Block Design. *Biometrics,* 9, pp127-140.

John Mandel (1954). Chain Block Designs with Two-Way Elimination of Heterogeneity. *Biometrics,* June 1954, pp 251-271.

** NOTE: Anythng expressed in this paper does not represent the views or policies of the NJ Department of Health and Senior Services.

Sage and R: Using R via the Sage notebook

Karl-Dieter Crisman

Gordon College
karl.crisman@gordon.edu

Keywords: R, Sage, notebook, interface, GUI

This talk will briefly introduce Sage, and demonstrate the basics of using R through the Sage notebook user interface. We also aim to provide a starting point for deep and fruitful interactions between these two GPL-licensed programs, both of which have extensive research and educational communities.

Sage is free open-source mathematics software. It includes a huge amount of new code, as well as interfaces to other mature, high-quality open-source packages. R is one of the most well-known of these components, and yet one that is currently underutilized by most Sage users. One goal of this presentation is to change that by increasing the visibility of R to the Sage community.

However, Sage also can contribute to using R, via its 'notebook' graphical user interface. On any web browser, one can use the full power of Sage - and hence R - via input and output cells, as well as view graphics, create interactive mathlets, and do WYSIWYG text editing and LaTeX markup. This is trivial to do if one has a local installation of Sage, but can also be done remotely; since there are a number of free public Sage servers, and it is easy to host one, collaboration and pedagogy using R can be greatly enhanced. Hence, our other goal is to increase the visibility of Sage to the R community, particularly in helping Sage developers better support the most common potential uses of R via the notebook.

References

William Stein, et al. (2010). Sage: Open Source Mathematics Software, `http://www.sagemath.org/`.

Model Maker 1: Using Model Based Ontologies for Agent Based Estimation and Learning of R Packages

Jeff B. Cromwell[1*]

1. The Cromwell Workshop, Pittsburgh, PA 15217

Keywords: Modeling, Machine Learning, Neuroinformatics,data mining,BioInformatics

Model Maker is .NET 3.5 software application for statistical modeling using R on the Windows 7 platform. Model Maker was developed as a standalone bundle of services for the specification, estimation, validation and publication of advanced linear and nonlinear models for both univariate and multivariate space-time data. Because of the difficulties in writing scripts to do advance statistical analysis in statistical and mathematical packages such as R, S-Plus and MATLAB, Model Maker minimizes the amount of the programming effort needed by a researcher to conduct these types of space-time analytics. Using a visual programming technique adopted in pipeline approaches of Microsoft Robotics Studio and S-Plus Enterprise Miner along with ontologies and agents, researchers using Model Maker can drag and drop resources, i.e. data, transformations, models and published output onto a Model Sketch canvas and mix data sources, models, variables, to experiment with different research designs. Furthermore, at the core of the application is an agent based programming model that permits the use of skill-based ontologies to extract method information from CRAN packages. We present the Model Maker tool and show how modeling ontologies in both fMRI and EEG research can be used to simulate the actions and interactions of autonomous agents that lead to article publication in TeX markup.

Using R to Assess Mathematical Sense-Making in Introductory Physics Courses

Brian A. Danielak[1,*], **Andrew Elby**[1,2], **Eric Kuo**[2], **Michael M. Hull**[2], **Ayush Gupta**[2], **David Hammer**[1,2]

1. Department of Curriculum and Instruction, University of Maryland, College Park
2. Department of Physics, University of Maryland, College Park

*Contact author: briandk@umd.edu

Keywords: Teaching, Learning, Physics Education, Mathematics Education

We study university students' reasoning in physics and engineering courses. Our theoretical perspective builds from research on students' intuitive "epistemologies"—the ways they understand knowledge and knowing, specifically with respect to the role of mathematics in sense-making about physical phenomena [1]. Students in conventional courses are known to treat formulas as algorithms to use for finding answers, rather than as having conceptual meaning [4]. Students who frame mathematics as expressions of ideas, we expect, are better able to move between different kinds of representations (graphs, equations, their own intuition, diagrams, causal stories) as they conceptualize and reason about physical situations. We call such fluency a kind of "physical/mathematical sense-making" [3], and seek to support it in our instruction and assessment.

In this paper, we used R exclusively to clean, visualize, and analyze student assessment outcomes from three different instructional treatments of student sections in an introductory physics course.

- Instructor 1 was teaching a large section of students ($n = 142$) using a sense-making approach for the first time.
- Instructor 2 taught ($n = 146$) students, drawing from over 15 years of his experience developing sense-making pedagogy.
- Instructor 3 was a highly-rated instructor using a traditional approach with his students ($n = 138$)

Instructors 1 and 2 specifically encouraged their students to reason conceptually, and to create arguments and counterarguments for how physical situations might play out. All three sections were given identical final exams containing multiple choice and free-response items. We identified five out of the eight multiple choice items as "sense-making-oriented," and predicted *a priori* that Instructor 2's students would outperform students from the other two sections on that subset. We then conducted a one-way analysis of variance on the normalized five-item subscores of section 1 ($M = 0.679, SD = 0.252$), section 2 ($M = 0.723, SD = 0.243$), and section 3 ($M = 0.591, SD = 0.272$). The ANOVA was statistically significant, $F(1, 424) = 7.84$, with $p < 0.01$ at $\alpha = 0.05$. Post-hoc Tukey HSD tests indicate Instructor 1's students significantly outperformed Instructor 3's students ($d = 0.347$), and Instructor 2's students significantly outperformed Instructor 3's students ($d = 0.485$) at $\alpha = 0.05$. We further note those differences are not significant across the raw scores of all eight multiple choice items. In our presentation, we will also report results of continuing work of that dataset, including: (1) identifying epistemologically-oriented patterns in students' free-response data, and exploring free-response/multiple choice patterns within and across students. (2) Using students' scores on the Maryland Physics Expectations Survey to examine the validity and reliability of our analyses. [2].

References

[1] David Hammer and Andrew Elby, *Tapping epistemological resources for learning physics*, The Journal of the Learning Sciences **12** (2003), no. 1, 53–90.

[2] Edward F. Redish, Jeffrey M. Saul, and Richard N. Steinberg, *Student expectations in introductory physics*, American Journal of Physics **66** (1998), 212–224.

[3] Alan H Schoenfeld, *On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics*, Informal reasoning and education, 1991, pp. 311–343.

[4] Bruce L. Sherin, *How students understand physics equations*, Cognition and Instruction **19** (2001), no. 4, 479–541.

Flexible report generation and literate programming using R and Python's docutils module

Abhijit Dasgupta[1,*]

1. Principal Statistician, ARAASTAT, Germantown, Maryland 20876 *Contact author: adasgupta@araastat.com Currently a consultant at NIAMS, NIH, Bethesda MD 20892.

Keywords: Literate programming, Python, OpenOffice.org, brew, Hmisc

Literate programming has been well-established in R using `Sweave`[1]. There are two major drawbacks to `Sweave` – (i) it requires one to use LaTeX for the source, and (ii) the end-product is in a non-editable format like PDF. Many of us and our collaborators typically write using Microsoft Office or similar desktop software, and it has been difficult to produce compatible literate programming output from R. Some attempts have been made to allow this using **R2HTML**[2] and **odfWeave**[3], and some good suggestions have been made on converting `Sweave`-produced documents to Word-compatible formats (see [4]). However, the results are often not satisfactory. The requirement of using LaTeX as the source code for **Sweave** has also hindered its widespread use for automatic report generation and documentation, since LaTeX has quite a steep learning curve.

reStructured Text(rSt)[5] is a text markup syntax which is easy to read, learn and format. It can be parsed using the **docutils**[6] module of Python[7] into various formats, including LaTeX, PDF, XML, HTML, and ODF (the last using the available Python script `rst2odt`[8]). reStructured text provides a very powerful, flexible and extensible platform for creating web pages and stand-alone documents, and is the preferred method for generating Python documentation. User-defined style files can be incorporated into the parsing of the *rSt* source into the final format, making it very customizable. *rSt* files converted into ODF (OpenDocument Format) can then be read by OpenOffice.org Writer (`http://www.openoffice.org`) and by Microsoft Word using either the conversion facilities of OpenOffice.org or the Sun ODF Plugin (`http://www.sun.com/software/star/odf_plugin/`). *rSt* and **docutils** have the ability to format complex tables and incorporate figures, which are the two principal needs for a literate programming platform for R. They also can incorporate fairly involved formatting into the final document as well. Source code in LaTeX or XML or HTML can be incorporated into a *rSt* document for further formatting depending on the output format (LaTeX or ODF/XML or HTML, respectively).

This work proposes to use *reStructured text* as the source platform for literate programming in R using the templating package **brew**[9] in R, and using **docutils** scripts to convert the resulting document into ODF, PDF, LaTeXor HTML formats. Tables can easily be generated using a minor modification of `print.char.matrix` in the **Hmisc**[10] library. Figures can be incorporated either as PDF or PNG depending on the final format. The resultant ODF document appears superior to the results obtained using **Sweave** and various conversion methods, producing a professional-looking document for publication and collaboration. The resultant PDF document using default templates also produces very clean tables and figures, though different from the results of PDFLaTeXand **Sweave**. The entire process can be automated using either a Makefile or a script in R. I will present examples of using this approach utilizing the `summary.formula` scripts from **Hmisc**, which produce quite complex tables.

There are three main advantages of this approach over **Sweave** and LaTeX. First of all, *reStructured text* can be quickly generated and visually formatted using a standard text editor, and doesn't require a steep learning curve to produce well-formatted documents using **docutils** scripts. The flexibility of using source code depending on the output format for further formatting or customization or inclusion of mathematics is available. Secondly, this approach allows the direct creation of native ODF files which are readable and editable by Microsoft Word, using automatically generated (and even complex) tables and figures. This allows easy transmission of the report to collaborators or publishers who regularly use Microsoft Word or related desktop tools. Thirdly, the same source file (if it does not contain specialized source code like LaTeX or XML) can be used to produce LaTeX, PDF, XML or HTML output using scripts in **docutils**, thus allowing flexibility and further potential for customization of the output.

A R package for this approach is in development.

References

1. Leisch, F (2008) Sweave User Manual.
   `http://www.stat.uni-muenchen.de/~leisch/Sweave`

2. Lecoutre, E (2003). The R2HTML Package. R News, Vol 3. N. 3, Vienna, Austria.

3. Kuhn, M and Weaston, S (2009). odfWeave: Sweave processing of Open Document Format (ODF) files. R package version 0.7.10.

4. Harrell, FE (2010). Converting Documents Produced by Sweave.
   `http://biostat.mc.vanderbilt.edu/wiki/Main/SweaveConvert`

5. reStructuredText: Markup Syntax and Parser Component of Docutils (2006).
   `http://docutils.sourceforge.net/rst.html`

6. Docutils: Documentation Utilities: Written in Python, for General- and Special-Purpose Use.
   `http://docutils.sourceforge.net`

7. The Python Languange Reference (2010).
   `http://docs.python.org/reference/`

8. Kuhlman, D (2006) Odtwriter for Docutils.
   `http://www.rexx.com/~dkuhlman/odtwriter.html`

9. Horner, J (2007). brew: Templating Framework for Report Generation. R package version 1.0-3.

10. Harrell, FE and others (2009) Hmisc: Harrell Miscellaneous. R package version 3.7-0.
    `http://CRAN.R-project.org/package=Hmisc`

Using R for the Visualisation of Computer Experiments

Neil Diamond[1*]

[1] Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia
*Contact author: neil.diamond@buseco.monash.edu.au

Keywords: Computer Experiments, Visualisation

An important aspect of analysing computer experiments (see, for example, Santner et. al. 2003) is the visualisation of results. For example, wireframe or contour plots of the predicted response against two primary inputs can be displayed for various sets of conditioning inputs; or main effect, interaction, or joint effects plots can be displayed.

In this talk, a new R package **VizCompX** that assists in the visualisation of computer experiments is described. The package can be used stand-alone in R, or in conjunction with the Kepler workflow engine (Kepler Core, 2010) via the Nimrod suite of tools (Monash eScience and Grid Engineering Laboratory, 2010) that automates the formulation, running, and collation of the individual experiments.

References

Santner T.J., Williams, B.J., and Notz, William.I. (2003). *The Design and Analysis of Computer Experiments.* Springer: New York.

Monash eScience and Grid Engineering Laboratory (2010). The Nimrod Toolkit.,
`http://messagelab.monash.edu.au/Nimrod/`.

Kepler Core (2010). Kepler Project,
`http://www.kepler-project.org/`.

Using R in an Event Driven Service Architecture

Zubin Dowlaty[1,*], **Deepak Bysani**[2]

1. Head of Analytics Technology, Mu Sigma Inc.
2. Lead Developer, Mu Sigma Inc.

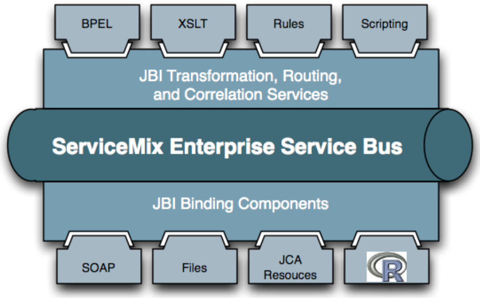
* Contact author: zubin.dowlaty@mu-sigma.com

Keywords: Enterprise Service Bus (ESB), Web Services, Real Time Streaming Data, Event Driven Architectures, Java Business Integration (JBI), Message Orientated Middleware (MOM), Real Time Predictive Scoring, High Frequency Trading (HFT)

Real time analytics is an important trend occurring within the applied analytics space. Various industry applications are finding compelling use cases for combining real time messages with an analytics engine to generate intelligent events. Examples such as real time monitoring of a supply chain, real time customer relationship management, and financial trading are areas finding traction.

The purpose of our discussion will be to review an open source implementation of an event driven architecture with R as the embedded analytics engine. We will overview the best practice architecture (see Figure 1) and demonstrate how a low latency high performance messaging engine coupled with R as the analytics engine can be created in practice. Packaging R in a JBI (Java Business Integration) container will be discussed and how the ESB (Enterprise Service Bus) can be used to integrate various business components, rules, and web services. The objective of this technology is to have R integrated into an enterprise system in order to inject advanced analytics into the message stream. A high frequency trading application using this technology will be demonstrated.

Figure 1


BPEL
XSLT
Rules
Scripting
JBI Transformation, Routing, and Correlation Services
ServiceMix Enterprise Service Bus
JBI Binding Components
SOAP
Files
JCA Resouces
R

References

Thomas Davenport (2009). *Realizing the Potential of Retail Analytics*, Babson Executive Education Working Knowledge Research Center.

Gartner Research (2007). *Hype Cycle for Business Intelligence and Performance Management*, Gartner Research.

Jeffrey A. Ryan (2009). *Real Time Market Data and Trade Execution with R*, http://cran.r-project.org/web/packages/IBrokers/vignettes/RealTime.pdf

Bruce Snyder (2007). Service *Orientated Integration with Apache Service Mix*, http://servicemix.apache.org/articles.data/SOIWithSMX.pdf

Simon Urbanek (2009). Rserve R Server, version .6-0, *http://cran.r-project.org/web/packages/Rserve/index.htm*

Read, write, format Excel 2007 (xlsx) files

Adrian A. Drăgulescu[⋆]

[⋆]Contact author: adrian.dragulescu@gmail.com

Keywords: Excel 2007, rJava, xlsx

The package `xlsx`[1] makes possible to interact with Excel 2007 files from R. While a power R user usually does not need to use Excel or even avoids it altogether, there are cases when being able to generate Excel output or to read Excel files into R is useful. For example, in an office environment where you need to collaborate with co-workers who use Excel as their primary tool. Or, to use Excel's formatting capabilities for reporting small to medium sized data sets.

The approach taken in the **xlsx** package is to use a proven, existing API between Java and Excel 2007 and use the **rJava** [2] package to link Java and R. The advantage of this approach is that the code on the R side is compact, easy to maintain and extend, even for people with little Java experience. All the heavy lifting of reading/writing xlsx files is being done in Java. The Java code used by **xlsx** is a project of the Apache Software Foundation[3]. So our approach benefits from a mature software project with many developers, test suites, and users that report issues on the Java side. While it is possible to interact directly with xlsx files from R as shown by the package **RExcelXML**[4], doing this in R is a big task that our approach avoids.

With the package **xlsx**, besides reading and writing xlsx files, you can programatically control cell properties to set text color and background color, set a cell font and data format. You can add borders, hide/unhide sheets, add/remove rows, add/remove sheets, etc. You can split panes, freeze panes, auto size columns, merge cell regions, work with cell comments and more. You can control the printer setup by setting the landscape mode, setting the number of copies, adjust the page margins, or if the copies should be in color.

References

[1] Adrian A. Drăgulescu (2010). *Read, write, format Excel 2007 (xlsx) files.* R package version 0.1.1. `http://CRAN.R-project.org/package=xlsx`

[2] Simon Urbanek (2009). *rJava: Low-level R to Java interface.* R package version 0.8-1. `http://CRAN.R-project.org/package=rJava`

[3] Apache POI project `http://poi.apache.org/`.

[4] Duncan Temple Lang (2009). *RExcelXML: Read and manipulate new-style (Office '07) Excel files.* `http://www.omegahat.org/RExcelXML/`

Rcpp: Seamless R and C++ integration

Dirk Eddelbuettel
`edd@debian.org`

Romain François
`romain@r-enthusiasts.com`

Submitted to *useR! 2010*

Abstract

The **Rcpp** package simplifies integrating C++ code with R. It provides a consistent C++ class hierarchy that maps various types of R objects (vectors, functions, environments, ...) to dedicated C++ classes. Object interchange between R and C++ is managed by simple, flexible and extensible concepts which include broad support for popular C++ idioms from the Standard Template Library (STL). Using the **inline** package, C++ code can be compiled, linked and loaded on the fly. Flexible error and exception code handling is provided. **Rcpp** substantially lowers the barrier for programmers wanting to combine C++ code with R.

We discuss and motivate the two APIs provided by **Rcpp**: the older 'classic' API introduced with the early releases of the package, and the 'new' API that results from a recent redesign. We provided simple examples that show how **Rcpp** improves upon the standard R API, demonstrate performance implications of different program designs and show how R can take advantage of modern C++ programming techniques, including template metaprogramming (TMP). The combination of modern C++ together with the interactive environment provided by R creates a very compelling combination for statistical programming.

Keywords: Foreign function interface, R, C++, Standard Template Library (STL)

RQuantLib: Bridging QuantLib and R

Dirk Eddelbuettel
edd@debian.org

Khanh Nguyen
knguyen@cs.umb.edu

Submitted to *useR! 2010*

Abstract

RQuantLib is a package for the R language and environment which connects R with QuantLib (http://www.quantlib.org), the premier open source library for quantitative finance. Written in portable C++, QuantLib aims at providing a comprehensive library spanning most aspects of quantitative finance such as pricing engines for various instruments, yield curve modeling, Monte Carlo and Finite Difference engines, PDE solvers, Risk management and more. At the same time, R has become the preeminent language and environment for statistical computing and data analysis—which are key building blocks for financial modeling, risk management and trading. So it seems natural to combine the features and power of R and QuantLib. **RQuantLib** is aimed at this goal, and provides a collection of functions for option and bond pricing, yield curve interpolation, financial markets calendaring and more.

RQuantLib was started in 2002 with coverage of equity options containing pricing functionality for vanilla European and American exercise as well as for several exotics such as Asian, Barrier and Binary options. Implied volatility calculations and option analytics were also included. Coverage of Fixed Income markets was first added to **RQuantLib** in 2005. Yield curve building functionality was provided via the DiscountCurve function which constructs spot rates from market data including the settlement date, deposit rates, futures prices, FRA rates, or swap rates, in various combinations. The function returns the corresponding discount factors, zero rates, and forward rates for a vector of times that is specified as input. In 2009, this functionality was significantly extend via the FittedBondCurve function which fits a term structure to a set of bonds using one of three different popular fitting methods ExponentialSplines, SimplePolynomial, or NelsonSiegel. It returns a data.frame with three columns date, zero.rate and discount.rate which can be converted directly into a **zoo** object and used in time series analysis or as further input for bond pricing functions. Bond pricing for zero coupon, fixed coupon, floating rate, callable, convertible zeros, convertible fixed coupon, and convertible floating coupon bonds are supported. These functions return, when applicable, the NPV, the clean price, the dirty price, accrued amount based on the input dates, yield and the cash flows of the bond.

RQuantLib is the only R package that brings the quantitative analytics of QuantLib to R while connecting the rich interactive R environment for data analysis, statistics and visualization to QuantLib. Besides providing convenient and easy access to QuantLib for R users who do not have the necessary experience in C++ to employ QuantLib directly, it also sets up a framework for users who wants to interface their own QuantLib-based functions with R.

Keywords: QuantLib, fixed income, yield curve, bond pricing, option pricing, quantitative finance, R, C++

An algorithm for Unconstrained Quadratically Penalized Convex Optimization

Steven Ellis[1]

[1] NYSPI at Columbia University

Estimators are often defined as the solutions to data dependent optimization problems. So if a statistician invents a new estimator, perhaps for an unconventional application, he/she may be faced with a numerical optimization problem. In looking for software to solve that problem the statistician may find the options few, confusing, or both.

A common form of objective function (i.e., function to be optimized) that arises in statistical estimation is the sum of a convex function and a (known) quadratic complexity penalty. A standard paradigm for creating kernel-based estimators leads to exactly such an optimization problem. Suppose that the particular optimization problem of interest is of this sort and unconstrained. Unfortunately, even generic off-the-shelf software specifically written for unconstrained convex optimization is difficult to find. So the statistician may have to fall back upon a general optimizer like BFGS, which may or may not deliver good performance on the particular problem he/she is facing.

This paper describes an optimization algorithm designed for unconstrained optimization problems in which the objective function is the sum of a non-negative convex function and a known quadratic penalty. The algorithm is described and compared with BFGS on some penalized logistic regression and penalized L$\hat{(}$3/2) regression problems.

Real-time processing and analysis of data streams

John Emerson, Michael Kane, and Bryan Lewis

This talk will propose a general framework for the processing and analysis of real-time data streams. It uses multiple R processes, shared memory, and signalling via the packages of the Bigmemory Project (http://www.bigmemory.org) and esperr (an R package in development for using Esper, which is a JAVA-based complex event manager). Examples include stock tick data and video processing.

A Simple Visualization of S & P 500 Index Performance

Nadeem Faiz, PhD[1] Contact author: nadeem_faiz@yahoo.com

1. Representing self, this work has no affiliation to author's employer, Agilent Technologies Inc.
2. Abstract submitted for Poster session in area of Visualization & Graphics

* **Keywords:** S&P Index, *R* Graphics, common investors, boxplots, polar plots

The historic performance of the Standard and Poors 500 (S&P Index) is calculated and represented visually using a simple algorithm implemented in R. The graph allows a casual observer to compare the index, on any chosen date, to the historical index performance. Information on over,under-performing the Index is easily seen.

Boxplot parameters such as `range`, `min`, `max`, `mean`, `median` and `quantile` points are calculated for each index value for various time intervals such as 30 days, 90 days, 1, 2, 3, 5 years, and so on. Box plot parameters are normalized by the range and plotted on a polar plot, with with each unit radii representing one time interval. Calculating these same changes for any one index point, and plotting on the graph allows a visual comparison.

Algorithm in pseudo code

Start with a dataset of the S&P Index from 1/3/1950 in descending chronological order

```
Index[] = read.table( S&P Index csv data file)
```

Create the visual plot of the S&P Index

```
for each interval in {30 day, 90 day, ..., 5 year} repeat the following
```

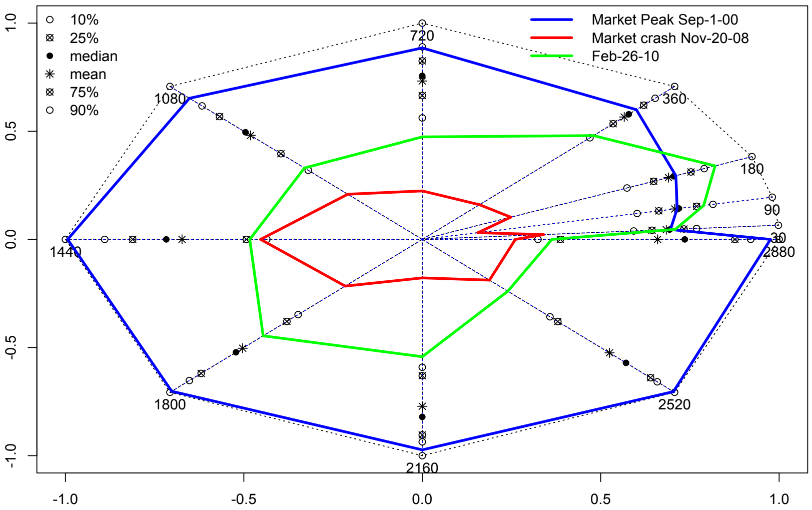
Calculating the historical performance data

```
   for each day in SandP Index, repeat the following
      change[day] = ( Index[day] - Index[day - interval] )/interval
   calculate mean(change[]) repeat for quantile(0,.1,.25,.5,.75,.9,1)
```

Creating the polar plot

```
   normalize each parameter, for eg: mean/range, median/range, etc.
   plot normalized parameters along radius at 360*interval/max(interval)
```

A Visualization of S&P 500 - from 1950 to present



References

Peter Dalgaard (2008). Introductory Statistics with R, Springer, 2nd edition, August 2008.

Yahoo Inc. (2010). S& P Index Data available at *Yahoo Finance* Website, http://finance.yahoo.com/

Note from author to reviewers:

R was well suited to crunching the S&P Index dataset (a large 800KB ascii text file). It had the graphics functionality to create polar plots from basic elements. The actual R program is relatively small, and no special library functions outside what's loaded by default in R 2.10.1 were used.

I started learning R in December 2009, this work was a simple project towards testing my progress.

I intend to use this graph to gauge if recent Index performance is approaching outliers of historical performance.

Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data

Michael P. Fay[1*]

1. National Institute of Allergy and Infectious Diseases
*email: mfay@niaid.nih.gov

Keywords: binomial tests, Blaker's exact test, exact McNemar's test, Fisher's exact test, Poisson test

There is an inherent relationship between two-sided hypothesis tests and confidence intervals. A series of two-sided hypothesis tests may be inverted to obtain the *matching* 100(1-α)% confidence interval defined as the smallest interval that contains all point null parameter values that would not be rejected at the α level. Unfortunately, for discrete data there are several different ways of defining two-sided exact tests, and the most commonly used two-sided exact tests are defined one way, while the most commonly used exact confidence intervals are inversions of tests defined a different way. This can lead to inconsistencies where the exact test rejects but the exact confidence interval contains the null parameter value. The R packages **exactci** and **exact2x2** provide several exact tests with the matching confidence intervals avoiding these inconsistencies as much as is possible. Examples are given for binomial and Poisson parameters and the paired and unpaired 2×2 tables.

Deducer: A Graphical useR interface for everyone

Ian Fellows

While R has proven itself to be a powerful and flexible tool for data exploration and analysis, it lacks the ease of use present in other software such as SPSS and MINITAB. An easy to use graphical interface can help new users accomplish tasks that would be out of their reach otherwise, and can improve the efficiency of expert users by replacing fifty key strokes with five mouse clicks. With this in mind, Deducer presents dialogs that are as understandable as possible for the beginner to understand, yet contain all (or most) of the options that an experienced statistician performing the same task would want. An Excel-like spreadsheet is included for easy data viewing and editing. Deducer is based on Java's Swing GUI library, and can be used on any common operating system. The GUI is independent of the specific R console and can easily be used by calling a text based menu system. Graphical menus are provided for the JGR console and the Windows RGui. In this talk we will explore the GUI's data manipulation and analysis using real world data.

mritc—A package for MRI tissue classification

Dai Feng, Merck & Co., Inc.
Luke Tierney, The University of Iowa

MRI tissue classification (segmentation) is a process during which anatomically meaningful labels are assigned to each voxel of an MR image. It is critical in the study of brain disorders such as Alzheimer's disease.

The package provides several functions to conduct MRI tissue classification. The methods include using the normal mixture model fitted by the EM algorithm, the hidden Markov normal mixture model at the voxel level fitted by the Iterated Conditional Mode algorithm, the Hidden Markov Random Field EM algorithm, or the Bayesian method, the higher resolution model fitted by Markov chain Monte Carlo, and the Gaussian partial volume hidden Markov random field model fitted by the modified EM algorithm. Initial values for different approaches can be obtained through multilevel thresholding using a fast algorithm for Otsu's Method.

For the data, the "Analyze", "NIfTI", and raw byte file formats are supported. Facilities for visualization of classification results and various performance evaluation indices are provided.

To improve speed, table lookup methods are used in various places and some computations are performed by embedded **C** code and **C** using Open MP to parallelize loops.

Estimation of the Area-Under-the-Curve of Mycophenolic Acid using population pharmacokinetic and multi-linear regression models simultaneously.

Michał J. Figurski[1,*] , Leslie M. Shaw[1]

1. Biomarker Research Laboratory, Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA.
* Contact author: figurski@mail.med.upenn.edu

Keywords: Mycophenolic Acid, pharmacokinetics, NONMEM, multi-linear regression model

Mycophenolic Acid is an immunosuppressive drug administered to patients after organ transplantation in order to prevent rejection of the transplanted organ. Because of the complicated pharmacokinetics of this drug, as well as its side effects, patient exposure to the drug is often monitored [1]. The best available pharmacokinetic parameter to be monitored is the Area Under the Curve (AUC) on the plot of measured drug concentrations versus time, over the entire dosing interval. This parameter is tedious to achieve analytically, because it requires multiple blood draws and extended patient stay in the facility. Several methods exist, that allow estimation of AUC using fewer blood draws obtained in a shorter time interval, as compared to full dosing-time interval AUC [2].

In this study we simultaneously used two techniques: multi-linear regression models developed using SAS [2], and population-pharmacokinetic models developed using NONMEM [3] to estimate the AUC of mycophenolic acid in a semi-automated way.

We have created an *R* script that automatically loads patient data stored in a 'csv' file, prepares the dataset for *NONMEM* simulation and invokes a *NONMEM* run. Next, it loads the output from simulation and estimates the mycophenolic acid AUC in two ways: using an appropriate multi-linear regression equation and using `trapz` function (**caTools** library) from *NONMEM*-simulated concentrations. Finally, the script produces a graphical representation of the results of simulation, with embedded table of results (**plotrix** library) and all important patient characteristics. This is all performed without any user interaction neither with *R* nor with *NONMEM.*

Our script is suitable for use in any clinical toxicology laboratory by unqualified personnel, who would only be required to enter the data into a spreadsheet and run the script. The only prerequisite for this method to work is a *NONMEM* license. Our script provides the user with results of estimation of the Area Under the Curve using two methods, quickly and with minimum effort.

References

1. L.M. Shaw, M. Figurski, M.C. Milone, J. Trofe, R.D. Bloom. *Therapeutic drug monitoring of mycophenolic acid.* Clin J Am Soc Nephrol, 2007, 2(5):1062-72
2. M.J. Figurski, A. Nawrocki, M.D. Pescovitz, R. Bouw, L.M. Shaw, *Development of a predictive limited sampling strategy for estimation of mycophenolic acid AUC in patients receiving concomitant Sirolimus or Cyclosporine.* Ther Drug Monit, 2008, 30(4), 445-55
3. M.J. Figurski, L.M. Shaw. *Estimation of Mycophenolic Acid AUC by means of population pharmacokinetics and multi-linear regression.* Ther Drug Monit, 2009, 31(5), 655 (abstract)

How to Effectively Visualize & Quantify a 3-D Image (or How Long Do We Mix to Get Homogeneous Product)?

Tom Filloon (Statistics), Dave Dunlop (Beauty)   Procter & Gamble - Cincinnati, Ohio

Abstract

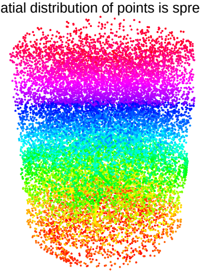
To determine if we have a homogeneous product after mixing, we need to understand the spatial (3D) distribution of particles and how to evaluate whether particles are randomly dispersed or not. It will be shown how nearest neighbor distances can be used for evaluating product homogeneity.

References

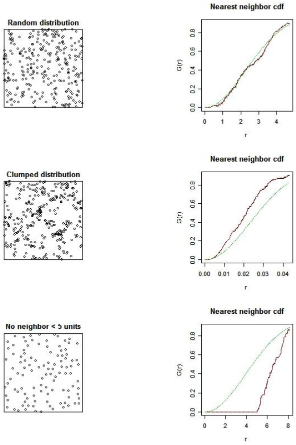
1. Ripley, B **Spatial Statistics**.
2. Cressie, N **Statistics for Spatial Data**
3. TG Filloon SLR April 2007

Introduction

How long does one need to mix a large mixing tank before the resulting product is homogeneous? Computer simulations are run to perform virtual mixing experiments. A large number of virtual particles (~20,000) are tracked over time while a computer simulates how they would move in a large cylindrical mixing tank in operation [*see picture below*]. Then after a given amount of time, one can visualize the spatial distribution of points by knowing their exact locations. The 'statistical' task then is to somehow quantify this 3-dimensional data into a summary measure as to whether the spatial distribution of points is spread out enough.



Spatial Statistics – Point Processes

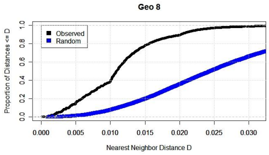


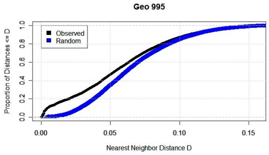
One approach to analyzing such data is to determine the average number of points within radius r of a random point, which yields a function K(r). If the K function is too large for small values of r (relatively to random Poisson process), this would indicate clustering /clumping of data.

Another more intuitive measure (**that will be used here**) is to determine, for each data point, the distance to its 'nearest neighbor' and then to compare this distribution of nearest neighbor distances to what would be expected if the data were from a random Poisson process. We will construct the empirical cumulative distribution function of these nearest neighbor distances and define it as G(r). Hence, G(r) represents the proportion of data points whose nearest neighbor is less than or equal to r units away, and will take the classical stair-step form.

3-D Nearest Neighbor Distributions

2 Simulations
– notice how Geo 995 simulation has been better mixed (closer to random distribution)





Computational Approach

Fortunately, via R-news (*thanks go to NHH Roger Bivand, for making me aware of UMD David Mount's work*) , I was able to get an off-CRAN R package (ann) that does inter-point distances calculations efficiently (C code) and hence, this implementation is done via the R software. An additional R package (rgl) allows for interactive 3-D plotting of the points to enable visualization of data and this statistical summary all in the same software. Furthermore, since the R software is free, I have loaded all of this capability onto the collaborator's desktop for his ease of use as this project moves forward.

Summary

A user-friendly, fast approach for interactive visualization & understanding of a spatial distribution of points was freely available within the R software (libraries *rgl, ann*). This approach is simple and generalizable to any 3-D (or 2-D!) problems.

RProtoBuf: Protocol Buffers for R

Romain François
romain@r-enthusiasts.com

Dirk Eddelbuettel
edd@debian.org

Submitted to *useR! 2010*

Abstract

Protocol buffers are a flexible, efficient, automated mechanism for serializing structured data—think XML, but smaller, faster, and simpler. Users define how they want the data to be structured once in a `proto` file and then use special generated source code to easily write and read structured data to and from a variety of data streams and using a variety of officially supported languages— Java, C++, or Python—or third party implementations for languages such as C#, Perl, Ruby, Haskell, and now R via the **RProtoBuf** package.

The **RProtoBuf** package implements R bindings to the C++ protobuf library from Google. It uses features of the protocol buffer library to support creation, manipulation, parsing and serialization of protocol buffers messages. Taking advantage of facilities in the **Rcpp** package, **RProtoBuf** uses S4 classes and external pointers to expose objects that look and feel like standard R lists, yet are managed by the underlying C++ library. These objects also conform to the language-agnostic definition of the message type allowing access to their content from other supported languages.

As the protocal buffers library does not offer any built-in support for networked access to protocol buffer streams, we intend to take advantage on ongoing changes to the R system to expose a native R server. This is work-in-progress, but we hope to be able to report on this at the conference.

Keywords: R, C++, Serialization, Data Interchange, Data Formats

Blogging about R

Tal Galili

This talk is a basic introduction to blogs: why to blog, how to blog, and the importance of the R blogosphere to the R community.

Because R is an open-source project, the R community members rely (mostly) on each other's help for statistical guidance, generating useful code, and general moral support.

Current online tools available for us to help each other include the R mailing lists, the community R-wiki, and the R blogosphere. The emerging R blogosphere is the only source, besides the R journal, that provides our community with articles about R. While these articles are not peer reviewed, they do come in higher volume (and often are of very high quality).

According to the meta-blog **www.R-bloggers.com,** the (English) R blogosphere has produced, in January 2010, about 115 "articles" about R. There are (currently) a bit over 50 bloggers who write about R, with about 1000 subscribers who read them daily (through e-mails or RSS). These numbers allow me to believe that there is a genuine interest in our community for more people - perhaps you? - to start (and continue) blogging about R.

In this talk I intend to share knowledge about blogging so that more people are able to participate (freely) in the R blogosphere - both as readers and as writers. The talk will have three main parts:

1) **What is a blog**
2) **How to blog** – using the (free) blogging service **WordPress.com** (with specific emphasis on R)
3) **How to develop readership** - integration with other social media/networks platforms, SEO, and other best practices

Also, my intention is to tailor the talk according to audience needs. If you are considering attending the talk, please e-mail me at **tal.galili@gmail.com** (with the subject line "blogging about R") so that I can receive your feedback through a short survey and also send you an update with the final talk outline.

* * *

Tal Galili founded www.R-bloggers.com and blogs on www.R-statistics.com

Analyzing the Operational RNA Code for Amino Acids - Using R

Tal Galili[1], Shaul Shaul[2], Yoav Benjamini[3]

It has been suggested that the nucleotide sequence of the tRNA acceptor stems specify an operational RNA code for amino acids. In the last 20 years several attributes of the putative code have been elucidated for a small number of model organisms. To gain insight about the ensemble attributes of the code, we used R to analyze the 6350 tRNA sequences from 102 Bacterial and 55 Archaeal species.

Our study found that the RNA codes, in both Archaea and Bacteria, differs substantially from the genetic code in the degree of degeneracy and specificity. We found instances of taxon-specific alternative codes, i.e., identical acceptor stems encrypting different amino acids in different species, as well as instances of ambiguity, i.e., identical acceptor stems encrypting two or more amino acids in the same species. Further investigation revealed that the species-specific degree of difference in the RNA code holds useful information, allowing us to distinguish between Archaea originating from different ecological environments and conditions.

In our research we employed various existing R facilities, performing data importing (Biostrings), cleaning and preparation (reshape), classification and regression tree (rpart), clustering (cluster), visualization (lattice) etc.
Also, we self developed (or implemented), in R, algorithms for tasks such as the comparing of hierarchical clusters, the creating of a distance matrix between Archaeas based on the code in their tRNA acceptor stems, performing simulation studies testing how well the "uniqueness property" is held by the organisms, based on their tRNA code.

In this talk I will provide (a brief) biological background needed in order to understand the above discoveries, and present how we used R various packages and statistical methods, while devising and implementing new methods (in R), in order to support our investigation.

Reference

Shaul S, Berel D, Benjamini Y, Graur D.(2010), **"Revisiting the operational RNA code for amino acids: Ensemble attributes and their implications", RNA.** 2010 Jan;16(1):141-53. Epub 2009 Dec 1.,

Correspondence should be addressed to: Tal.Galili@gmail.com

[1] Department of Statistics and Operations Research, Tel Aviv University, Israel.
[2] Department of Zoology, Tel Aviv University, Israel.
[3] Department of Statistics and Operations Research, Tel Aviv University, Israel.

R to LaTeX / HTML

Christophe Genolini[1,2,3,⋆], **Bernard Desgraupes**[3], **Lionel Riou França**[1,2,4]

1. Inserm, U669, Paris, France
2. Univ. Paris-Sud and Univ. Paris Descartes, UMR-S0669, Paris, France
3. Modal'X, Univ. Paris Ouest Nanterre La Défense, France
4. Phisquare: Phisquare Institute, 75001 Paris

⋆Contact author: genolini@u-paris10.fr

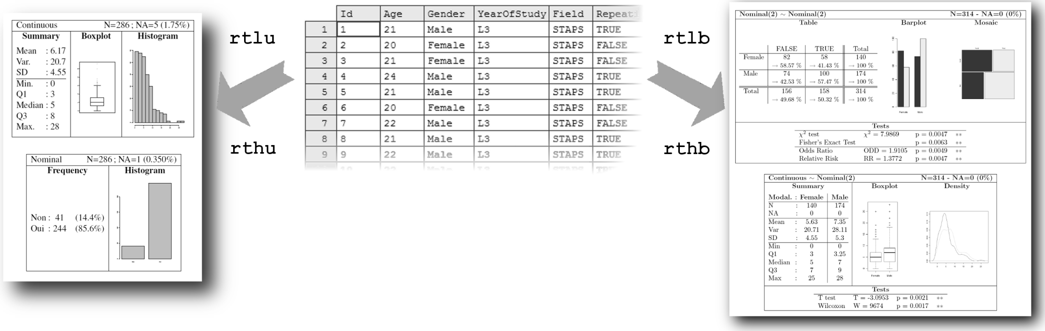
Keywords: Interface to other languages, Univariate analysis, Bivariate analysis, LaTeX, HTML

The package **r2lh** (R to LaTeX or HTML) provides facilities to export some R analysis (univariate and bivariate) in a LaTeX or HTML format. The main goal of this package is to facilitate the preliminary exploration of data by running automaticaly some classical analysis.

Univariate analysis (functions `rtlu` and `rthu`) describes a single variable X. According to X nature (**nominal**, **ordinal**, **discrete** -numeric with few modalities- or **continuous** -numeric with many modalities-), it computes frequencies, mean, standard deviation, quartiles, boxplot, barplot and histogram. The univariate analysis has been presented in [1].

Bivariate analysis (function `rtlb` and `rthb`) considers two joint variables $Y \sim X$. A first part is descriptive: frequencies, mean, standard deviation, quartiles of Y relatively to each modalities of X when applicable, juxtaposed boxplot, barplot, densities. A second part is more about testing the existence of a link between Y and X. It computes both parametric and non-parametric tests, trying many possible test. It also display some graphics helping the user to decide which test is more relevant (qqplot, densities...)

To illustrate its way of working, **r2lh** uses a data set resulting from enquiries led by some second year students of the University of Paris Ouest that had decided to investigate the "Exam Cheating in French Universities" [2].



Continuous N=286 ; NA=5 (1.75%)

Summary

Mean	: 6.17
Var.	: 20.7
SD	: 4.55
Min.	: 0
Q1	: 3
Median	: 5
Q3	: 8
Max.	: 28

Nominal N=286 ; NA=1 (0.350%)

Frequency

Non :	41	(14.4%)
Oui :	244	(85.6%)

	Id	Age	Gender	YearOfStudy	Field	Repeati
1	1	21	Male	L3	STAPS	TRUE
2	2	20	Female	L3	STAPS	FALSE
3	3	21	Female	L3	STAPS	FALSE
4	4	24	Male	L3	STAPS	TRUE
5	5	21	Male	L3	STAPS	TRUE
6	6	20	Female	L3	STAPS	FALSE
7	7	22	Male	L3	STAPS	FALSE
8	8	21	Male	L3	STAPS	TRUE
9	9	22	Male	L3	STAPS	TRUE

Nominal(2) ~ Nominal(2) N=314 - NA=0 (0%)

Table

	FALSE	TRUE	Total
Female	82 → 58.57 %	58 → 41.43 %	140 → 100 %
Male	74 → 42.53 %	100 → 57.47 %	174 → 100 %
Total	156 → 49.68 %	158 → 50.32 %	314 → 100 %

Tests

χ^2 test	$\chi^2 = 7.9869$	p = 0.0047	**
Fisher's Exact Test		p = 0.0063	**
Odds Ratio	ODD = 1.9105	p = 0.0049	**
Relative Risk	RR = 1.3772	p = 0.0047	**

Continuous ~ Nominal(2) N=314 - NA=0 (0%)

Summary

Modal. :	Female	Male
N :	140	174
NA :	0	0
Mean :	5.63	7.35
Var :	20.71	28.11
SD :	4.55	5.3
Min :	0	0
Q1 :	3	3.25
Median :	5	7
Q3 :	7	9
Max :	25	28

Tests

T test	T = -3.0953	p = 0.0021	**
Wilcoxon	W = 9674	p = 0.0017	**

References

1 Christophe Genolini, Lionel Riou Frana (2009) "R to LaTeX, Univariate Analysis" *in proceedings, UseR! 2009, The R User Conference 2009*, (Agrocampus-Ouest, Rennes, France), July 2008, pp 60.

2 Christophe Genolini (2007). "EPO2007: Exam cheating at French University" `http://christophe.genolini.free.fr/EPO/EPO2007-Fraude.php`

Introducing computational thinking with free software in a math for liberal arts course

Panayotis Giannakouros[1,2,★], **Lihua Chen**[1]

1. James Madison University
2. University of Missouri–Kansas City
★Contact author: giannapx@jmu.edu

Keywords: Teaching, iPoIDE, Emacs

This poster reports on incorporating a suite of free software, including R, into an introductory math class for the liberal arts. The course was designed to give an appreciation of mathematics to students with no college math background in majors that may not require them to take another math class. In this case, the flexible course goals and the features of **GNU Emacs** and free software were combined to promote learning relevant to computational thinking.

GNU Emacs is a powerful text editor that has been actively developed for over 25 years. It is built around a full programing language, Emacs Lisp, and ported to every major computing platform [1]. Beyond its text editing capabilities, **Emacs** is capable of serving as a platform-independent computing environment. This computing environment combined with the free software licenses under which many powerful programs are released has made possible the pre-configured portable cross platform software suite used in this course [2]. With this suite, setup for the course was reduced to copying a software folder to a student accessible drive.

In the course, computer use was motivated by the topics through which the course evolved. We turned to the computers to illustrate concepts, to work on puzzles, and as means to write up assignments. After the students accepted the unfamiliar computing interface, they passed among a number of programs in the software environment seamlessly and with steadily increasing confidence.

In conclusion, the course required minimal computing setup, had good learning outcomes, and course evaluations were excellent. The approach taken in this course could be incorporated into an introductory statistics course or developed into a course on computational thinking.

References

[1] Free Software Foundation, *GNU Emacs–GNU project–Free Software Foundation(FSF)*, http://www.gnu.org/software/emacs, 1996–2010.

[2] Panayotis Giannakouros, *Statlive*, http://www.statlive.org, 2007–2010.

Designing a Flexible GUI for R

Sheri Gilley[1,2,*]

1. REvolution Computing
*Contact author: sheri@revolution-computing.com

Keywords: GUI, user interface,UI design

One of REvolution's main goals is to make data analysis done in R accessible to a wider community of people. To that end, we're in the process of designing and building a graphical user interface for R. The goals of our user interface are the following:

1. Support the entire workflow of data analysis
2. Work on all platforms
3. Easy to use for a person who does not know how to program in R
4. Aid in learning for a beginner in R programming
5. Easy to extend for someone who is an experienced R programmer

We developed personas and use cases, followed by a series of prototypes. Prototypes were then presented to various potential users and evaluated. At this talk I will present some of the more important use cases and show how they were satisfied in the prototype, as well as a demo our progress in building of this new user interface.

bild: a package for BInary Longitudinal Data

M.Helena Gonçalves[1,2,⋆], **M.Salomé Cabral**[1,3], **Adelchi Azzalini**[4]

1. Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
2. Departamento de Matemática, FCT, Universidade do Algarve, Portugal
3. Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Portugal
4. Dipartimento di Scienze Statistiche, Università di Padova, Italy

⋆Contact author: mhgoncal@ualg.pt

Keywords: Binary longitudinal data, Exact likelihood, Marginal models, Markov Chain, Random effects.

The software tools that we propose are aimed at the analysis of binary longitudinal data from the point of view of likelihood inference, which requires complete specification of a stochastic model for the individual profile. Denote by y_{it} $(t = 1,\dots,T_i) \in \{0,1\}$ the response value at time t from subject i $(i = 1,\dots,n)$, and by Y_{it} its generating random variable whose mean values is $\mathbb{P}\{Y_{it} = 1\} = \theta_{it}$. Associated to each observation time and each subject, a set of p covariates, x_{it}, is available. In our formulation the parameter of interest is the marginal probability of success, that is related to the covariates via a logistic regression model,

$$\operatorname{logit} \mathbb{P}\{Y_{it} = 1\} = x_{it}^{\top}\beta. \tag{1}$$

The dependence structure of the process corresponds to a second order Markov Chain. This set-up leads to consideration of the joint distribution of three components of the process at the time, (Y_{t-2}, Y_{t-1}, Y_t) say. Our choice is to impose the constraints

$$OR(Y_{t-1}, Y_{t-2}) = \quad \psi_1 \quad = OR(Y_{t-1}, Y_t) \tag{2}$$

$$OR(Y_{t-2}, Y_t|Y_{t-1} = 0) = \quad \psi_2 \quad = OR(Y_{t-2}, Y_t|Y_{t-1} = 1). \tag{3}$$

where ψ_1 and ψ_2 are two positive parameters. In the context of binary processes, dependence is more conveniently measured by odds ratios rather than correlations, and conditions (2)–(3) provide a parametrisation whose interpretation is similar to the partial autocorrelation of a Gaussian process, transferred to the odd-ratio scale. The problem is finding the $p_{hj} = \mathbb{P}\{Y_t = 1|Y_{t-2} = h, Y_{t-1} = j\}, h, j = 0, 1$, satisfying the above-stated conditions, see [1] for details. This software allows the presence of individual random effects by adding the component $b_i \sim N(0, \sigma^2)$ in (1) leading to the logistic model with random intercept

$$\operatorname{logit} \mathbb{P}\{Y_{it} = 1\} = x_{it}^{\top}\beta + b_i, \tag{4}$$

where the b_i's are assumed to be sampled independently from each other. We reparameterise $\omega = \log\sigma^2$ both for numerical convenience and to improve accuracy of the asymptotic approximation to the distribution of MLEs. One dimensional integrals are computed using adaptive Gaussian quadrature. A frequent problem with longitudinal studies is the presence of missing data, this software allows for a simple pattern of missing data, details available in [2]. We considered a form of residuals of a fitted model, to be used for diagnostic purposes. Graphical analysis of residuals is difficult even for the simple case of logistic regression of binary data, due to the extreme discreteness of the binary data. To alleviate the problem of discreteness, we aggregate residuals across individuals at each given time point. The package, called **bild**, is a S4-methods package and provides R functions for parametric and graphical analysis of binary longitudinal data. The functions of **bild** have been written in R language, except for some FORTRAN routines which are interfaced through R. The main function performs the fit of parametric models via likelihood methods. Serial dependence and random effects are allowed according to the stochastic model chosen: independence, MC1 (1st order Markov Chain), MC2 (2nd order Markov Chain), MC1R (1st order Markov Chain with random effects) or MC2R (2nd order Markov Chain with random effects). Missing values and unbalanced data are automatically accounted for computing the likelihood function. Six plots are available in plot methods: Residuals vs Fitted, Residuals vs Time, ACF residuals, PACF residuals, Parametric fit and Individual mean profiles.

References

[1] M. Helena Gonçalves and Adelchi Azzalini (2008). Using Markov chains for marginal modelling of binary longitudinal data in an exact likelihood approach. *Metron*, LXVI, 157–181.

[2] M. Helena Gonçalves (2002). *Likelihood methods for discrete longitudinal data.* PhD thesis, University of Lisbon.

PMML Execution of R Built Predictive Solutions

Alex Guazzelli[1*] , Kostantinos Stathatos[1] , Michael Zeller[1]

1. Zementis, Inc.

* Contact author: Alex.Guazzelli@zementis.com

Keywords: PMML, Model Deployment, ADAPA, Predictive Solutions, Predictive Applications

The rule in the past was that whenever a model was built in a particular development environment, it remained in that environment forever, unless it was manually recoded to work somewhere else. This rule has been shattered with the advent of *PMML* (Predictive Modeling Markup Language). Defined as an XML-based language used to represent predictive data mining models, it was specified by the Data Mining Group, an independent group of leading technology companies. By providing a uniform standard to represent predictive models, *PMML* allows for the exchange of predictive solutions between different applications and various vendors. The *R PMML* package, which is currently available through CRAN (the Comprehensive *R* Archive Network), exports *PMML* for a variety of modeling techniques which include: neural network models, support vector machines, decision trees, regression models, association rules and clustering models. Besides *R*, many statistical tools also support the standard; these include, for example, tools from *KNIME, SAS, IBM/SPSS*, and *TIBCO*.

Once exported as *PMML* files, models are readily available for deployment into an execution engine for scoring or classification. *ADAPA* is one example of such an engine. It takes in models expressed in *PMML* and transforms them into web-services. Models can be executed either remotely by using web-services calls, or via a web console. Users can also use an Excel add-in to score data from inside Excel using models built in *R*.

R models have been exported into *PMML* and uploaded in *ADAPA* for many different purposes. Use cases where clients have used the flexibility of *R* to develop and the *PMML* standard combined with *ADAPA* to deploy range from financial applications (e.g., risk, compliance, fraud) to energy applications for the smart grid. The ability to easily transition solutions developed in *R* to the operational IT production environment helps eliminate the traditional limitations of *R*, e.g. performance for high volume or real-time transactional systems and memory constraints associated with large data sets.

References

A. Guazzelli, K. Stathatos, and M. Zeller (2009). Efficient Deployment of Predictive Analytics through Open Standards and Cloud Computing. *SIGKDD Explorations Newsletter*, 11/1, 32-38.

A. Guazzelli, M. Zeller, W. C. Lin, G. Williams (2009). PMML: An Open Standard for Sharing Models. *The R Journal*, 1/1, 60-65.

Data Mining Group (2009). *PMML version 3.2*, http://www.dmg.org/pmml-v3-2.html.

G. Williams, M. Harshler, A. Guazzelli, M. Zeller, W. Lin, H. Ishwaran, U. B. Kogalur, and R. Guha. (2009). *PMML: Generate PMML for various models*. http://rattle.togaware.com/. R package version 1.2.7.

R. Pechter (2009). What's PMML and What's New in PMML 4.0? *SIGKDD Explorations Newsletter*, 11/1, 19-25.

Parallelizing a Computationally Intensive Financial R Application with Zircon Technology

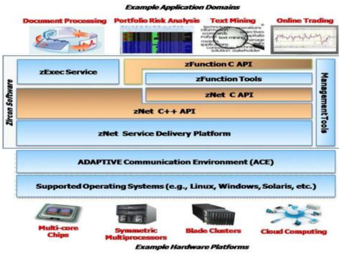
By: Ron Guida, Director, Zircon Computing, LLC

Abstract

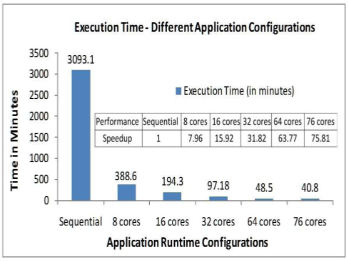
Statisticians, analysts, scientists, and engineers require massive processing power to conduct data analysis, predictive modeling, visualization, and other complex tasks. This poster describes how Zircon substantially improved the performance of a representative complex computational finance application by integrating the Zircon adaptive ultra high performance computing software platform and tools with the R programming language and environment. This integrated solution uses distribution and parallelization to reduce the total computation time of the R-based application from 3,093 minutes to 40 minutes on an off-the-shelf, commodity multiprocessing platform.

The poster presentation shall describe the case study in which computationally intensive financial models built by Garrett Asset Management (GAM), written in the R programming language, were optimized using the Zircon ultra high performance software and UltraCloud™ enabled in the IBM CoD resource center, to gather linear performance results, while also providing ease-of-use and ease-of-deployment functionality to the GAM user. Solution topology and performance results are shown in detail.

Zircon Software Topology

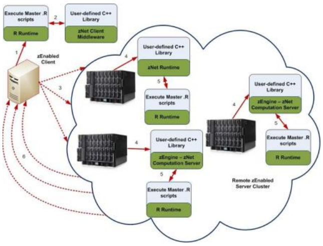

Example Application Domains
Document Processing
Portfolio Risk Analysis
Text Mining
Online Trading
Zircon Software
zExec Service
zFunction C API
zFunction Tools
zNet C API
zNet C++ API
zNet Service Delivery Platform
Management Tools
ADAPTIVE Communication Environment (ACE)
Supported Operating Systems (e.g., Linux, Windows, Solaris, etc.)
Multi-core Chips
Symmetric Multiprocessors
Blade Clusters
Cloud Computing
Example Hardware Platforms

Performance Results



Performance	Sequential	8 cores	16 cores	32 cores	64 cores	76 cores
Speedup	1	7.96	15.92	31.82	63.77	75.81

UltraCloud™ Solution


Execute Master .R scripts
R Runtime
2
User-defined C++ Library
zNet Client Middleware
1
zEnabled Client
3
4
User-defined C++ Library
zNet Runtime
5
Execute Master .R scripts
R Runtime
4
User-defined C++ Library
zEngine – zNet Computation Server
5
Execute Master .R scripts
R Runtime
6
4
User-defined C++ Library
zEngine – zNet Computation Server
5
Execute Master .R scripts
R Runtime
Remote zEnabled Server Cluster

Using Rwave to detect synchrony of influenza between U.S. states

Christian Gunning[1,*]

1. Biology Department, University of New Mexico
*Contact author: xian@unm.edu

Keywords: Spatio-temporal, Wavelets, Time series, Rwave, Influenza

The continuous wavelet transform (CWT) is a powerful tool for analyzing non-stationary spatio-temporal data. Compared to the discrete wavelet transform (DWT), the CWT offers a well-defined relationship between scale and frequency, and a time-scale decomposition that is independent of translations and truncations of the original time series [1]. To our knowledge, **Rwave** is the only active R package that implements the CWT.

Influenza epidemics in temperate regions display marked yearly seasonality, with peak incidence occurring during the winter months. The spatio-temporal course of yearly epidemics has significant public health consequences, but is non-stationary and difficult to predict. Here, I use the **Rwave** package to investigate synchrony of influenza infections between U.S. states over 6 years. I use weekly estimates of influenza incidence published by Google [4,5]. I compute the cross-wavelet spectrum between all states with complete records, excluding Hawaii (36 states total), and examine the distribution of phases at frequencies around 1 year. I use a Wilcox test to determine the significance of these phase differences, and estimate the difference in phase from zero using the `wilcox.test`'s pseudomedian estimate. I then examine latitude, longitude, and population of the comparison states as predictors of estimated phase difference using linear regression. Of these, latitude is the only significant predictor of phase difference.

Previous work has inferred synchrony of influenza primarily from timing of epidemic peak (e.g. [2] and [3]). The CWT, as used here, allows inspection of synchrony over the full period of record. As such, this method accounts for synchrony of epidemic onset and decay, as well as inter-epidemic dynamics, and allows a richer exploration of underlying disease dynamics.

References

[1] B. Cazelles, M. Chavez, D. Berteaux, F. Ménard, J.O. Vik, S. Jenouvrier, and N.C. Stenseth, *Wavelet analysis of ecological time series*, Oecologia **156** (2008), no. 2, 287–304.

[2] K.M.L. Charland, D.L. Buckeridge, J.L. Sturtevant, F. Melton, B.Y. Reis, K.D. Mandl, and J.S. Brownstein, *Effect of environmental factors on the spatio-temporal patterns of influenza spread*, Epidemiology and Infection **137** (2009), no. 10, 1377–1387.

[3] B.S. Finkelman, C. Viboud, K. Koelle, M.J. Ferrari, N. Bharti, and B.T. Grenfell, *Global patterns in seasonal activity of influenza A/H3N2, A/H1N1, and B from 1997 to 2005: viral coexistence and latitudinal gradients*, PLoS One **2** (2007), no. 12.

[4] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant, *Detecting influenza epidemics using search engine query data*, Nature **457** (2008), no. 7232, 1012–1014.

[5] google.org, *Google flu trends*, 2009.

An Excel Interface for Functions in the **metRology** Package

William F. Guthrie[1,*], Hung-kung Liu[1]

1. Statistical Engineering Division, National Institute of Standards and Technology
* Contact author: will.guthrie@nist.gov

Keywords: software deployment,

This talk describes the development and use of an Excel interface developed using *RExcel* for the functions in the new **metRology** package for statistical metrology. The interface uses a combination of *Visual Basic* macros and *RExcel* worksheet and macro functions to allow the user to access templates for different statistical analyses that are common in metrological work. This interface makes it easy for scientists, engineers, and metrologists who need to carry out statistical analyses to harness the power of *R* without the high overhead costs associated with learning a new software package. The software will be illustrated using examples from NIST work. Lessons learned when the software was used during hands-on workshops for uncertainty analysis in Argentina and Brazil will also be discussed.

References

Thomas Baier and Erich Neuwirth (2007) Excel :: COM :: R, *Computational Statistics* 22/1, pp. 91-108.

Export pivot table to R using RExcel

Keith Halbert, Richard M. Heiberger, Erich Neuwirth

Keywords: Microsoft Excel, R Through Excel, RExcel, pivot tables

Pivot tables are featured in Microsoft Excel as the preferred method for summarizing and subsetting large amounts of data. Excel's tradmark pivot tables are referred to as PivotTables.

Previously, an R user who wished to analyze a given Excel PivotTable in R would be required to export the complete data to R and manually create a table within R itself. Given that the desired result already exists within Excel, this exercise consumes time and introduces potential transfer error.

We have added to RExcel the functionality of automatically selecting and exporting a PivotTable from Excel to R. This action creates within R an object of class structable, with a name assigned by the user via dialog box. With this capability we can directly display an Excel pivot table using the mosaic plot in the **vcd** package or barchart in the **lattice** package.

References

[1] Erich Neuwirth, with contributions by Richard Heiberger, Christian Ritter, Jan Karel Pieterse, , and Jurgen Volkering, *Rexcelinstaller: Integration of r and excel, (use r in excel, read/write xls files)*, 2009, R package version 3.0-18.

Statistical Analysis of Cell Population Data

Michael Halter (NIST, CSTL), Daniel R. Sisan (NIST, CSTL), K.M. Mullen (NIST, MSEL), Z.Q. John Lu (NIST, ITL)

Abstract: Characterizing a population of cells in culture is an important problem in cell standards development. Cell population data can easily be obtained from flow cytometry and automated microscopy. Distribution models can be useful tools for interpreting cell population data and, in principle, should capture both the cell-to-cell variability as well as satisfy the stationarity property for continuously cultured cells. M. Halter et al (2009, J. Theo. Biology **257**, pp.124-130) developed a cell volume distribution model for flow cytometry data which relates to parameters of cell growth rates and division times. The accompanying R package *cellVolumeDist* released by K.M. Mullen et al (2009) in cran.r-project.org contains more efficient statistical fitting methods for fitting parametric multinomial distribution data. Following single cells in time by live cell microscopy can provide insight into the biological variability exhibited by cellular populations, though, reliable cell cycle dependent data is experimentally challenging to acquire. The goal of this talk is to discuss analysis of green fluorescence protein data from live cell images and development of cell population models to facilitate understanding of promoter mechanisms in cell production processes.

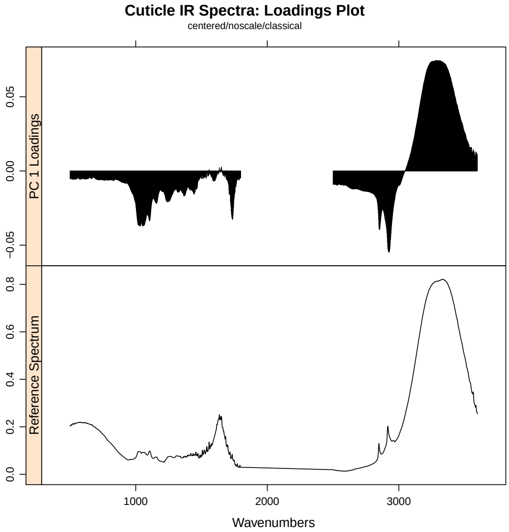
ChemoSpec: an R Package for the Chemometric Analysis of Spectroscopic Data

Bryan A. Hanson[1,2*]

1. Dept. of Chemistry & Biochemistry, DePauw University, Greencastle IN USA
2. The development of **ChemoSpec** was supported by the Faculty Development Committee at DePauw University

*Contact author: hanson@depauw.edu

Keywords: chemometrics, spectroscopy, multivariate, metabolomics



ChemoSpec is a collection of functions for plotting spectra (such as NMR & IR spectra) and carrying out various forms of top-down exploratory data analysis, including hierarchical cluster analysis (HCA), principal components analysis (PCA) and model-based clustering (from **mclust**). S3 classes are used and the data is stored in a `Spectra` object created during data import. Two-dimensional and several 3-dimensional methods are provided for visualizing score plots, including interactive plots and graphical MANOVA methods. Robust methods appropriate for this type of high-dimensional data are employed and diagnostics plots are available. These functions rely heavily on methods and functions discussed in Varmuza & Filzmoser, as well as ideas from the **ggobi** group. **ChemoSpec** is designed to facilitate comparison of samples from treatment and control groups such as typically found in ecological or medical investigations. It is designed to be user friendly and suitable for people with limited background in R, as it was written to support undergraduate research projects. **ChemoSpec** functions will be demonstrated using data from the author's research on plant stress using metabolomics.

References

Bryan A. Hanson (2009) ChemoSpec, `http://github.com/bryanhanson/ChemoSpec`.

K. Varmuza & P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press (2009).

A Plot Method for `"htest"` Objects

Richard M. Heiberger[1,*], **G. Jay Kerns**[2]

1. Temple University
2. Youngstown State University

*Contact author: rmh@temple.edu

Keywords: plot method, `"htest"`, hypothesis tests, confidence intervals

The numerical results of many statistical tests in R are stored in an `"htest"` object. The print method for the class displays a table. We have written a generic `plot.htest` function for the class and constructed the plot methods for normal, t, chi-square, and F tests. The plot methods call the graphing functions in the **HH** package. The hypothesis graphs display

1. the density function for the null hypothesis with critical bounds and shaded areas for the rejection region,
2. the location of the observed value with shading for the p-value, and
3. a second density for the alternative hypothesis with shaded areas for the Type II error.

The confidence interval plots show

1. the density with parameters set at the observed value of the statistic,
2. the confidence interval, and
3. shaded areas for the confidence level.

The axes are labeled in both the data units ($\bar{x}$, or s^2, or s_x^2/s_y^2) and in standardized units. We also wrote menu items for the **RcmdrPlugin.HH** package that can be used with the **Rcmdr** point-and-click GUI.

References

[1] Fox, J. et al. (2010). Rcmdr: R Commander. R package; additional contributors: Michael Ash, Theophilius Boye, Stefano Calza, Andy Chang, Philippe Grosjean, Richard Heiberger, G. Jay Kerns, Renaud Lancelot, Matthieu Lesnoff, Samir Messad, Martin Maechler, Erich Neuwirth, Dan Putler, Miroslav Ristic, Peter Wolf.;
`http://www.r-project.org`,
`http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/`.

[2] Heiberger, R. M. (2010a). HH: Statistical Analysis and Data Display: Heiberger and Holland. R package,
`http://www.r-project.org`;
contributions from Burt Holland and G. Jay Kerns.

[3] Heiberger, R. M. (2010b). RcmdrPlugin.HH: Rcmdr support for the HH package. R package,
`http://www.r-project.org`;
contributions from Burt Holland.

An Intermediate Course in Statistical Computing

Richard M. Heiberger[1,*]

1. Temple University
*Contact author: rmh@temple.edu

Keywords: Course, Statistical Computing

I have just completed teaching the latest incarnation of a course on Statistical Computing for second year graduate students in Statistics. This talk summarizes my course and includes responses to comments made by Peter Dalgaard in his talk at the UseR 2009 conference. My topics include

1. Elementary numerical analysis. I begin with FAQ 37, with focus on the floating point representation of numbers, prevention of disasterous cancellation, and the mechanics of the QR decomposition.
2. Parsing. Precedence of operations, Polish notation, parsing and the use of the same parser for arithmetic, `?plotmath`, and model formulas.
3. Interprocess communication. I use `system`, `shell`, and for Windows `RExcel` and `statconnDCOM`.
4. Graphics. I discuss the construction of regular graphics and lattice graphics. Focus is on the incremental development of complex graphical displays and understanding the tools needed for such development.
5. Packages. The course requires each student to construct a small R package. The goal is to understand the discipline and the mechanism of packaging software so others can use it. We cover methods, function design, debugging, and documentation with `.Rd` files.

In-database analytics with *R*

Brian Hess[1], Michele Chambers[2]

1. Brian is Principal Mathematician and Director of Advanced Analytics at Netezza
2. Michele is Director of Advanced Analytics Product Management at Netezza

Keywords: In-database analytics, data mining, scoring, high performance computing, parallel algorithms

In-database analytics is a hot topic and not typically associated with *R*. Join this sessions to learn about the latest thinking in combing R and in-database analytics from the market leader in data warehouse appliances, Netezza. In this session, you’ll learn:

- What are in-database analytics?
- How does in-database analytics processing help you?
- Can in-database analytics be used for data mining as well as scoring?
- How can you take advantage of a massively parallel architecture to speed up embarrassingly parallel algorithms as well as heroic computations?

This session will highlight how Netezza partners and customers are using R along with in-database analytics to

References

http://www.netezza.com/data-warehouse-appliance-products/twinfin-i.aspx

Analytics at Scale with *R*

Brian Hess[1], Michele Chambers[2]

1. Brian is Principal Mathematician and Director of Advanced Analytics at Netezza
2. Michele is Director of Advanced Analytics Product Management at Netezza

Keywords: In-database analytics, data mining, scoring, high performance computing, parallel algorithms

As *R* grows in commercial acceptance, companies are seeking ways to migrate their ad-hoc analysis into production deployment. Depending on the organization and the problem being addressed, there may be a desire to learn the model on data beyond a sample as well as applying the model to large scale data. Scaling up analytics takes the following basic forms:

- Data Intensity
  - Depth of data (ie: number of transactions, deeper level of transactions or more history)
  - Width of data (ie: number of factors, dimensions, features)
- Computational Intensity
  - Computational complexity (ie: k-means, PCA, heroic computations, matrix, linear algebra)
  - Model complexity (ie: number of experiments, simulations, more initial conditions)
- Parallel Intensity
  - Combination of computational intensity on data intensive problems

In order to address these emerging requirements, commercial companies supporting *R*, need to provide building blocks to make migration of *R* models and algorithms easier including but not limited to:

- Programming constructs (ie: iterators, foreach, etc.)
- Storage constructs (ie: striping data, etc.)

In this session, we want to engage with the *R* community to think about how best to separate the data layer from the processing layer in order to facilitate the commercial viability of *R* for large scale analytics.

References

http://www.netezza.com/data-warehouse-appliance-products/twinfin-i.aspx

Cloud-R: toward a community-backed R in the cloud

Hsin-Ying Hsieh[1], Kun-Hsien Lin[1] & Sun-Chong Wang[1,2]

[1]Systems Biology and Bioinformatics Institute, National Central University, Chungli Taoyuan 32001 Taiwan, [2]Epigenetics Laboratory, Centre for Addiction and Mental Health, Toronto Ont M5T 1R8 Canada

Keywords: R, cloud-computing, parallel computing, distributed computing

http://epigenomics.ncu.edu.tw/Cloud-R

Cloud-*R* is a web-based platform for R embodying the latest cloud computing concept [1]. As the volume of genomic data continues to explode, a computation environment with larger memory and more CPUs for R than a personal computer is in need. Cloud computing, aided by wider Internet penetration and faster Web communications, represents a paradigm shift in computation toward deployment of applications to remote server computers. We propose to let users run their R programs through web browsers. Cloud-*R* is such a Web server that provides R utilities over the Internet. A basic requirement of Cloud-*R* design is that user experience of Cloud-*R* be identical to that of regular R. Users, after free registration, enjoy the following conveniences: (1) Cache of client environment settings; (2) Space for uploading and downloading data; (3) Graphs in PNG/PDF format for downloading. Cloud-*R* adheres to the idea of open software. More importantly, to the goal of virtually limitless computational resources for R, users' computational resources (i.e. computers) can be connected to Cloud-*R* [2]. The contributions of users' hardware can be made and withdrawn at anytime by themselves. Such an 'open resource' model increases the utilization of otherwise idling computers, benefiting the R community at large. Examples of running R at Cloud-*R* are demonstrated in the Cloud-*R* homepage, so are step-by-step instructions of connecting and dis-connecting user's computers to Cloud-*R*.

References

[1] K. H. Lin, H. Y. Hsieh and S. C. Wang, "Cloud-*R*: an R biostatistical computation and graphics environment in the cloud" is submitted to NAR.

[2] REvolution Computing with support, contributions from Pfizer and Inc. nws: R functions for NetWorkSpaces and Sleigh. R package version 1.7.0.0. http://nws-r.sourceforge.net/.

Scalable linear algebra with the **nzmatrix** package

Mario E. Inchiosa[1,*], Cezary Dendek[1], Przemysław Biecek[1]

1. Netezza Corporation
* Contact author: minchiosa@netezza.com

Keywords: Linear algebra, large matrices, ScaLAPACK, high performance computing, in-database analytics

R is an excellent platform for computing with matrices: it is easy to use, and it efficiently supports many useful matrix operations and linear algebra computations, for example via the **Matrix** package. *R* even provides partial access to the ScaLAPACK library of linear algebra routines for distributed-memory high performance computing, via the **RScaLAPACK** package. However, all matrix data feeding into and out of a calculation using these packages must fit within *R*'s memory. The **bigmemory** and **ff** packages lift the memory restriction, but they provide virtually no linear algebra support.

We will present a new package, **nzmatrix,** which enables the *R* user to perform a wide range of linear algebra computations and manipulations on very large matrices using fully distributed and parallelized memory, processing, and database storage. The package supports distributed matrix addition, subtraction, multiplication, inversion, transposition, linear equation solving, linear least squared problems, QR decomposition, LU decomposition, Cholesky decomposition, SVD decomposition, eigenvalues and eigenvectors, reshaping, reduction, and many other functions. **Nzmatrix** scales to hundreds of gigabytes of RAM, hundreds of processing cores, and hundreds of terabytes of database storage, and beyond.

The *R* environment serves as the front end for performing interactive or scripted operations with **nzmatrix** matrices in a manner very similar to that used for native *R* matrices. Matrices are transparently stored in a Netezza Performance Server (NPS) database. **Nzmatrix** provides user-friendly *R* wrapper functions for in-database stored procedures that access MPI, PBLAS, and ScaLAPACK routines for matrix-oriented high performance computing. Matrix data transfer directly from the database to the distributed matrix engine and back. Distributed matrices, in whole or in part, can be converted to native *R* matrices, and vice versa.

Use of and Using R as an Object Oriented Language

John James

Keywords: lists and objects

The benefit of encapsulating data and methods in OO programming are well established. Although R has some aspiration as an Object Oriented Language, it is reasonable to say that its use as such is discouraged. One reason for this is that its mechanism for storing complex types within objects is awkward: use has to be made of lists to essentially ensure distinctness when creating vectors and other hierarchies of these objects.

However lists themselves are very flexible, and arguably too flexible, for use within code that has both to be maintained and also where the integrity of the managed objects must be maintained.

This paper proposes a simple template mechanism, analogous to C++ Template implementation that enables the creation objects that are robust in that they rely on key R concepts but none the less enforce the type rules intended by their creator.

Some simple examples are given in parsing free format text.

steReoscopy revisited

Landon Jensen[1,*]

1. Micron Technology, Inc.
* Contact author: lsjensen@micron.com

Keywords: 3D, Stereoscopy, Polarization, Visualization

3D technology has been making news recently with new hits on the big screen and new gadgets for the home theater. While stereoscopy in *R* is nothing new (consider the example included in the `cloud` function from the **lattice** package), we aim to revisit and explore additional passive 3D technologies (including superimposed images produced from dual projectors and viewed through polarized glasses) for an entertaining look at enhanced data visualization.

Panel on Starting & Building a Local R User Group

Drew Conway, New York RUG, www.meetup.com/nyhackr/
John Nash, Ottawa Gatineau RUG, stat.ethz.ch/mailman/listinfo/r-ug-ottawa
Szilard Pafka, Los Angeles RUG, www.meetup.com/LAarea-R-usergroup/
Jim Porzak, San Francisco Bay Area RUG, http://www.meetup.com/R-Users/

In the past year the a number of local R User Groups (RUG's) have started or ramped up their membership. This panel discussion will focus on actual experiences:

- Building interest in your community
- The kick off meeting
- Finding speakers
- Finding sponsors
- Promotion tricks
- Organizational tricks & traps
- Joys of RUG'ing

Audience participation & questions will be encouraged!

Call for panelists: If you would like to participate on the panel, please contact Jim Porzak.

Model simulation and decision analysis with the SimR package in R

Joseph Kahn*, Peter Danenberg, Sourav Das, Derek Ayers, Richard Nixon, Joyee Ghosh
Novartis Pharmaceuticals, East Hanover, NJ, USA.

Objectives: To better enable modeling and simulation to inform clinical development decisions, we developed an automated simulation package in R incorporating good practices from decision analysis: the package enables quantitative comparison of decisions; it determines influential uncertainties; and it maintains transparency regarding source information and probabilistic dependencies.

Methods: Performing model simulation and decision analysis with SimR requires:

1. Assigning model input parameters and meta-data (e.g., units, value ranges, prior probability distributions and source/pedigree), along with the selection of decisions/strategies. These are stored in a spreadsheet or in .csv files for easy editing and review;
2. Writing an R function to carry out a single model run, (the only code required to be written and validated by the user);
3. Calling the SimR package in R that plans and runs simulations using appropriate random seeds, archives results, and produces analysis tables and plots.

Results: Multiple decision analysis projects are being carried out at Novartis pharmaceuticals in R with the SimR package. For increased speed, the package can run parallel simulations in a grid environment. “Tornado” plots of one-way sensitivity help determine which input parameters are most salient. Two-way sensitivity plots display interaction effects from input parameters to model outputs. Bar charts and cumulative distribution plots compare performance and risks among strategies.

Conclusions: Modeling and simulation using the SimR package is grounded in decision analysis best practices. Its benefits include:

1. Faster model development and validation with use of functions for bookkeeping and production of analysis tables and graphics;
2. Facilitation of rapid communication and model archiving with standardized inputs and results formats that free users to focus on insightful comparisons;
3. Fast run times and greater precision with multiple simulations automatically run in parallel on a grid computing network.

References:

[1] Ross Ihaka and Robert Gentleman (1996). “R: A Language for Data Analysis and Graphics,” *Journal of Computational and Graphical Statistics*, **5** (3): 299—314.

Computers and the Teaching of Statistics

Daniel Kaplan

Macalester College, Saint Paul, Minnesota kaplan@macalester.edu

Keywords: Statistics education, bootstrapping, simulation, modeling

In 1978, Brad Efron published a paper, *Computers and the Theory of Statistics: Thinking the Unthinkable*, about how computing has influenced the theory of statistics. Of course, he had in mind the revolutionary power of high-speed computers to perform calculations. I will talk about an other revolutionary aspect of computing: the idea of a computation as a transformation of inputs to outputs, the notion of "abstraction," the packaging of algorithms in ways that can be easily communicated and used, and the development of notation and languages to support this. For computer programmers, these are everyday, somewhat elementary ideas, but they have hardly made their way into mathematics or statistical education. This is unfortunate. By changing the "instruction set" we provide our students from "square-roots" and "sums" to higher level operations of "fitting" and "randomization" and "p-value" (among others), we can provide our students with greater insight into statistical reasoning and a greatly enhanced power to make use of sophisticated techniques. The syntax of R, in marked contrast to other commonly used statistics packages, provides strong support for bringing such computational abstraction to teaching statistics. I'll talk about the approaches I've taken to doing this in introductory statistics, and applications to teaching about confounding, experimental randomization, coverage and power.

References

Bradley Efron (1978). Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, 21, 460–480.

Daniel T Kaplan (2007). Computing and Introductory Statistics, *Technology Innovations in Statistics Education*, 1, Article 5,
`http://www.escholarship.org/uc/item/3088k195`.

The IPSUR package: an Introduction to Probability and Statistics Using R

G. Jay Kerns[1,*]

1. Youngstown State University
*Contact author: gkerns@ysu.edu

Keywords: Sweave, teaching, probability, statistics, introductory

IPSUR stands for Introduction to Probability and Statistics Using R, ISBN: 978-0-557-24979-4, which is both a textbook and an R package written for an undergraduate course in probability and statistics. Attendees of the class for which the book was written include mathematics, engineering, and computer science majors. The approximate prerequisites are two or three semesters of calculus and some linear algebra in a few places.

IPSUR is FREE, in the GNU sense of the word. This entails both opportunities and challeges, and in this presentation we will discuss a little bit of both.

IPSUR contains several interrelated parts: the *Document*, the *Program*, the *Package*, and the *Ancillaries*. The Document is what the end-user reads. The Program (essentially a gigantic `Sweave` vignette) provides an efficient means to modify the Document. The Package is an R package that houses the Program and the Document, and makes it easy for students to use. Indeed, after installation of the **IPSUR** package the student needs only issue

```
library(IPSUR)
read(IPSUR)
```

to begin study. Finally, the Ancillaries are extra materials that reside in the Package and were produced by the Program to supplement use of the Document. In this talk, we briefly detail each of these in turn.

Since IPSUR is *free*, instructors are *free* to tailor it to suit the needs of their individual classes of students. We will spend part of the time outlining steps instructors can take to accomplish this goal.

References

Kerns, G. J. (2010). The IPSUR Project Page on R Forge. http://ipsur.r-forge.r-project.org/.

Kerns, G. J. (2010). Introduction to Probability and Statistics Using R. First Edition.

Automatic R-script generation for Monte Carlo Simulations

Rüdiger Kessel[1,*]

1. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899
* Contact author: ruediger.kessel@nist.gov

Keywords: Monte Carlo Simulation, automatic *R*-script generation

The Supplement 1 to the guide of the expression of uncertainty in measurement (BIPM 2006) describes how the Monte Carlo method can be used to evaluate the uncertainty of measurement. We have developed a method to automatically generate *R*-scripts for arbitrary measurement problems which carry out such simulations. Based on a text file containing a description of the measurement problem, an R-script is generated and executed without user interaction.

We discuss different aspects of our approach to automate the generation of *R* scripts including passing command line parameters, determining the execution order and data transfer between programs. The automatic script generation is demonstrated in practice for some examples.

The automatic *R*-script generation was developed to support a framework to validate other simulation software. We discuss how *R* can support this framework.

Automatic *R*-script generation can also be used in commercial applications. Since it is clearly defined which part of the system is open-source and which part is proprietary, it allows the integration of both worlds to improve usability without license violation.

References

BIPM (2006). *Expression of Uncertainty in Measurement Supplement 1: Numerical Methods for the Propagation of Distributions*, JCGM Working Group of the Expression of Uncertainty in Measurement

Adaptive Nonparametric Statistics with Applications to Gene Expression Data

John Kloke, Joseph McKean, Patrick Kimes, Hilary Parker

Keywords: gene expressions, genetics

Rank-based methods offer the analyst a robust alternative to traditional likelihood or least squares methods. In this talk we briefly review adaptive nonparametric methods for the two sample location problem. Next we discuss an adaptive estimation scheme for estimating a shift parameter. In the remainder of the talk we illustrate these methods via an R package we are developing and apply them to gene expression data.

Extracting within-experiment precision of horticultural experiments useful for meta-analysis

Guido Knapp, Bimal Sinha, Dihua Xu

For combining results from independent experiments, it is essential that information about the precision of the estimates of treatment effects is available. In publications of horticultural experiments, the results of multiple comparisons tests are often reported without sufficient information about the precision of the experiments. Based on limited information of the precision of an experiment such as treatments with the same letter are not significantly different, we develop a method for extracting a possible range of the precision of the experiment which can then be used for meta-analysis. The procedure is demonstrated using a real data example where alternatives to methyl bromide are studied in pre-plant soil fumigation. We also provide an R program which computes the possible range of the precision.

Tests in Modeling Continuous Multivariate Distributions Using Copulas

Ivan Kojadinovic[1], **Jun Yan**[2,*]

1. Department of Statistics, University of Auckland
2. Department of Statistics, University of Connecticut

*Contact author: jun.yan@uconn.edu

Keywords: goodness of fit, multivariate independence, pseudo-observations, rank-based tests, serial independence

Copula models have become popular in many areas such as finance, insurance, and hydrology. Improper use of copula, including that allegedly destroyed the Wall Street in October, 2008, has raised concerns about application of copulas. Hypothesis tests provide guard against abuse of copulas. A copula model is needed only when independence is rejected. A specific form of copula cannot be applied to give interpretation unless it passes goodness-of-fit tests. We presents tests for independence and goodness-of-fit of copula models, among other recent development including selected extreme value copulas, in the R package **copula**. Some implementation details are documented. Usage of the tests are illustrated with real examples.

Infrared Spectrometric Purity Control of Chemical Substances using R

Fayaz Kondagula[1,*], **Karl Molt**[1]

1. Universität Duisburg-Essen, Department of Chemistry, 47048 Duisburg, Germany
*Contact author: fayazpharm@yahoo.co.in

Keywords: Infrared Spectroscopy, Difference Spectroscopy, Purity Control

Infrared Spectra of organic compounds offer a wealth of bands which are characteristic for certain structural units. Therefore IR spectroscopy is widely used as "fingerprint method" for identifying unknown compounds. A typical IR spectrum contains about 4000 data points and computer software is needed for processing the spectral data. The instrument manufacturers normally provide proprietary software for this purpose. But this is often limited to a restricted number of common applications. To enable a more flexible and universal numerical and statistical evaluation of spectral data, the authors have developed methods to directly read the spectra into R.

On this basis a method was developed for controlling the purity of chemical substances. If a spectrum of a pure reference substance is known, there are two ways for determining the purity of a potentially contaminated sample. The first is difference spectroscopy and its principle is very simple. After obtaining the spectrum of the sample the reference spectrum is subtracted from it. The resulting difference spectrum is then due to the impurity contained in the sample. However due to noise, baseline shifts and atmosperic disturbances IR spectra normally are far from perfect and so a specific algorithm ("dynamic difference spectroscopy") was developed for determining the optimal difference factor, i.e. the factor by which the reference spectrum has to be multiplied before subtraction to obtain an optimal compensation. This works by calculating a whole series of difference spectra with decreasing difference factors and monitoring how the integral of the difference spectra changes. Shortly before it becomes significantly larger than zero, the optimal compensation has been reached. For finding this critical point the `diff` function was used.

The second way to check for impurities is to calculate the correlation coefficient r between the sample and the reference spectrum [1, 2, 3]. This is performed by regressing the sample on the reference spectrum with `lm`. The correlation coefficient however has the disadvantage that the progress of its deviation from one is very slow with increasing contamination. We therefore transformed the r values into Fisher's z coefficients (using the package **survcomp**) and could show that these react much more sensitive to small impurities [4].

Both methods are demonstrated with the example of a certain plasticizer (Palatinol N) contaminated by increasing quantities of another type of plasticizer (Palatinol 911 P).

References

[1] Horst Weitkamp und Dieter Wortig, Vollautomatische Identitätsprüfung von Arzneimitteln durch rechnergekoppelte IR-Spektroskopie. Mikrochimica Acta 1983 II, 31-57.

[2] Robert A. Hoult, Patent number US5023804: Method and apparatus for comparing spectra.

[3] Peter R. Griffiths and Limin Shao, Self-Weighted Correlation Coefficients and Their Application to Measure Spectral Similarity, Appl. Spectroscopy. 63, 2009, 916-919.

[4] Fayaz Kondagula and Karl Molt, Infrared Spectrometric Purity Control of Organic Liquids and Water, Clean 2009,37(12),955-962.

Simple Bayesian Networks on Netezza Box

By Mieczyslaw Klopotek and Przemyslaw Biecek and Justin Lindsey
Netezza Corporation, Polish Academy of Sciences

The NZ Analytics cartridge contains a bundle of implementations of useful statistical and data mining algorithms. When installed, they can be accessed either as SQL functionality or as R functions, depending on the interface you use. Diverse implementation technologies (stored procedures, user-defined functions, user-defined aggregates) were used depending on the nature of the solved problem. The algorithms implemented or wrapped into stored procedures are accessible also via an NZ-R interface. This presentation concerns the Bayesian Network implementation.

A **Bayesian Network** is a method of representing a joint probability distribution in many variables in a compact way. The representation consists of a directed acyclic graph structure DAG with conditional probabilities of a node given its parents attached to each node, $P(X_i \mid \pi(X_i))$. We talk about a simple Bayesian Network if each node has only one parent. Though this assumption is a significant simplification, it has been found useful for problems in a large number of variables. In spite of the simplicity of the case, the efficient approach of Chow/Liu [1] is of prohibitive memory complexity (quadratic in the number of variables, so 5,000 variables is a practical limitation for 1GB memory), hence ways to overcome the memory limitations need to be sought. Though various space- and time saving improvements have been proposed [2,4], they prove to be not useful under massively parallel database systems in which data is stored record-wise, because they restrict the number of dependency computations and not the number of passes through the database which most time-consuming.

To be able to compute BNs from data restricting the number of passes through the database, a new approach, based on insights from [3], is being proposed in this paper, with the following steps:

- Step 1: Take the first N variables for which we can fit their sufficient statistics into the memory
- Step 2: Build a Chow/Liu tree form them
- Step 3: Forget the sufficient statistics except those related to edges in the obtained tree (their amount will be linear in the number of edges)
- Step 4: Take the next portion of say M variables so that the sufficient statistics of the tree edges plus the sufficient statistics for the matrix N x the number of nodes in the tree will fit into the memory
- Step 5: Apply the iterations of the algorithm IT (starting with step 3) for the M new variables.
- Step 6 If any variables are left, go back to step 3, otherwise terminate (that is apply the rest of the Chow/Liu procedure of orienting edges and computing the conditionals from original data).

A correctness proof will be provided in the paper.

References

[1] Chow, C. K., Liu, C. N.: Approximating discrete probability distributions with dependence trees, IEEE Transactions on Information Theory, IT-14, No.3, 1968, pp. 462-467

[2] M.A.Kłopotek: A New Bayesian Tree Learning Method with Reduced Time and Space Complexity. *Fundamenta Informaticae*, 49(no 4)2002, IOS Press, pp. 349-367. –

[3] M.A.Kłopotek: A New Space-Saving Bayesian Tree Construction Method for High Dimensional Data *Demonstratio Mathematica*, Vol. 35, No. 3 (2002)pp. 671-684

[4] Meila, M.: An accelerated Chow and Liu algorithm: fitting tree distributions to high-dimensional sparse data. http://citeseer.nj.nec.com/363584.html

Kod pola został zmieniony

The caret Package: A Unified Interface for Predictive Models

Max Kuhn[1*]

1. Pfizer Global R&D *Contact author: max.kuhn@pfizer.com

Keywords: Machine Learning, Resampling, Parameter Tuning, Feature Selection

The **caret** package, short for **C**lassification **A**nd **RE**gression **T**raining, contains numerous tools for developing predictive models using the rich set of techniques available in R. The package focuses on simplifying model training and tuning across a wide variety of models. It also includes methods for pre-processing training data, calculating variable importance, feature selection and model visualization. An example from computational chemistry is used to illustrate the functionality on a real data set.

Structured Text Access and Analysis

Bill Ladd[1]

1. RecordedFuture, Inc

* Contact author: bill.ladd@recordedfuture.com

Keywords: JSON, Spotfire

The Web contains a vast amount of unstructured information. Web users access specific content of interest with a variety of Websites supporting unstructured search. The unstructured search approaches clearly provide tremendous value but are unable to address a variety of classes of search. RecordedFuture is aggregating a variety of Web-based news and information sources and developing semantic context enabling more structured classes of search. In this presentation, we present initial methods for accessing and analyzing this structured content. The **RJSONIO** package is used to form queries and manage response data. Analytic approaches for the extracted content include normalization and regression approaches. R-based visualization approaches are complemented with data presentation capabilities of Spotfire.

.

Placing the Power of R into your Hands

Andrew Lampitt

Keywords: business intelligence

Jaspersoft provides a web-based, open and modular product for the evolving business intelligence needs of organizations. Jaspersoft is experiencing high growth serving market sectors in education, pharmaceutical/healthcare, telecommunications, and technology. Select customers to be covered in this session include University of Nebraska, Healthport, Telvent, QED Financial Systems, Virgin Money, eBuilder, HandySoft, and British Telecom. Jaspersoft's open source business intelligence suite is the world's most widely used BI software, with more than 10 million product downloads worldwide, an estimated 150,000 production deployments, and more than 11,000 commercial customers in 100 countries. This session introduces Jaspersoft, the commercial company behind JasperReports and iReport, the world's most downloaded open source reporting engine and graphical report designer, as well as Jaspersoft's suite of commercial and open source reporting and data analysis products, built on top of the foundations of JasperReports, iReport, and other open source components. Jaspersoft's software is rapidly updated by a community of more than 120,000 registered members working on more than 350 projects. Jaspersoft's modular approach allows integration with R in a flexible way so that R charts and results may be easily distributed to a large audience in a friendly web user interface. Learn how Jaspersoft can put the power of R into the hands of end-users through published and interactive reports and dashboards, and how data can be further explored in Jaspersoft using point & click reporting and multi-dimensional data analysis tools.

Animated Statistical Graphics using R

by Autumn Posey Laughbaum
Undergraduate at Montana State University, Mathematical Sciences

Abstract: Visualizing data and statistical results is crucial for all scientists. For data collected over multiple dimensions, such as space and time, effective visualization requires more sophisticated and unique tools. For example, changes over space and time can be conveyed through the display of several contour (e.g. topographical) plots at different moments in time. A more intuitive and streamlined approach is to display the change over time with an animation. Allowing for a dynamic time variable helps us visualize and understand the changes in the data more intuitively. We can also extend this concept to visualize error associated with statistical prediction or estimation. Typically, such information is provided in two separate plots, a contour plot of the prediction and one for the standard error of the prediction. An animation would streamline this process by providing all of the information, the prediction and standard error, in one dynamic plot. Functions to do this are not yet widely available in statistical software, thus I extend the current methods by changing the dynamic variable from time into a variable that quantifies the margin of error associated with prediction or estimation over space. Effectively, I create an animated spatial prediction interval plot to convey the information. In addition creating functions for the two animations mentioned above, I also develop computer code containing functions for displaying change in temporal data, as well as creating teaching mechanisms for introductory statistics. The code runs in the statistical software R and can be easily accessed through an R package, a streamlined group of functions. R is free open source software, runs on all platforms, and is widely used in many disciplines. The animations that the user creates can be run within R and also saved as Flash files. Examples of my functions are available on my website: http://studentweb.montana.edu/autumn.laughbaum/research.html. Animations are easy to understand and foster a level of excitement in the viewer. My animation functions not only streamline current methods of visualization, but they also provide more intuitive resources for statisticians and non-statisticians alike.

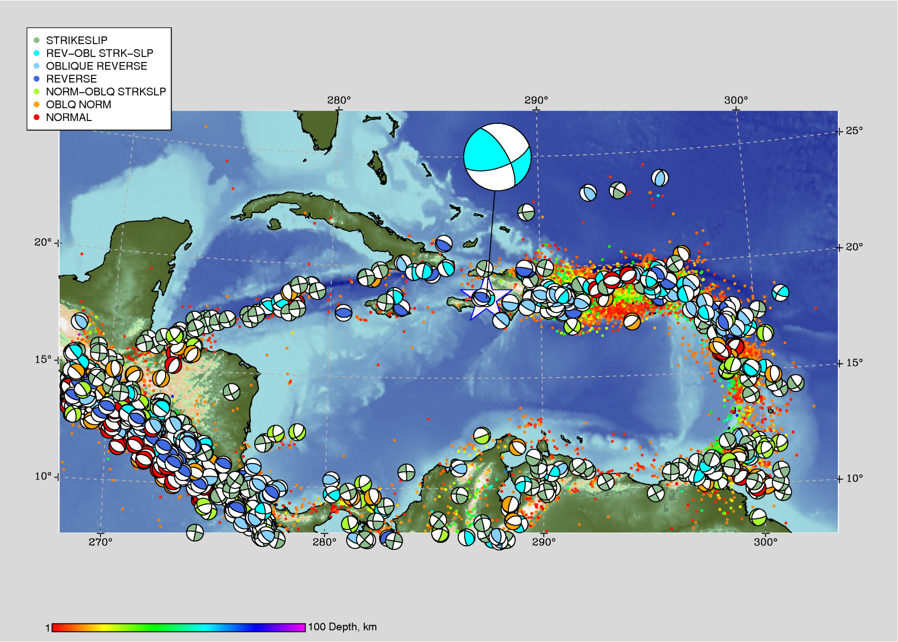
The Haiti Earthquake: Seismological Analysis Using R

Jonathan M. Lees[1]

1. University of North Carolina, Chapel Hill
*Contact author: Jonathan M. Leesjonathan.lees@unc.edu

Keywords: Earthquake, Hazard, Spatial Analysis, Time Series,

The catastrophic earthquake in Haiti on January 12, 2010 focused the world on geologic hazards in economically depressed regions of North America. In this presentation I will illustrate how R and several contributed packages can be used to quickly extract information on the distribution of earthquakes, provide graphical tools for visualization, and investigate wave propagation phenomena associated with this disaster. Contributed packages discussed here include **RSEIS**, for manipulation of seismic waveform data, **RFOC** for analysis of earthquake focal mechanisms, **GEOmap** for analysis of spatial data. Interactive parts of the analysis are supported by a simple GUI package, **RPMG**. The suite of programs is aimed at exploratory analysis that illuminates geologic/tectonic issues related to great earthquakes. Emphasis is put on graphical visualizations of complex data, from time series to spatial distributions of earthquake meta-data.



Random KNN Classification and Regression

Shengqiao Li[1], **Donald A. Adjeroh**[2] **and E. James Harner**[1,⋆]

1. Department of Statistics, West Virginia University
2. Lane Department of Computer Science and Electrical Engineering, West Virginia University

⋆Contact author: jharner@stat.wvu.edu

Keywords: Classification, Machine Learning, K Nearest Neighbor, High Dimensional Data

High dimensional data, involving thousands of variables, are becoming increasingly available in various applications in biometrics, bioinformatics, chemometrics, and drug design. While such high dimensional data can be readily generated, successful analysis and modeling of such datasets is highly challenging. We present Random KNN, a novel generalization of traditional nearest-neighbor modeling. Random KNN consists of an ensemble of base k-nearest neighbor models, each constructed from a random subset of the input variables. We study the properties of the proposed Random KNN, including its theoretical convergence. Using different datasets, we perform an empirical analysis of its performance, and compare its results with those from recently proposed methods for high dimensional datasets. It is shown that Random KNN provides significant advantages in both the computational requirement and classification performance. The Random KNN approach can be applied to both qualitative and quantitative responses, i.e., classification and regression problems, and has applications in statistics, machine learning and pattern recognition. The Random KNN algorithms are implemented in the *rknn* R package.

References

Shengqiao Li, E. James Harner and Donald A. Adjeroh (2010). Random KNN. Submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

metRology - a new R package for statistical metrology

Hung-kung Liu[1,*], **Steve Ellison**[2], **Alan Heckert**[1], **William Guthrie**[1], **Antonio Possolo**[1], **Daniel Samarov**[1], **James Yen**[1]

1. National Institute of Standards and Technology, USA
2. Laboratory of the Government Chemist, UK

*Contact author: liu@nist.gov

Keywords: Software deployment

The techniques for statistical modeling, computation and data analysis that have, particularly during recent years, been developed or adapted for use in metrology often are sufficiently involved to prevent manual application, and instead require that the metrologist employ implementations in computer software. This talk describes a new joint initiative at NIST's Statistical Engineering Division and Laboratory of the Government Chemist, UK to facilitate statistical and mathematical computation in measurement science by producing a new R package, **metRology**, that is freely available to all. The broad goals of the project are:

1. To provide access to a wide range of powerful statistical and graphical methods for the analysis of metrological data, exploiting the model-oriented constructs that R provides;

2. To accelerate the development of extensible, scalable, and interoperable software for metrology;

3. To promote the production and dissemination of high-quality documentation that is a key component of reproducible research;

4. To provide training in R emphasizing computational and statistical methods for the analysis of metrological data.

R is provided with a command line interface, which is the preferred user interface. However, metrologists who will only use some features of **metRology** occasionally would probably benefit from a graphical user interface (GUI). A menus/dialog boxes GUI and a spreadsheet GUI for R will be introduced.

Using *R* for Data Simulation and Regression of Isothermal Titration Calorimetry of Proteins with Alternative Conformations

Yingyun Liu[1,*]

1. Department of Biology, Johns Hopkins University, Baltimore, MD 21218

* Contact author: yingyunliu@jhu.edu

Keywords: protein, ligand, ITC, conformation, regression

All proteins have alternative conformations. For some, the partition function is such that some conformations are much more probable than others. A compound ligand can also have significant alternative conformations. When isothermal titration calorimetry (ITC) experiments are carried out with such proteins and ligands, data analysis can become complicated. A model system of a ligand binding to only one of two alternative conformations of a protein was analyzed in the *R* language environment, with extensive use of the package **minpack.lm**. A multitude of symmetry is found for this reaction system between the total ligand concentration and the total protein concentration. Reverse ITC titrations with the same experimental parameters yielded exactly the same data for this system, according to the analysis and simulations. Regression of simulated ITC data for the system with the simple binary binding model all converged when binding could be observed, though there are situations where binding can not be observed even when there is significant binding between the ligand and the one binding-capable protein conformation. In typical cases, the apparent enthalpy and binding constant obtained from regression agree with the values calculated with the Gibbs-Helmholtz equation using the simulation parameters. In cases where the binding curve is not typical, the apparent enthalpy and binding constant from regression can be off from calculated values with accompanying abnormal stoichiometry. In such cases, changing concentrations of reactants in different ITC runs can affect the regression behavior, and can help in obtaining more accurate thermodynamic reaction parameters. These simulations provide the reaction signatures of the modeled system, which can be helpful when deciphering a reaction mechanism or interpreting ITC data. The simulations also demonstrated the validity and usefulness of the Gibbs-Helmholtz equation in a situation where there is no temperature change.

References

Timur V. Elzhov, Katharine M. Mullen (2009). *CRAN-Package minpack.lm* http://cran.r-project.org/web/packages/minpack.lm/index.html

Freire, E (1998). Statistical thermodynamic linkage between conformational and binding equilibria. *Advances in Protein Chemistry* **51**: 255-279.

Perola, E., and Charifson, P.S. (2004). Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *Journal of Medicinal Chemistry* **47**: 2499-2510.

Taneva, S.G., Moro, F., Velázquez-Campoy, A., Muga, A.(2010). Energetics of nucleotide-induced DnaK conformational states. *Biochemistry* **49**: 1338-1345.

Fitting Multiphase Regression Models in R with Applications in Microarray Experiments

Z.Q. John Lu, Statistical Engineering Division, Information Technology Lab, NIST, Gaithersburg, Maryland, USA ;
Florian Potra, NIST and University of Maryland, Baltimore County, USA ;
Alex Jing Wang, NIST SURF Program and Cornell University, Ithaca, New York, USA.

Multiphase regression models have been used in many statistical applications. Here we present a new application by recasting multiphase regression as a model-based approach to defining performance metrics in a measurement process. Performance metrics such as background, detection limit, linearity range, efficiency (linear slope) are commonly used to gauge measurement performance of a bioassay, see for example, Satterfield, Lippa, Lu, Salit, J. Res. Natl. Inst. Stand. Techol. **113**, 157-174 (2008). Fitting multiphase regression model to measurement data curves presents an interesting and challenging optimization problem, because the underlying objective function is not continuous with respect to the threshold or intersection point of two linear lines. A simple iterative procedure is proposed for solving the nonlinear least square problem in which a clever explicit solution is given at each step. Indeed, we will show that the objective function is a piecewise rational function of threshold, and through both simulated data and spike-in microarray Affymetrix experiments, we illustrate the use of R functions developed for this new application.

Sparse Model Matrices for Generalized Linear Models

Martin Mächler[1,2,*], **Douglas Bates**[1,2]

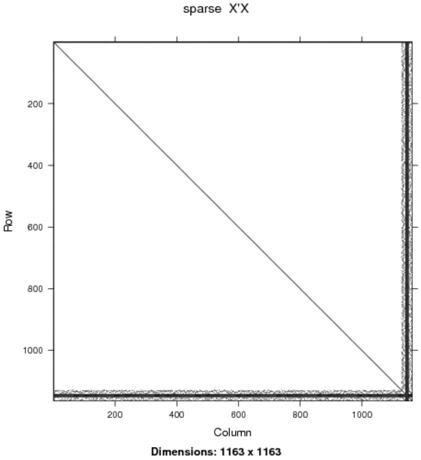
1. ETH Zurich, Switzerland and University of Wisconsin, Madison, USA
2. R Core Development Team *Contact author: maechler@stat.math.ethz.ch

Keywords: sparse matrices, linear models, GLM, mixed effects, large data

Using sparse model matrices for linear, generalized linear, and also (generalized) linear mixed effect models can be very advantageous particularly, when most predictor variables are categorical (`factor`s). We will demonstrate using such sparse matrices via function `sparse.model.matrix()` from the **Matrix** package and use sparse Cholesky decompositions to fit large linear and generalized linear models (GLM) efficiently. We will compare these with the corresponding functionality in R packages **biglm** and **speedglm**.

However, speed is not everything, and a not so well-known fact is that S and R have had smart "home-made" low-level code underlying `lm()`, for situations of ill-conditioned ("quasi-singular") design matrices $\mathbf{X}$. The benefit of this pivoting code has been the easy "identification" of components $\hat{\beta}_j$ to be set to `NA` such that other parts remained stably defined. Traditional high-quality code for matrix decompositions and least squares computation, as provided, e.g., by LAPACK, would not automatically provide the same functionality, nor do the analogous libraries for sparse computations.

We will explore some of these computational challenges in a mixed-effect model of moderately large data sets (e.g., $n = 73421$, $p = 1163$, $q = 2972$). Notably, the computation of standard errors for the fixed effects can pose problems which seem harder to solve than the parameter (maximum likelihood) estimation itself.



References

Douglas Bates and Martin Maechler (2005 ff). Introduction to the Matrix package (and other vignettes). `http://cran.r-project.org/web/packages/Matrix/`

Douglas Bates (2005 ff); Diverse vignettes on using Mixed-Effect models `http://lme4.r-forge.r-project.org/`

Douglas Bates (2010). **lme4: Mixed-Effects Modelling with R**; Springer, useR! series.

Thomas Lumley (2009). `biglm`: bounded memory linear and generalized linear models. R package version 0.7. `http://CRAN.R-project.org/package=biglm`

Martin Maechler and Douglas Bates (2009). Sparse Matrices in package Matrix and applications; slides from useR! 2009, Rennes. `http://matrix.r-forge.r-project.org/slides/2009-07-10-Rennes/MM-talk.pdf`

Generalized additive models to nomial responses using bivariate Kernel: an solution to spatial analysis.

A.C.C.N. Mafra*, R. Cordeiro, L.B. Nucci, C. Stephan (State University of Campinas – Unicamp, Campinas, São Paulo, Brazil, 13083-970)

Keywords: bivariate Kernel, generalized additive models, multinomial responses, spatial analisys

The spatial distribution of disease risk has always been a concern in Epidemiology. Particularly, in the last two decades, several techniques of spatial analysis for epidemiological data have been developed to estimate the variation of risk in space. However, several studies based on scales or multinomal response still use logistic regression models for data analysis without incorporating spatial effects. This is partially due to the reduced number of techniques and computational tools which allow treating adequately multinomial responses occurring in epidemiological data within the spatial setting. This work aims to contribute to overcoming this limitation. To do this, when the response of a case-control study is multinominal, the authors define the spatial risk[1] and present how to obtain, analyse and map it, including acquaintance of the estimates statistical significance. To obtain the risk is presented an algorithm to estimate GAM when the response is multinomial and the non-parametric function, to study the space, is estimated as a bidimensional kernel. The significance analysis of the estimated spatial risk is obted trough a simulation method[2]. All programming was made in software R2.7 and the maps were built on ArcMap 9.2. Data from a population-based case-control study of occupational accidents in a Brazilian city[3] were adjusted with the response variable classified in three categories: serious cases, mild cases and controls. Along the spatial analysis, other informations about the occupation and the employee were included. The analysis has found areas of significant increased relative risk and protection for occupational accidents that varied depending on the level of comparison. Some areas had twice the risk compared to the average of the region studied when considering serious accident. In parametric variables studied, different risk and protection factors were found for the two levels of the cases. This work brings the complete way to analyze data from case-control studies with multinomial response where the spatial risk has to be analyzed.

REFERENCES

1. Bithell J. An application of density estimation to geographical epidemiology. Statistics in Medicine 1990;9:691-701.
2. Kelsall JE,.Diggle PJ. Spatial variation in risk of disease: a nonparametric binary regression approach. Applied Statistics 1998;47:559-73.
3. Stephan C. Distribuição do risco de acidente do trabalho entre trabalhadores precarizados de Piracicaba. - Campinas, SP : [s.n.], 2008.

New developments for extended Rasch modeling in R

Patrick Mair, Reinhold Hatzinger

Keywords: social sciences

The eRm package allows for the computation of extended Rasch models. This package implements the Rasch model, the linear logistic test model (LLTM), the linear logistic model with relaxed assumptions (LLRA), the (linear) rating scale model (RSM and LRSM), and the (linear) partial credit model (PCM and LPCM). Conditional maximum likelihood estimation (CML) is used for item parameter estimation. Various model diagnostics such as Andersen's LR-test, item-fit and person-fit statistics, graphical model tests, exact model tests, etc. are available. In this talk we give an overview of recent developments and implementations.

Stochastic modeling and simulation in the design of multicenter clinical trials

Frank Mannino, Richard Heiberger, Valerii Fedorov

The design of multi center trials requires careful balancing and synchronization of closely related components such as selection of sample size, test statistics, treatment randomization method, customer and sponsor risks, number of centers, duration of trial, logistics of drug supply and manufacturing under uncertainty or randomness of input information. The effective way to understand how these aspects interact and affect each other is through the use of stochastic models and Monte Carlo simulations, as many of these models are difficult or practically impossible to handle analytically, especially when used in combination. By simulating our trials we can get a better sense of the properties including variability of statistical power for various patients outcomes under competing models, treatment imbalances, length of the trial, amount of drug needed, and overall costs. An R package and RExcel interface have been developed to handle these simulations in real time setting to quantify and support decision making in development of new drugs.

The bullwhip effect under a generalized demand process: an R implementation.

Marlene Marchena[1*]

1.Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil.
*Contact author: marchena@ele.puc-rio.br

Keywords: SCM, ARMA, forecasting, bullwhip effect, safety stock.

An important problem in Supply Chain Management (SCM) is the bullwhip effect, a phenomenon in which demand variability increases as one moves up the supply chain. In this paper we investigate a theoretical and practical application of Zhang (2004b) with the purpose of quantifying the bullwhip effect. The measure commonly used for this phenomenon is the ratio of the variance of the order process to that of the demand process.

We consider a single-item in a two-stage supply chain model, where the retailer employs an order-up-to replenishment policy, and we use the optimal forecasting procedure that minimizes the mean squared forecasting error. Using this model, we measure the bullwhip effect in the case of a stationary autoregressive moving average ARMA(p,q) demand process admitting an infinite moving average (MA) representation. In some particular cases we obtain explicit formulas for this measure. Finally, a R implementation is provided.

We program a function (`SCperf`) whose output gives numerical results for the bullwhip effect and other supply chain performance variables. It is well known that measuring the bullwhip effect is difficult in practice but our function overcomes this problem thanks to the help of a R function (`ARMAtoMA`) which converts an ARMA process into an infinite MA process. It leads to a simple but powerful tool which can be helpful for the study of the bullwhip effect and other supply chain research problems.

Our contributions to this subject can be described as follows: first, this study hopes to improve the understanding of time series techniques. On the other hand, we show that for certain types of demand processes the use of the optimal forecasting method considered in the model leads to significant reduction in the safety stock level. This highlights the potential economic benefits resulting from the use of time series analysis. Next, the function `SCperf` might be used to complement other managerial support decision tools. Finally, the code is given, which makes (together with the fact that R is a freeware) the whole research reproducible by everyone. It may be as well modified for specific tasks.

References

Duc, T. T.H., Luong, H.T. and Kim, Y.D., (2008). A measure of the bullwhip effect in supply chains with a mixed autoregressive moving average demand process. *European Journal of Operational Research*, 187, 243–256.

Luong, H.T. and Phien, N.H., 2007. Measure of the bullwhip effect in supply chains: the case of high order autoregressive demand process. *European Journal of Operational Research*, 183, 197-209.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
`http://www.R-project.org`.

Zhang, X., (2004b). Evolution of ARMA demand in supply chains. *Manufacturing and Services Operations Management*, 6 (2), 195-198.
`http://personal.georgiasouthern.edu/~xzhang/research/msomDemandEvolution.pdf`.

R package wfIMA: Wavelet-Functional Indexes of Magnetic Activity

Inga Maslova[1,*]

1. Department of Mathematics and Statistics, American University, Washington, D. C.
*Contact author: Maslova@american.edu

Keywords: Space physics, magnetic storm, wavelets, functional data

An R package for space physics applications, **wfIMA** is developed. It consists of several major functions that compute indices of the magnetic storm activity and estimate the Solar quiet daily variation. Both indexes are widely used in geophysical community. A novel approach based on wavelet and functional principal component analysis is used in order to develop an applied statistical software. This package implements the ideas introduced in [1] and [2]. It utilizes the functions of **fda** and **waveslim** packages. Magnetic storm activity indices and quiet daily Solar variations are computed automatically without any subjective human intervention using the most recent magnetometer data available. This allows to produce indexes in near-real time, which would be an attractive alternative to the current index. This package will be publicly available at Comprehensive R Archive Network (CRAN). It would be a very important tool for the geophysical community and would address the need of software that balances the parameter flexibility with reliable results.

References

[1] I. Maslova, P. Kokoszka, J. Sojka, and L. Zhu, *Removal of nonconstant daily variation by means of wavelet and functional data analysis*, Journal of Geophysical Research **114** (2009), A03202, doi:10.1029/2008JA013685.

[2] ———, *Estimation of sq variation by means of multiresolution and principal component analyses*, Journal of Atmospheric and Solar-Terrestrial Physics **72** (2010), 625 – 632.

Rdsm: Distributed (Quasi-)Threads Programming in R

Norman Matloff[1⋆]

1. Dept. of Computer Science, University of California, Davis
⋆Contact author: matloff@cs.ucdavis.edu

Keywords: parallel programming, threads, shared-memory, collaborative applications, Web data collection

`Rdsm` provides a threads-like programming environment for R, usable both within a multicore machine and across a network of multiple machines. The package gives the illusion of shared memory, again even across multiple machines on a network.

Consider for instance the assignment `y <- x`. In a message-passing setting such as `Rmpi`, `x` and `y` may reside in processes 2 and 5, say. The programmer would write code

send x to process 5

to run on process 2, and write code

receive data item from process 2
set y to received item

to run on process 5. By contrast, in a shared-memory environment, the programmer would merely write

set y to x

vastly simpler.

Accordingly, many in the general parallel processing community find that the shared-memory approach makes code easier and faster to develop, and easier to maintain and extend. See for example Chandra (2001), Hess *et al* (2003) etc.

`Rdsm` should have a wide variety of applications, such as

- Performance programming, in “embarrassingly parallel” settings.
- Parallel I/O applications, e.g. parallel Web data collection.
- Collaborative tools.

Again, these applications can be done via message-passing too, but we argue that the shared-memory paradigm makes these applications easier to develop, maintain and extend.

We will also compare `Rdsm` to other parallel R packages, in terms of paradigm, flexibility and convenience.

Finally, two general points about parallel R and parallel programming in general will be presented. First, it will be argued that for non-embarrassingly parallel situations, the nature of R presents a fundamental obstacle to performance, so that one is essentially forced to have R call parallel C code, say with direct threads or via OpenMP. Second, a simple mathematical argument will be presented regarding parallelization of `for` loops, showing that in most cases it is not profitable to micromanage allocations of iterations to processes.

References

Chandra, Rohit (2001), *Parallel Programming in OpenMP*, Kaufmann, pp.10ff (especially Table 1.1).

Hess, Matthias *et al* (2003), Experiences Using OpenMP Based on Compiler Directive Software DSM on a PC Cluster, in *OpenMP Shared Memory Parallel Programming: International Workshop on OpenMP Applications and Tools*, Michael Voss (ed.), Springer, p.216.

Trade Cartograms: a Graphical Method for Dyadic Datasets

Benjamin Mazzotta[1*]

1. Fletcher School, Tufts University
* benjamin.mazzotta@tufts.edu

Keywords: International trade, cartogram, country dyads,

This paper proposes to create a complete set of annual cartograms for world trade by partner country from the IMF's Direction of Trade Statistics, with one plot for each country's trade relationships in a given year. The primary method is the Newman-Gastner diffusion cartogram algorithm, implemented variously as Mark Newman's *cart* software (Gastner and Newman 2004), and as Tom Gross' ArcGIS script. Dyadic datasets are often aggregated due to the difficulty of presenting the information graphically. This set of cartograms would constitute a set of reference images akin to the S. Louis Federal Reserve's FRED charts for national economic data series.

Cartograms have been widely exploited for the purpose of comparing distributions of national variables. For example, *The Atlas of the Real World* is a collection of cartograms of national population, education, infrastructure, and many other statistics. Cartograms retain basic geographic information that other graphical methods, such as balloonplots, lack. Conversion of dyadic datasets to national aggregates is the default for GIS, resulting in a graphic that ignores that variation across each nation's partners. One cartogram per country, as I propose, would instead highlight the variation across partner countries.

Diffusion cartograms are computationally tractable with desktop computers. A variety of methods exist for extremely simple implementations of this approach, notably ScapeToad and Frank Hardisty's Cartogram Site. Statisticians can present intertemporal and international comparisons with a complete series of cartograms for each country. Simple animations are possible. Worldmapper's animations are even more creative, comparing the distribution of national populations conditional on per capita income. R's **GIS** package transformed the IMF trade data into GIS layers. I used ArcGIS for mapping.

References

Dorling, Daniel, Mark Newman, and Anna Barford. 2008. *The Atlas of the Real World*. Thames & Hudson, October 27.

Gastner, Michael T., and M. E. J. Newman. 2004. Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America* 101, no. 20 (May 18): 7499-7504. doi:10.1073/pnas.0400280101. http://www-personal.umich.edu/~mejn/cart/

Gross, Tom. *Cartogram Geoprocessing Tool version 2*. ESRI Support Center. http://arcscripts.esri.com/details.asp?dbid=15638.

Hardisty, Frank. Cartogram Generator. http://people.cas.sc.edu/hardistf/cartograms/.

International Monetary Fund, *Direction of Trade Statistics*. Washington, D.C. http://imfstatistics.org.

OECD Statistics Directorate. 2008. *Total Trade in Value by Partner Countries (1960-2007)*.

ScapeToad - cartogram software by the Choros laboratory. http://scapetoad.choros.ch/.

Worldmapper: The world as you've never seen it before. http://www.worldmapper.org/faq.html.

SPRINT: a Simple Parallel INTerface to High Performance Computing and a Parallel R Function Library.

Muriel Mewissen[1], Thorsten Forster[1], Terry Sloan[2], Savvas Petrou[2], Michal Piotrowski[2], Bartek Dobrelewski[2], Peter Ghazal[1], Arthur Trew[2], Jon Hill[3]

1. Division of Pathway Medicine, The University of Edinburgh, Chancellor's Building, 49 Little France Crescent, Edinburgh, EH16 4SB, UK.
2. Edinburgh Parallel Computing Centre, James Clerk Maxwell Building, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, Uk.
3. AMCG, Earth Science and Engineering, Imperial College, London, SW7 2AZ, UK.

Keywords: High Performance Computing, Bioconductor, Correlation, Permutation, Microarray

The analysis of post genomic data is increasingly becoming harder to perform on standard computing infrastructures due to the sheer amount of data involved requiring more disk space and longer processing times. High Performance Computing (HPC) is an obvious answer to the need for more computing power. Access to computer clusters is common now with HPC resources becoming available to all through local or national initiatives such as the UK supercomputing service HECToR. However, the transition from general computing, such as *R* Language and Environment for Statistical Computing, to parallel computing is not straight forward. Software application and tools have to be adapted to take advantage of the extra computing power. SPRINT aims to provide bioinformaticians using *R/Bioconductor* to analyse microarray data with easy access to HPC providing maximum performance but requiring minimal expert knowledge and minimal changes to existing *R* scripts. The SPRINT framework consists of an HPC harness and a library of parallelized R functions. SPRINT is very flexible; it runs on a range of HPC systems and allows the addition of user contributed functions. It handles functions that are trivial to parallelize, functions that are non trivial to parallelize and functions generating very large output.

The SPRINT parallel harness is written in *C* and uses the Message Passing Interface (**MPI**) library. It takes as input the *R* script and data to be analyzed. The use of the parallel IO support of the **MPI** library helps ensure that the results can be output in parallel providing great scalability. The use of **ff** objects from the **ff** package which allows the manipulation of large objects on file almost as if they were in memory, also removes the limitation on the size of the data that can be successfully analyzed. SPRINT can therefore handle very large amount of data increasing further its scalability.

The SPRINT library currently includes a parallel implementation of functions that have been highlighted as bottlenecks in the analysis of post genomic data by a user requirement survey. These include a Person pair-wise correlation (`cor` from **stats**) and a permutation test function (`mt.maxT` from **multtest**). Benchmarking runs on the HECToR Cray XT system have shown an almost perfect scaling to 512 processors.

References

J. Hill (2008). SPRINT: A new parallel framework for R. *BMC Bioinformatics*, 9, 558.

(2010). *SPRINT: A new parallel framework for R*,
http://www.r-sprint.org/.

(2010). *ECDF Edinburgh Compute and Data Facilities*,
http://www.ecdf.ed.ac.uk/.

(2010). *HECToR: UK National Supercomputing Service*,
http://www.hector.ac.uk/.

R-TREE: Implementation of Decision Trees using R

Margaret Miró-Juliá[1,⋆], Arnau Mir[1] and Monica J. Ruiz-Miró[2]

1. Departamento de Ciencias Matemáticas e Informática, Universidad de las Islas Baleares, SPAIN
2. Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, SPAIN

⋆Contact author: margaret.miro@uib.es

Keywords: R-software, classification, attribute selection, information gain.

The work presented here deals with the application of intelligent approaches to data analysis using R. The use of intelligent strategies allow for the construction of efficient patterns or structures for data classification. Knowledge discovery in Databases can be viewed as a three phase process: the data processing phase, the data mining phase and the evaluation phase. The main elements are the original data base, the transformed or processed data, the patterns or models of classified data and the knowledge extracted from the data. The data processing phase selects from the original data base a data set that focuses on a subset of attributes or variables on which knowledge discovery has to be performed; it also removes outliers and redundant information. The data mining phase converts this target data into useful patterns. Among all available classification methods, decision trees were selected for their simplicity and intuitiveness. The evaluation phase proves the consistency of the decision tree by means of a testing set.

Decision tree structures provide a common and easy way to organize data. Every decision tree begins with a root node. Each node in the tree evaluates an attribute from the data and determines which path it should follow. Classification using a decision tree is performed by routing from the root node until arriving at a leaf node. Decision trees can represent different types of data. The simplest and most familiar is numerical data. It is often desirable to organize nominal data as well. Nominal quantities are formally described by a discrete set of symbols. The type of data organized by a tree is important for understanding how the tree works at the node level. Recalling that each node is effectively a test, numeric data is often evaluated in terms of simple mathematical inequality, whereas nominal data is tested in Boolean fashion.

Decision tree induction algorithms function recursively. First, an attribute must be selected as the root node. In order to create the most efficient tree, the root node must effectively split the data. Each split attempts to pare down a set of instances (the actual data) until they all have the same classification. The best split is the one that provides what is termed as the most information gain. Information in this context comes from the concept of entropy from information theory developed by Claude Shannon. Although information has many contexts, it has a very specific mathematical meaning relating to certainty in decision making. Information can be expressed as a mathematical quantity as follows: $I = -\sum_{i=1}^{m} p_i \log_2 p_i$. Information gain is defined as information before splitting minus information after splitting. Ideally, each split in the decision tree should bring us closer to a classification.

The aim of this paper is to present an algorithm written with R: the R-tree algorithm. This algorithm will implement a decision tree by calculation of the information gain of the different attributes available in the data frame. This will allow for the selection of the attribute that offers the best split at each level of the tree. An important feature of this algorithm is that it handles both nominal and numerical data.

References

Fayyed U., Piatetsky-Shapiro G., Smyth P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence, AI Magazine Fall 96*, 37–54.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, vol. 1, 81–106.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 4, 379–423.

Using R for Active Learning and Self-assesment in an e-Learning Environment

Arnau Mir[1,⋆], Margaret Miró-Juliá[1] and Monica J. Ruiz-Miró[2]

1. Departamento de Ciencias Matemáticas e Informática, Universidad de las Islas Baleares, SPAIN
2. Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, SPAIN

⋆Contact author: arnau.mir@uib.es

Keywords: Education, self-assessment, R-software.

Over the past years, the way to teach mathematics and statistics has changed drastically. Nowadays it is unthinkable to teach statistics without the help of a statistical package. Moreover, the rapid pace of technological change has increased the importance of mathematics and statistics in science. Because of this, a project based on innovative learning and teaching techniques has been developed. The project aims to ensure that basic skills are mastered while developing conceptual thinking and modeling skills; to improve student's mathematical and statistical knowledge using an appropriate software; and to develop assessment methods that focus on higher level abilities and not just routine application of standard methods.

R-QUEST is an innovative system for e-learning based on self-assessment methods that guide the student throughout the learning process. It is an e-learning tool that helps students learn mathematics and statistics in an autonomous manner. The cornerstone of R-QUEST is R, R is used not only as a way to learn statistics but as a tool to better understand scientific concepts. At our university, the use of R has been introduced as a transversal competency in first year Biology and Biochemistry courses. Since we are dealing with first year students, it is out of the question to provide them with a manual and expect them to learn R without assistance. Moreover, programming languages can only be learnt through practice. The grading of weekly assignments can be quite a burden for the professors. But assessment plays a key role in determining student approaches to learning, since student's performance is driven by assessment. Therefore, it is essential to assess those aspects we want students to master.

Due to the above mentioned facts, R-QUEST a tool implemented on the Moodle Educational Platform has been developed. R-QUEST provides weekly lessons that consider one or two problems that must be solved using R. To help students learn R on their own (outside of the classroom), handouts on required commands are available on-line together with quizzes that allow students to practice and consolidate concepts. The number of questions in each quiz varies from 2 to 9, a typical quiz has 5 questions. In this first stage, each question has between 10 and 15 variants that are randomly assigned each time the quiz is opened by the student. The quizzes are adaptive in the sense that questions can be answered until the correct answer is found. At this moment, R-QUEST has a pool of 120 questions, each with 10-15 variants, our goal is to reach 300 questions.

Direct manual question introduction into the Moodle Educational Platform is a cumbersome and arduous task, specially when mathematical formulas are required. To ease this task, a Python package that transforms a quiz written in LaTeX into GIFT has been designed. The quizzes must be written following a specific format that provides a question template and specifies the parameters that handle the question variants. Once the set of quizzes is in GIFT format they are imported to the Moodle Educational Platform, where they are made available to the students.

The R-QUEST methodology makes learning R an easy chore. Furthermore, R-QUEST can be used by the student in an autonomous manner. R-QUEST helps the students to better understand topics of Mathematics and Statistics courses. The professor does not have to play an active role in the learning process since it is an e-learning tool implemented on the Moodle Educational Platform.

References

Crawley, Michael J. (2007). *The R Book*, John Wiley and Sons Ltd., West Sussex, England.

Lutz, M. (2009). *Learning Python, 4th edition*, O'Reilly Media, Inc. Sebastopol, CA.

Weiss, D. J. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educatoinal Measurement*,21, 361–375.

Integrated Development of the Software with Literate Programming: An Example in Plant Breeding

X. Mi , H.F. Utz, A.E. Melchinger*

Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, Germany

* Contact author: melchinger@pz.uni-hohenheim.de

Keywords: Multi-stage Selection, QTL, Resources Optimization, Portfolio Analysis, Literate Programming

The resource is always a limiting factor in a plant breeding program. The breeders try to optimize the allocation of limited resources, such as time, money, locations and so on, in order to enhance genetic gain. Furthermore, there is a continuous information explosion, in the present era: (1) generation of huge amount of information and data, such as molecular markers, biometric data of various genotypes in different environments, cross validation and meta analysis outputs and (2) new algorithms, new generations of computers, software and data bases developed for solving our problems and storing the data and information. To meet the challenges, large multi-inter-discipline scientific teams have to operate together to generate huge amount of data sets, manage the data sets and undertake the required analysis.

In this situation, an efficient software management technology is required: (1) to handle large data sets quickly, (2) should be well documented, (3) should be easily implementable and amenable to the incorporation with new features, such as C-port and object oriented, (4) should be popular with plenty users and (5) should be easily interpretable by the partners. We choose the management software Sweave and StatWeave with literate programming technology and this technology generates programs like writing a scientific paper, to build our package *PlabPortfolio*. This package is applied in plant/animal breeding for maximizing the gain of a multi-stage selection procedure under certain restrictions (e.g. for a given annual budget or certain risk limits at each stage). By implementing the R-package *mvtnorm*, which is one of the core packages of R and calculates the multi-variate normal distribution, the number of independent variables in the multi-normal regression model is increased from 3 to 1000. This makes it possible to use huge amounts of marker and QTL information. A variance analysis function, which achieves an efficient distribution of the budget between the QTL tests and field trials, will be implemented into the project.

References

Tallis, G. M. (1961). Moment generating function of truncated multi-normal distribution. *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, 23(1):223

Utz, F. (1969). Mehrstufenselecktion in der Panzenzuechtung. *PhD thesis*, University Hohenheim.

Knuth, Donald E. (1992). Literate Programming. California: Stanford University Center for the Study of Language and Information. ISBN 978-0937073803.

Hothorn, T., Bretz, F., and Genz, A. (2001). On multivariate t and Gauss probabilities in R. *R News*, 1(2):27-29.

Mi, X., Miwa, T. and Hothorn, T. (2009). Implement of Miwa's analytical algorithm of multi-normal distribution *R News*, accepted.

Mi, X. (2008). Model Selection Procedure with Familywise Error Rate Control for Binomial Order-Restricted Problems. *PhD thesis*, University Hannover.

R-Adamant: Financial technical analysis made easy

Fausto Molinari[⋆], **Martina Salvadori**[⋆⋆]

[⋆]Contact: fausto.molinari@ymail.com, [⋆⋆]Contact: martinasalvadori@libero.it

Keywords: Finance, Econometrics, GUI, HPC, Database

Most college students and resource groups know very well, how frustrating and time consuming is loading the code and the scripts manually and check the package dependencies of R.

After spending several years, looking for a more user friendly and scalable softwares we decided to develop our own package for R, customized for financial technical analysis, containing almost 620 formulas, organized as simple formulas, indexes, models, tests and oscillators.

Of course keeping with the r-project tradition, our package will be released under the open source GNU license.

Full compatibility is guaranteed for all the Os, supported by R and any part of the code may be easily debugged in order to fix any programming error.

The release for the UseR conference will contain almost 620 formulas and the first stable release publicly available will contain 1200 formulas.

The estimated target for the first class, business solution will be of 2304 formulas, in addition to frequent updates and high level technical and statistical support.

A customizable GUI will improve the productivity and ease of use for beginners and also advanced users, who can benefit of the all in one package, independent of any other 3rd party software.

Our team and the open source community will provide support to develop customized packages to Universities, research groups and individuals to address any specific need. All the released formulas will be fully documented in a specific manual.

It will also have a host of extensions and advanced features such as, parallel computing operations and interaction with grid and HPC applications.

Moreover, the data could be retrieved locally or remotely from csv or txt files and any other database source of the most popular databases (PostgreSQL, MySQL, SQLlite, MS SQL, Oracle, Sybase, etc...) and displayed in chart or as image file.

References

Johnston J. (1984), Econometric Methods - third edition, McGraw-Hill

Harvey A.C. (1993), Time Series Models 2nd Edition, The MIT Press.

Crawley M.J. (2007), The R Book, Wiley.

R Development Core Team (2009), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria
`http://www.R-project.org`

Eddelbuettel D. (2010), CRAN Task View: High-Performance and Parallel Computing with R
`http://cran.r-project.org/web/views/HighPerformanceComputing.html`

Parallel Computing with R using GridRPC

Junji Nakano[1,*], **Ei-ji Nakama**[2]

1. The Institute of Statistical Mathematics, Tokyo, Japan
2. COM-ONE Ltd. Ishikawa, Japan

[*]Contact author: nakanoj@ism.ac.jp

Keywords: Grid computing, Heterogeneous clusters, Ninf-G

Parallel computing becomes popular presently to achieve massive calculations. We have several techniques to perform parallel computing with R, for example, **Rmpi** and **snow**. These packages provide flexible and stable parallel computing mechanisms, and are especially suitable for a cluster of homogeneous computers in an intra-network. Grid computing, on the other hand, appeared recently to use simultaneously several heterogeneous computer clusters which are located far away and connected by Internet. **GridR** is one package to use R in such a grid environment. We propose to use GridRPC, a remote procedure call API for grid computing, for adding parallel computing functions to R. We provide a package to realize **snow**-like functions by utilizing Ninf-G, a reference implementation of the GridRPC API.

References

Ninf administration group (2008). Ninf: A Global Computing Infrastructure, `http://ninf.apgrid.org/`.

Schmidberger, M, Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L. and Mansmann, U. (2009). State of the Art in Parallel Computing with R. *Journal of Statistical Software*, August 2009, Vol. 31, Issue 1.

Prototyping Preventive Maintenance Tools with R

Erich Neuwirth[1,*] **, Julia Theresa Csar**[1]

1. University of Vienna
* Contact author: erich.neuwirth@univie.ac.at

Keywords: preventive maintenance, incremental quantile estimation, condition monitoring

Machinery in factories is constantly monitored. A lot of data, like vibrations, is collected to track the machines' performance and condition. In this paper we present a method of timely analysis and visualization of these data. The representation derived from the data can be used to detect unusual behavior that could lead to malfunction if unnoticed, or for long-time monitoring to detect and track erosion or abrasion.

Our algorithm is based on John Chamber's algorithm for Incremental Quantile Estimation in univariate data. We adapted the method to cope with multivariate data.

As an example, we will use data collected from a coffee maker. The noise of the crushing mill of the coffee machine was recorded. Energy content of different frequency regions was used as the input to our multivariate algorithm. After applying the algorithm to this dataset, confidence intervals were calculated for each status. Now the grinding texture of the coffee machine can be evaluated based on its noise while running and the critical coffee bean charging level can be recognized early. The visualization and calculation was made using *R*.

Application to different and more complex machinery provides the possibility of preventive maintenance beyond the routine maintenance schedule and detection of erosion during periods of operation.

An important aspect of our algorithm is that we compute a condensed representation of the data in real time and therefore this algorithm works with relatively low storage requirements.

References

John M. Chambers (2006). Monitoring Networked Applications With Incremental Quantile Estimation. *Statistical Science,* 2006, Vol. 21, No. 4, 463-475.

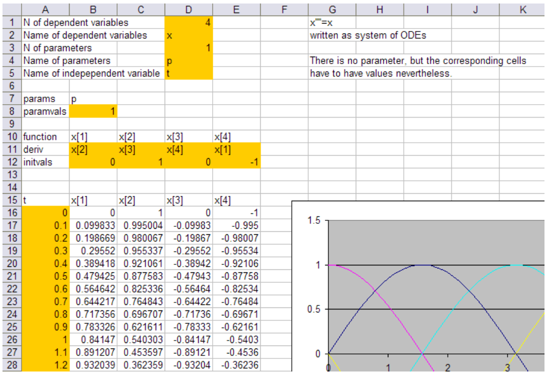
Investigating ODEs with R and spreadsheets

Erich Neuwirth [1*]

1. University of Vienna
* Contact author: erich.neuwirth@univie.ac.at

Keywords: numerical integration, interactively controlling model parameters, differential equations, integrating R and spreadsheets

Spreadsheets offer nice interactive tools for investigating simple numerical problems including ordinary differential equations (ODEs), but they lack sophisticated methods, especially integrations methods for ODEs. *R* has some very powerful packages for ODEs, especially **odeSolve** and **deSolve**. Using spreadsheets add-ins like **RExcel**, one can study the quality of numerical integrator and also perform sensitivity analysis on parameters and initial conditions for ODEs. The user studies ODEs in an environment as displayed here:



	A	B	C	D	E	F	G	H	I	J	K
1	N of dependent variables			4			x''''=x				
2	Name of dependent variables			x			written as system of ODEs				
3	N of parameters			1							
4	Name of parameters			p			There is no parameter, but the corresponding cells				
5	Name of indepependent variable			t			have to have values nevertheless.				
6											
7	params	p									
8	paramvals	1									
9											
10	function	x[1]	x[2]	x[3]	x[4]						
11	deriv	x[2]	x[3]	x[4]	x[1]						
12	initvals	0	1	0	-1						
13											
14											
15	t	x[1]	x[2]	x[3]	x[4]						
16	0	0	1	0	-1						
17	0.1	0.099833	0.995004	-0.09983	-0.995						
18	0.2	0.198669	0.980067	-0.19867	-0.98007						
19	0.3	0.29552	0.955337	-0.29552	-0.95534						
20	0.4	0.389418	0.921061	-0.38942	-0.92106						
21	0.5	0.479425	0.877583	-0.47943	-0.87758						
22	0.6	0.564642	0.825336	-0.56464	-0.82534						
23	0.7	0.644217	0.764843	-0.64422	-0.76484						
24	0.8	0.717356	0.696707	-0.71736	-0.69671						
25	0.9	0.783326	0.621611	-0.78333	-0.62161						
26	1	0.84147	0.540303	-0.84147	-0.5403						
27	1.1	0.891207	0.453597	-0.89121	-0.4536						
28	1.2	0.932039	0.362359	-0.93204	-0.36236						

The symbolic representations of equations in spreadsheet programs and in the package **deSolve** is quite different, and the *R* package notation is much closer to textbook representations. Therefore we designed a dialog baesd user interface which allows the user to specify the problem in notation close to textbook standards. The spreadsheet tool heavily draws upon the power odd the numerical methods from the R packages, but also builds upon the convenience of spreadsheets (easy to use controls and self updating graphs) to create a simple to use learning environment for a well defined mathematical topic.

References

Erich Neuwirth . Deane Arganbright, *The Active Modeler: Mathematical Modeling with Microsoft Excel.* Duxbury Press, 2003

Karline Soetaert, Thomas Petzoldt, R. Woodrow Setzer (2010). *Solving Differential Equations in R: Package deSolve*. Journal of Statistical Software, 33(9), 1–25.

Using ontologies for R functions management

Olivier Corby[1], Caroline Domerg[2], Juliette Fabre[2], Catherine Faron-Zucker[3], Alexandre Granier[2], Isabelle Mirbel[3], Vincent Negre[2], Pascal Neveu[2⋆], Anne Tireau[2]

1. INRIA Sophia Antipolis Mditerrane, Sophia Antipolis, France
2. INRA, Montpellier, France
3. I3S, University of Nice, Sophia Antipolis, France

⋆Contact author: Pascal.Neveu@supagro.inra.fr

Keywords: Capitalization of R Functions, Ontology, Knowledge Management, Semantic Web.

The research work of biologists often requires the development of a lot of scripts or functions to manipulate and analyse experimental data. In the laboratory LEPSE specialized on the analysis and modelling of plant responses and adaptation to variable environmental stresses, dozens of R functions are produced every year, concerning various fields such as genetic analyses, high throughput data phenotyping or environmental interactions and involving several databases. As a result, there is an important turn-over of function authors and users which generates different problems like re-using, sharing or understanding of these functions. In this context, in the framework of the DESIR project[1], we have initiated a knowledge management action aiming to capitalize, organize, share and valorize these functions through the development of a knowledge-based repository of these functions.

Given the great diversity of the functions produced, their associated documentation is heterogeneous and it is not pertinent to organise them into packages. We decided instead to index them with some formalized knowledge describing them, in order to retrieve them by formal reasoning. For this purpose, we developed an ontology providing a controlled and structured vocabulary that captures the concepts and properties necessary to describe R functions. This ontology comprises concepts and properties to describe functions – like "Author", "Intention", "Argument", "Value" – as well as the relations between functions – like "hasForRCoreCall", "canBeUsedAfter", "isAdaptedFrom", "looksLike". As a result functions can be retrieved according to a wide range of criterii: thier author and/or the graphics produced, their intention(s) (e.g. perform multidimensional exploration), the function(s) they call (e.g. the "lm" R core function or a specific function of the repository) –more generally, it is relevant to generate the call graph of one function to understand it–, the functions from which they are adapted –this makes easier the maintenance of the repository–, the functions after or before which they should be used –this helps to construct chainings of treatments–, their similarity with other functions, etc.

To formalize both the ontology and the annotations of R functions, we adopted the Semantic Web models: The annotations are represented into the Resource Description Framework[2] (RDF) and the ontology in the Ontology Web Language[3] (OWL). As a result we are able to semantically retrieve R functions by expressing queries in the SPARQL language[4]. We developed a semantic web application for the repository and search of annotated R functions. It relies upon the semantic engine Corese (Corby et al. 2004) dedicated to ontological query answering on the Semantic Web: Corese enables to interpret and process SPARQL queries on RDF annotations and OWL ontologies. Our application provides an environment for (1) *storage and annotation*: a prototype of Web user interface allows authors to upload R functions (one function per file) and to describe them in a few minutes; and (2) *powerful search*: users can find and get R functions with a global and accurate understanding and receive suggestions to support their search.

To conclude, we have built a semantic repository of annotated R functions to centralize and share R functions for biologists. It capitalizes expert know-hows that would otherwise oftenly be lost or become non-usable because of a lack of documentation and description. We are convinced that this kind of repository developed for the LEPSE could benefit a much wider community of R function authors and users and be adapted to handle other programming languages.

References

O. Corby, R. Dieng-Kuntz and C. Faron-Zucker (2004). Querying the Semantic Web with Corese Search Engine, ECAI 2004, Proc. of the 16th Eureopean Conference on Artificial Intelligence, (Valencia, Spain), August 2004, 705–709.

[1] http://www-sop.inria.fr/edelweiss/projects/desir/wakka.php?wiki=ColorDesirHomePage

[2] http://www.w3.org/RDF/

[3] http://www.w3.org/2004/OWL/

[4] http://www.w3.org/TR/rdf-sparql-query/

David v. Goliath: How to build an R presence in a corporate SAS environment

Derek McCrae Norton[1,*]

1. Equifax, Inc

*Contact author: Derek.Norton@equifax.com

Keywords: Business Analyitcs, Open Source, SAS

SAS has been growing and gaining market share since 1976. In 1991 that began to change as a new competitor, R, was born. Now 19 years later R is a viable alternative to SAS, but there is still resistance in the corporate environment.

How does one break through the resistance?

By following some simple steps (simple doesn't necessarily mean easy), you can build a strong R following at your corporation. This work will provide those steps as well as some methods to implement them.

- Start Small.
- Spread the word.
- Show the ROI.
- Focus on what R does better than SAS.
- . . . Tune in to find more.

Remember, Goliath was toppled by one stone from one small boy. . . There was an entire army waiting, but the tipping point was one boy.

Using R for data management in ecophysiology Information Systems

Caroline Domerg[1], Juliette Fabre[1], Vincent Nègre[1*], Anne Tireau[2]

1.UMR LEPSE, INRA, Montpellier, France
2. UMR MISTEA, INRA, Montpellier, France
* Contact author: vincent.negre@supagro.inra.fr

Keywords: Information System, Database, Web interface, Data Management, Plant Biology

In the Laboratory of the Ecophysiology of Plants under Environmental Stresses (LEPSE[1]) at INRA, Montpellier (France) three experimental set-ups allow the study of the effect of genotype x environment interactions on plant ecophysiological traits: (i) a field network for maize populations, (ii) the PHENODYN semi-automated platform for maize phenotyping (high-throughput) including two environments, a greenhouse and a growth chamber, and (iii) the PHENOPSIS automated platform for *Arabidopsis thaliana* phenotyping (high-throughput).

As these experimental devices generate an important amount of data of different types (numeric, images) and natures (phenotypic, environmental, genetic), information systems were developed around each device for the collection of data and metadata, their storage and organization in *MySQL* databases, and their extraction, visualization and analysis via Web interfaces developed in *PHP* and *HTML* (Cincalli DB[2], Phenodyn DB[3] and Phenopsis DB[4]).

We have used *R* and the **RODBC** package for the management of data at the different levels of the information systems. *R* scripts were developed to: (i) automatically insert online data issued from the platforms (growth measurements, weights of the pots, environmental data and irrigation data), (ii) manually insert offline datasets (such as phenotypic data measured on plants) via Web interfaces, (iii) transform datasets extracted from the databases in order to display them and render them available in downloadable files via Web interfaces, (iv) provide online tools for data visualization (environmental kinetics, growth curves) as a support for experiment monitoring or data exploration and (v) perform data analyses (such as growth modeling) and calculations of elaborated data. *R* scripts are either automatically ran for data insertion, or called in *PHP* programs of the Web interfaces for data extraction and transformation, data visualization and analysis. Some of them are available for download on the Web sites.

The poster presents the three experimental set-ups, the organization of their information systems and how we have used *R* at the different levels of these information systems.

References

Fabre J. (2008). Développement d'un Système d'Information de phénotypage d'Arabidopsis thaliana. *Cahier des techniques de l'Inra*, 65, 31-461.

[1] http://www1.montpellier.inra.fr/ibip/lepse/

[2] http://bioweb.supagro.inra.fr/cincalli/

[3] http://bioweb.supagro.inra.fr/phenodyn/

[4] http://bioweb.supagro.inra.fr/phenopsis/

UseR abstract 2010

The use of R, S+ and Spotfire in the Oracle Life Sciences Data Hub Environment

Michael O'Connell, PhD, TIBCO Spotfire
Subra Subramanian, Oracle Life Sciences

The Oracle Life Sciences Data Hub (LSH) is an environment that enables reproducible research and development across Life Science functional areas. It enables the use of analytic software including SAS, R, S+ and Spotfire. These analytic software applications are enabled in the LSH environment through "adapters". Such adapters enable access to data in LSH in a reproducible and refreshable way. TIBCO Spotfire version 3.1 includes tight connectivity to R and S+. The product incorporates a script repository where statisticians and others can register R and S+ scripts, and Spotfire authors can search, retrieve and utilize these scripts when setting up a Spotfire analysis workbook. This presentation will outline a systems approach to R, S+ and Spotfire analysis in the LSH environment. The R and Spotfire adapters to LSH will be described, along with the R/S+ repository and workflow in Spotfire. They system will be illustrated with the analysis of pharmaceutical research and development data using LSH, R/S+ and Spotfire.

Web development with R

Jeroen Ooms[1,2,*]

1. UCLA Dept. of Statistics
2. Revolution Computing

*Contact author: jeroen@revolution-computing.com

Keywords: Web Applications, Cloud computing

Web applications are becoming increasingly important in offering software to the user. The main advantage above classical client-based software is that new or improved applications can easily be made available to a wide audience. Furthermore web applications are by design server-based, and therefore make more efficient use of resources and are easier to maintain. R is a particularly suitable back-end for web applications. A user-friendly graphical interface reduces the barrier for many researchers, and R's high computing requirements would benefit greatly from a scalable environment. Furthermore R integrates nicely with databases and latex, and which opens the door for online data management, data sharing and reporting services.

This presentation is meant to give an overview of different approaches, architectures and software that could be of interest when embedding R in the web. We will go over some tools that you can use to connect a HTTPD to R, show their advantages and disadvantages, and illustrate some examples of different projects that have embedded R in web applications. Furthermore, the speaker will share some of his personal experiences and vision for the future.

References

Jeffrey Horner (2009). rapache: Web application development with R and Apache, `http://biostat.mc.vanderbilt.edu/rapache/`.

Simon Urbanek. Rserve - Binary R server, `http://www.rforge.net/Rserve/`.

Revolution Computing, `http://www.revolution-computing.com/`.

Jeroen Ooms, `http://www.jeroenooms.com`.

Red-R: A visual programming and data analysis framework

Anup Parikh[1,*] and Kyle R Covington[1]

1. Baylor College of Medicine, One Baylor Plaza, Houston Texas, USA
* Contact author: anup.parikh@gmail.com

Keywords: graphical user interface, visual programming, bioinformatics

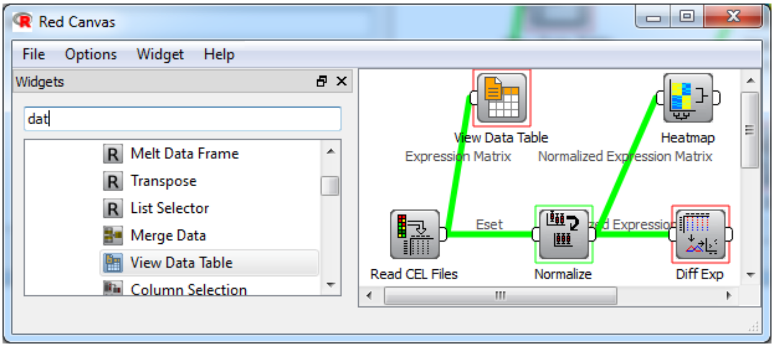
Cross-discipline collaboration is major driving force for new innovation. As an example, the collaboration between bioinformaticists (computational users) and biologists (non-computational users) to generate and analyze high-throughput datasets has revolutionized the life sciences. However, as the volume of data and technical complexity of the analysis increases, non-computational users can not participate in the data exploration and analysis. Furthermore due to the complexity, data analysis pipelines are difficult to understand and share, even with other computational users. This lack of powerful yet user friendly tools greatly limits analysis interpretability and the ability of non-computational users to participate in the data analysis process.

To address these challenges we have developed Red-R, a user friendly visual programming and data analysis framework for *R*. Red-R makes the advanced functionality of *R* available to the non-computational users by hiding the computational complexity behind a visual programming interface. Analyses are performed by visually linking a series of widgets together that read, manipulate, and interactively display data. These pipelines, representing both the data and analysis, can be easily shared with others. Creating a visual representation of the analysis greatly increases interpretability and gives non-computational users the ability to explore the data and analysis without understanding *R* code. Along with the visual representation, all the R code is readily available for review. This functionality in Red-R can create more effective collaborations by allowing users to perform complex analyses and share them so that others can easily understand the analysis and further explore the data.

Red-R is an extension of Orange (http://www.ailab.si/orange), a data mining framework written in *Python* and *Qt*. Red-R accesses all the functionality and data in *R*, using the *Python* interface for *R* provided by RPy (http://rpy.sourceforge.net). This framework is highly flexible and can be extended to include virtually all the functionally *R* currently offers. To facilitate this, we have created a widget that parses the *R* code of a given function and generate the basic *Python*/*Qt* code to quickly create a new widget in Red-R. In addition to creating a visual interface to *R* functions, Red-R can harness the power of Qt graphics to create interactive visualizations.

The current version of Red-R provides functionality for data manipulation, basic statistics and advanced bioinformatics analysis from the **Bioconductor** package. Red-R also provides widgets to import existing *R* sessions and execute any *R* code. We hope to gain community support in extending the repository of widgets and providing analysis pipeline for common tasks. To that aim, our website (http://www.red-r.org) hosts documentation, example schemas, and shared resources for Red-R.

Red-R Interface: Differential expression analysis of Affymetrix microarrays.



Software for the joint modelling of longitudinal and survival data: the JoineR package

Pete Philipson[1,⋆], Ruwanthi Kolamunnage-Dona[2], Inês Sousa[3,4], Peter Diggle[4], Robin Henderson[5], Paula Williamson[2] and Gerwyn Green[4]

1. University College London, United Kingdom
2. University of Liverpool, United Kingdom
3. University of Minho, Portugal
4. Lancaster University, United Kingdom
5. Newcastle University, United Kingdom

⋆Contact author: p.philipson@ucl.ac.uk

Keywords: joint modelling, longitudinal, survival, random effects, transformation model

The joint modelling of longitudinal and survival data has seen a surge of interest in recent years. Wulfsohn and Tsiatis (1997) developed the methodology for a random effects joint model, and their work was built upon by Henderson et al (2000). This model considers a linear mixed effects sub-model for the longitudinal data and a Cox-based sub-model for the survival data, linking the sub-models through common random effects. A subsequent alternative approach (Diggle et al 2007) considered a fully parametric transformation model in which a transformed version of the time-to-event outcome is treated as conditionally normally distributed given the repeated measurements.

Software to accommodate joint models has, however, been outstripped by the methodological advances to the extent that the first R software emerged only recently (see package **JM**). The availability of electronically linked healthcare databases suggest that the lacunae in readily available software for joint modelling is set to become more apparent. The **JoineR** package attempts to fill some of this void, building on established separate modelling functions `coxph` and `lme` where appropriate.

The unique contribution of **JoineR** is in allowing flexibility in the form of the latent association that links the longitudinal and event time processes in the random effects model, as well as being the first package to provide functionality to fit the transformation model. Accompanying these novel joint modelling functions are several useful functions to transform, view and simulate data. The features of **JoineR** will be illustrated on both simulated and real data.

Aspects of the package have been trialled in a series of successful workshops and details of the overarching project can be found on the JoineR website (http://www.liv.ac.uk/joine-r/index.htm).

References

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53, 330–339.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1, 465–480.

Diggle, P., Sousa, I. and Chetwynd, A. G. (2007). Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture. *Statistics in Medicine*, 27, 2981–2998.

Rizopoulos, D. JM: Joint Modelling of Longitudinal and Survival Data `http://cran.r-project.org/web/packages/JM/index.html`.

The **cg** package for comparison of groups

Bill Pikounis[1,*] **, John Oleynick**[1]

1. Johnson & Johnson Pharmaceutical Research & Development
* Contact author: bpikouni@its.jnj.com

Keywords: resistance, robustness, censoring, strategy

In research of medicines, the comparison of treatments, test articles, conditions, administrations, etc. is very common. Studies are done, and the data are then most often analyzed with a default mixture of equal variance t-tests, analysis of variance, and multiple comparison procedures. But even for a presumed one-factor linear model to compare groups, more often than not there is the need to accomodate data which is better suited for expression of multiplicative effects, potential outliers, and limits of detection. Base *R* and contributed packages provide all the pieces to develop a much more comprehensive strategy. Such an approach includes exploration of the data, fitting models, formal analysis to gauge the magnitude of effects, and checking of assumptions. The cg package is developed with those goals in mind, using a flow of wrapper functions to guide the full analysis and interpretation of the data. Examples from our non-clinical world of research will be used to illustrate the package and strategy.

A Tool for Quantitative Analysis of Proteomics Data

Ashoka D. Polpitiya[1*], **Navdeep Jaitly**[2], **Konstantinos Petritis**[1]

1. Center for Proteomics, Translational Genomics Research Institute, Phoenix, AZ 85004.
2. Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4.

*Contact author: ashoka@tgen.org

Keywords: proteomics, normalization, statconnDCOM, R,C#

Proteomics involves the study of protein content of an organism, a complement of its genome. Mass spectrometry coupled with liquid chromatography techniques are the most popular methods used for proteomics studies. Although there are number of tools available for high-throughput microarray data processing (Gentleman et al. 2004), they are not ideally suited to address the issues specific to proteomics data. For example, in the most common approach called the "Bottom-up" proteomics where proteins are first enzymetically digested to obtain smaller peptides that are easier to detect in a mass spectrometer, inferring the protein quantities from the observed peptides is a unique challenge. Another major issue is the extent of missing values that is largely due to the amount of species near the threshold for detection leading to unbalanced datasets.

Inferno is a free, open source statistical tool designed to address the unique challenges associated with large scale proteomics studies. The graphical user interface (GUI) of Inferno is implemented in C# language and the core algorithms are implemented in R (R Development Core Team, 2009). statconnDCOM (Baier et al., 2010) is used for the connectivity between the .NET environment and R. Inferno runs on Microsoft Windows platform.

Inferno features many statistical plots such as boxplots, histograms, QQ plots, and correlation diagrams. A set of normalization algorithms such as LOWESS and linear regression, is also implemented for removing any systematic variation in the data. It also presents few methods to infer protein abundances from the observed peptide abundances and a comprehensive ANOVA scheme based on the **car** package (Fox, 2002), for selecting significant proteins in an experiment.

Inferno is available at http://inferno4proteomics.googlecode.com.

References

R Development Core Team (2009). R: A Language and Environment for Statistical Computing, http://www.R-project.org.

Baier T., Neuwirth E. (2010). Statconn DCOM, http://rcom.univie.ac.at/.

Gentleman RC, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics, Genome Biol., 5(10):R80.

Fox, J. (2002). An R and S-PLUS Companion to Applied Regression. Sage Publications, Thousand Oaks, CA, USA.

Statistical Issues in Accessing Brain Functionality and Anatomy

Jörg Polzehl[1,*], **Karsten Tabelow**[1,2]

1. Weierstrass Institute for Applied Analysis and Stochastics
2. DFG Research Center MATHEON “Mathematics for key technologies”

[*]Contact author: polzehl@wias-berlin.de

Keywords: imaging, functional MRI, diffusion weighted MRI

Among the imaging techniques used in neuroscience there are two, functional MR imaging (fMRI) and diffusion weighted MR imaging (DWI), that especially boosted the research field in the last two decades. Both noninvasive methods allow for an in-vivo analysis of brain activity and brain anatomy.

fMRI focuses on localizing cognitive functionality within the brain gray matter. In a typical designed experiment the proband/patient is exposed to specific stimuli. The recorded data then consists of time series of three-dimensional image volumes where in some voxels (volume elements) a signal response corresponding to the stimuli is observed. This response is related to brain activity via the blood oxygenation level dependent (BOLD) effect. Statistical issues in the fMRI data analysis include adequate modeling, multiple testing, noise reduction and comparisons in groups of subjects, to name just a few.

DWI probes microscopic structures well beyond typical image resolutions through water molecule displacement. It can be used in particular to characterize the integrity of neuronal tissue in the central nervous system. Diffusion weighted data can be viewed as five-dimensional, i.e., living on a 3D grid of voxels with information on water diffusivity measured in N_{grad} directions on the sphere, characterized by two angles. The most common model for such data is the diffusion tensor model (DTI). This model is used to describe main direction of diffusivity and to derive clinically relevant diagnostical quantities. Generalizations focus on modeling and estimation of an orientation density function (ODF) at each voxel. Information from both the tensor model and ODF based models can be used for fiber tracking, characterization of fiber bundles and finally in connection with results from fMRI studies for accessing brain connectivity.

We will illustrate some of the statistical problems that arise and show how they can be addressed within R using packages **fmri** and **dti**. Some open problems will also be discussed.

References

Nicole A. Lazar (2008). *The Statistical Analysis of Functional MRI data*, Springer Series for Biology and Health.

Susumu Mori (2007). *Introduction to Diffusion Tensor Imaging*, Elsevier.

Jörg Polzehl and Karsten Tabelow (2009), Structural Adaptive Smoothing in Diffusion Tensor Imaging: The R Package **dti**, *Journal of Statistical Software*, 31, 1-23.

Karsten Tabelow, Valentin Piëch, Jörg Polzehl and Henning U. Voss (2009), High-resolution fMRI: Overcoming the signal-to-noise problem, *Journal of Neuroscience Methods*, 178, 357–365.

Marketing Analytics in R

Jim Porzak
Ancestry.com
San Francisco, CA

Marketing analytics can be defined as the application of well known methods from statistics, data mining, and visualization to optimize marketing efforts – admittedly a rather broad topic.

This talk describes a basic tool kit aimed at modern marketing analysts who are expected to have basic statistical knowledge and are skilled in Excel and the other MS Office packages in the Windows environment, but are without any experience with R. These folks are at the front line of implementing, evaluating, and explaining the results of marketing campaigns and tests, both in the on-line and off-line worlds. Other aspects of marketing analytics, especially "customer intelligence," segmentation, and predictive modeling have been discussed at prior useR! conferences.

Fundamental to campaign and test analytics is an appropriate data structure to support the detailed description of campaigns with multiple test cells and, perhaps, multiple responses within cells. The tests could be simple "A/B" tests or more interesting factor-based "MVT" designs. Since no campaign is ever done in isolation, it is also important to include between-campaign metadata for comparisons and rollups across media, offers, segment, etc.

The tool kit supports the all the basic steps a campaign analyst needs to go through:

- Initial reality check that available sample sizes will be sufficient for useful conclusions,
- Setting up the campaign test design and metadata,
- Random assignment of subjects to test cells,
- Detailed analysis of campaign results, and
- Visualization of campaign results for executive presentations.

The tool kit is designed modularly with well defined interface layers between the internal and external data sets and between the core functions and the user interface. Initially, we use Ian Fellows' Deducr package for the user interface. A web based interface will be a future option.

IsoGeneGUI: a graphical user interface for analyzing dose-response studies in microarray experiments

Setia Pramana[1]**, Dan Lin**[1]**, Philippe Haldermans**[1]**, Ziv Shkedy**[1]**, Tobias Verbeke**[2]

1. Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, Diepenbeek, Belgium
2. OpenAnalytics BVBA, Heist-op-den-Berg, Belgium
*Contact author: setia.pramana@uhasselt.be

Keywords: dose-response, microarray , monotonic trend, graphical user interface.

The main objectives in drug-discovery studies in the pharmaceutical industry are to find a safe and efficacious dose or a dose range, and to establish a dose response relationship. The emerging of biomedical technologies leads to an integration of dose-response studies with microarray technologies. In this setting, the response are gene-expressions measured at a set of increasing dose levels. The aim of such a study is to identify a set of genes with monotonic increasing/decreasing mean expressions at increasing doses. Lin et al. (2010) discussed several testing procedures for dose-response studies of microarray experiments. These testing procedures which take into account the order restriction of means with respect to increasing doses, are Williams (Williams, 1971 and 1972), Marcus (Marcus, 1976), the likelihood ratio test (Barlow et al. 1972, and Robertson et al. 1988), the M statistic (Hu et al., 2005), and the modified M statistic (Lin et al., 2007).

The aforementioned methods are implemented in an R package called **IsoGene** (Pramana et al., 2010). The **IsoGene** package requires the user to have basic knowledge of R. To overcome this limitation, a user friendly interface called **IsoGeneGUI** is developed. It is a menu driven package and data analysis can be perform simply by selecting options from the menus of the package. The **IsoGeneGUI** is developed by using the R-Tcl/Tk interface implemented in the **tcltk** package (Dalgaard, 2001).

The inference is based on resampling methods to obtain the *p*-values (Ge et al., 2003). The multiplicity adjustment includes: Bonferroni, Holm (1979), Hochberg (1988), and Sidak procedures for controlling the family-wise Type I error rate (FWER), and the Benjamini-Hochberg (BH-FDR, Benjamini and Hochberg, 1995) and Benjamini-Yekutieli (BY-FDR, Benjamini and Yekutieli, 2001) procedures for controlling the False Discovery Rate (FDR), and the Significance Analysis of Microarrays (SAM, Tusher et al., 2001).

The package provides three options in its Analysis menu (1) Likelihood ratio test statistic, (2) Permutations, and (3) Significance Analysis of Microarrays (SAM). The package produces some default graphical displays as well as user-defined graphical output which can be save as different image types. The analysis results can be saved in R workspaces and/or excel files. The IsoGeneGUI package can be obtained from:

- R-forge site: https://r-forge.r-project.org/projects/isogenegui/,
- **IsoGene** project site: http://www.ibiostat.be/software/IsoGeneGUI/index.html.

At the **IsoGene** project site, users manuals and example data sets can be downloaded. Moreover, illustrative examples are provided at the site as well.

References

Lin, D., Shkedy, Z., Yekutieli, D., Amaratunga, D., and Bijnens, L. (Editors) (2010) Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R, Springer to be published in 2010.

Lin, D., Shkedy, Z., Yekutieli, D., Burzykowski, T., Gohlmann, H.,De Bondt , A., Perera, T., Geerts, T., and Bijnens L. Testing for trends in dose-response microarray experiments: A comparison of several testing procedures, multiplicity and resampling-based inference. Statistical Applications in Genetics and Molecular Biology 6(1) (2007), Article 26.

Pramana, S., Lin, D., Haldermans, P., Shkedy, Z., Verbeke, T., Gohlmann, H., Bondt, A. D., Talloen, W., and Bijnens., L. (2010). Analysis of dose-response studies in microarray experiments using the R IsoGene package. Journal of Statistical Software.

Navigator: creating and managing complex graphics in a production environment

Richard Pugh, John James

The simulation of the behaviour of chemicals in human and animal bodies is complex. There are many programs that assist in this understanding but understanding the output from these is also complicated. Naturally R is widely used in helping the interpretation of the outputs.

However, using R it is also easy to generate hundreds of graphs as part of project work. When R is used in production it is essential to manage the output: otherwise which graph was produced when gets lost is the urgent rush to finish the work. Then three months or three years later there is the need to review what happened: without a clear and explicit audit trail the events leading to a conclusion are lost.

The Navigator provides a web browser interface for managing, storing, viewing and printing R reports. After a user has executed the analysis, the Navigator application retrieves the various files and persists them in a relational database. By using Navigator a large number of runs can be assessed and characterised within a short space of time and in a standardized way. Reports can be printed in Microsoft RTF format to provide a permanent record of the information stored within the Navigator database.

This talk illustrates an approach to the industrialization of R and how it can be moved from the experimental bench to an industrial toolset.

TravelR: Travel Demand Modeling in R

Jeremy Raw[1]

1. Contact author: jeremy.raw@earthlink.net

Keywords: Travel Demand Modeling Highway Assignment

Travel demand modeling is widely applied for analysis of major transportation investments and air quality conformity. Yet this field has remained conservative in its methodology, and unable to keep pace with evolving policy analysis goals (TRB, 2007). Since the US Department of Transportation ended development of its Urban Transportation Modeling System (UTMS) in the mid-1980s, most agencies have depended on expensive and inflexible proprietary software platforms. Where advanced travel models have been implemented, they have been large customized software systems, with all the expense and risk inherent in such efforts (VDOT, 2009).

Travel model developers have begun experimenting with open source tools. The R statistical environment has drawn attention because of its fast vector and matrix processing, powerful graphics, and vast collection of statistical tools for estimating and applying statistical models. The Oregon Department of Transportation has implemented considerable portions of their travel demand modeling system in R (ODOT, 2010). The author has used R for model estimation at the Virginia Department of Transportation.

Until now, however, it has not been possible to build complete travel demand models in R, due primarily to the lack of solutions for the equilibrium traffic assignment problem (which allocates travel demand on congested transportation networks). Travel models written in R have remained dependent on proprietary traffic assignment routines.

This presentation introduces **TravelR**, which implements common algorithms for highway assignment and other basic travel demand modeling functions. The highway assignment function supports features that are important to practical large-scale models: turn penalties (used to simplify network coding), multiple interacting vehicle classes (such as cars and trucks operating with different cost functions on different network subsets), and select link analysis (to extract travel patterns through portions of the network). Written in C++ and R, it handles large problems with reasonable speed. **TravelR** also inter-operates with the R packages **igraph** and **sp**, allowing access to R tools for graph algorithms and geospatial analysis.

In addition to providing tools to build entire travel demand models in R, **TravelR** makes possible an R-based framework for travel demand modeling in which new methods can be explored, tested, and rapidly put into practice. The author hopes that an open source platform for travel demand modeling in R will accelerate the investigation and adoption of new travel modeling techniques that can respond to increasingly challenging environmental, social and financial issues in transportation systems.

References

Oregon Department of Transportation (ODOT) (2010). *(Travel) Modeling with R*, `http://www.oregon.gov/ODOT/TD/TPAU/R.shtml`.

Transportation Research Board (TRB) (2007). *Metropolitan Travel Forecasting: Current Practice and Future Direction*, Special Report 288, `http://onlinepubs.trb.org/onlinepubs/sr/sr288.pdf`

Virginia Department of Transportation (VDOT) (2009). *Implementng Activity-Based Models in Virginia*, `http://www.virginiadot.org/projects/resources/vtm/VTMRP09-01_Final.pdf`

NppToR: R Interaction for Notepad++

Andrew Redd[1,⋆]

1. Texas A&M University, Department of Statistics, College Station, TX, USA
⋆Contact author: aredd@stat.tamu.edu

Keywords: Editors, Interaction, Development

Notepad++ is one of the fastest, most powerful and most popular code and text editors for Windows. NppToR is a utility that adds interaction between Notepad++ and the windows R GUI. It can also define syntax highlighting rules and auto-completion. This enables Notepad++ to be used as an alternative to the built in editor on Windows.

References

Redd (2010). NppToR Project Pages,
`http://npptor.sourceforge.net`.

R and BI – Integrating R with Open Source Business Intelligence Platforms Pentaho and Jaspersoft

David Reinke, Steve Miller

Keywords: business intelligence

Increasingly, R is becoming the tool of choice for statistical analysis, optimization, machine learning and visualization in the business world. This trend will only escalate as more R analysts transition to business from academia. But whereas in academia R is often the central tool for analytics, in business R must coexist with and enhance mainstream business intelligence (BI) technologies. A modern BI portfolio already includes relational databeses, data integration (extract, transform, load – ETL), query and reporting, online analytical processing (OLAP), dashboards, and advanced visualization. The opportunity to extend traditional BI with R analytics revolves on the introduction of advanced statistical modeling and visualizations native to R. The challenge is to seamlessly integrate R capabilities within the existing BI space. This presentation will explain and demo an initial approach to integrating R with two comprehensive open source BI (OSBI) platforms – Pentaho and Jaspersoft. Our efforts will be successful if we stimulate additional progress, transparency and innovation by combining the R and BI worlds.

The demonstration will show how we integrated the OSBI platforms with R through use of RServe and its Java API. The BI platforms provide an end user web application which include application security, data provisioning and BI functionality. Our integration will demonstrate a process by which BI components can be created that prompt the user for parameters, acquire data from a relational database and pass into RServer, invoke R commands for processing, and display the resulting R generated statistics and/or graphs within the BI platform. Discussion will include concepts related to creating a reusable java class library of commonly used processes to speed additional development.

The Convergence Properties of the BLP (1995) Contraction Mapping and Alternative Algorithms in R

Jo Reynaerts[1,⋆], **Ravi Varadhan**[2], **John C. Nash**[3]

1. LICOS, Katholieke Universiteit Leuven
2. Center on Aging and Health, Johns Hopkins University
3. Telfer School of Management, University of Ottawa

⋆Contact author: Jo.Reynaerts@econ.kuleuven.be

Keywords: Nonlinear rootfinding, contraction mappings, fixed-point problems, (Quasi-)Newton methods

A number of real-world estimation problems require the solution of large sets of nonlinear equations. For some of these problems, fixed-point iteration schemes have been proposed, such as the Berry *et al.* (1995, BLP) Contraction Mapping algorithm used in estimating random coefficients logit models of demand for differentiated products. This is a fixed-point problem in J products by T markets, where one must invert a demand system to uncover a vector $\boldsymbol{\delta} \in \mathbb{R}^{J \times T}$ that captures the "mean" utility for each product $j = 1, \dots, J$ in each market $t = 1, \dots, T$, and that equates predicted market shares $\hat{s}$ with the actual observed market shares S, as in $\hat{s}(\boldsymbol{\delta}, \boldsymbol{\sigma}) = S$. With $\boldsymbol{\sigma}$ the vector of standard deviations of the distributions of individual tastes for product characteristics, the BLP Contraction Mapping thus involves computing $\boldsymbol{\delta}(\boldsymbol{\sigma}) = \hat{s}^{-1}(S; \boldsymbol{\sigma})$ using the following iterative scheme:

1. For each value of $\boldsymbol{\sigma}$, compute the next value for $\boldsymbol{\delta}$ using
$$\boldsymbol{\delta}^{h+1} = \boldsymbol{\delta}^{h} + \log(S) - \log\left(\hat{s}(\boldsymbol{\delta}^{h}, \boldsymbol{\sigma})\right). \tag{1}$$
2. Stop if $\|\boldsymbol{\delta}^{h+1} - \boldsymbol{\delta}^{h}\| \leq \epsilon$, where $\|\cdot\|$ can be either L_2 or L_∞ and $\epsilon =$ `1e-09` is the tolerance level.

Given its linear rate of convergence and the size of $\boldsymbol{\delta}$ (typically exceeding 1000 observations), this is a time-consuming procedure. By reformulating the fixed-point problem as a nonlinear rootfinding problem, this paper introduces alternative methods to speed up the convergence process, specifically (1) the classical Newton-Raphson (N-R) method, (2) the Broyden secant method (see e.g. Dennis and Schnabel, 1983), and (3) the derivative-free spectral agorithm (DF-SANE). In a Monte Carlo study representing various scenarios, we use the **BB** (Varadhan and Gilbert, 2009) and **nleqslv** (Hasselman, 2009) packages for the implementation in R (R Development Core Team, 2009). We find that DF-SANE is more than three times faster than the BLP Contraction Mapping and achieves convergence in nearly 100% of the simulations runs throughout different scenarios. As suspected for this large system of nonlinear equations, numerical N-R is very slow to converge compared with BLP. The analytical N-R method is substantially faster, with speed of convergence similar to DF-SANE.

A central issue in accelerating iterative algorithms is the trade-off between stability (global convergence) and speed. If the cost of increasing speed is measured in decreased stability, compared to other acceleration schemes the SQUAREM algorithm (Varadhan and Roland, 2008) pays a much smaller price while procuring a decent gain in speed. We are therefore also exploring this globally convergent algorithm, which provides for explicit trade-off between speed and stability, for accelerating the BLP Contraction Mapping.

References

Berry, S., Levinsohn, J. and Pakes, A. (1995), "Automobile Prices in Market Equilibrium," *Econometrica* 63(4), pp. 841–890.

Dennis, J.E. and Schnabel, R.B. (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall Series in Computational Mathematics, Prentice Hall, New Jersey.

Hasselman, B. (2009), *nleqslv: Solve Systems of Nonlinear Equations*, URL `http://CRAN.R-project.org/package=nleqslv`, R package version 1.5.

R Development Core Team (2009), *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.

Varadhan, R. and Gilbert, P. (2009), "BB: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function," *Journal of Statistical Software* 32(4), pp. 1–26, URL `http://www.jstatsoft.org/v32/i04/`.

Varadhan, R. and Roland, C. (2008), "Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm," *Scandinavian Journal of Statistics* 35(2), pp. 335–353.

A Grid Computing Environment for Design and Analysis of Computer Experiments

Yann Richet[1,2,*] , David Ginsbourger[1,3], Olivier Roustant[1,4], Yves Deville[1,5]

1. DICE consortium, "Deep Inside Computer Experiments", France
2. IRSN, Radioprotection and Nuclear Safety Insitute, Fontenay-aux-Roses, France
3. CHYN, Institute of Geology and Hydrogeology, University of Neuchatel, Switzerland
4. ENSM-SE, Graduate School of Engineering, Saint-Etienne, France
5. Statistical consultant, Chambery , France

* Contact author: yann.richet@irsn.fr

Keywords: grid computing, GUI, Rserve, kriging, EGO

Promethee[1] is a flexible and generic grid computing environment designed for numerical engineering. As a key feature, this tool provides a Graphical User Interface (GUI) for the management of parametric computing tasks with heavy software. From the user point of view, Promethee frontend GUI allows to:

1. edit and parametrize input files (of a given computing software),
2. launch remotely and simultaneously all calculations (thanks to Promethee backend grid),
3. parse and store all software output results.

This basic usage of Promethee holds interesting day-to-day benefits for engineering, and above all, turns the computing software usable as a math function, thus suitable for design of computer experiments frameworks. Latest developments of Promethee include the direct integration of algorithms included in R packages, such as **sensitivity**, **DiceDesign**[2], **DiceKriging**[2] and **DiceOptim**[2]. The key point of this R integration is to easily bring cutting-edge algorithms inside Promethee environment, so that any engineer using Promethee as its grid computing frontend may access powerful algorithms without any knowledge in R programming. This fast and easy wrapping of R packages inside one's grid significantly shortens the time-to-market delay for algorithms development projects, and thereby increases industrial interest and support for these R&D services. Obviously a proper use of most of these algorithms relies on users understanding of underlying mathematical background, and could not be dissociated from well suited documentation (mainly provided by packages documentation and vignettes) and training courses (to be embedded in service deliverables).

Technically speaking, Promethee embeds **Rserve**[3] *java* client as a library dependency, and the access to R codes and results is performed through the **Rserve** protocol, thus allowing remote execution (and a possible backend centralization) of R functions.

As an example, the integration of EGO (Efficient Global Optimization[4,5]) algorithm through **DiceKriging** and **DiceOptim** packages inside Promethee grid environment at IRSN is illustrated by a real-world application in nuclear safety assessment based on an industrial Monte-Carlo neutronic simulator[6].

References

[1] Promethee project (2009). "*Unobtrusive grid computing, efficient parametric modeling and deep robust design"*,
http://promethee.irsn.org/.

[2] DICE consortium (2006-2009). "*Deep Inside Computer Experiments"*, http://www.dice-consortium.fr/

[3] Rserve. "*Binary R Server"*, http://www.rforge.net/Rserve/

[4] D. Ginsbourger, R. Le Riche, and L. Carraro (2009), chapter "*Kriging is well-suited to parallelize optimization*", to appear in Computational Intelligence in Expensive Optimization Problems, Studies in Evolutionary Learning and Optimization, Springer.

[5] D.R. Jones (2001), "*A taxonomy of global optimization methods based on response surfaces"*, Journal of Global Optimization, 21, 345-383.

[6] J. Miss, O. Jacquet, F. Bernard, B. Forestier, W. Haeck, Y. Richet, (2008), *"First validation of the new continous energy version of MORET5 Monte Carlo code"*, PHYSOR 2008, Interlaken, Switzerland

spmR: an R package for fMRI data analysis based on the SPM algorithms.

Bjorn Roelstraete[1,*] **& Yves Rosseel**[1]

1. Department of Data Analysis, Ghent University, Belgium
*Contact author: Bjorn.Roelstraete@UGent.be

Keywords: fMRI, Statistical Parametric Map, Dynamic Causal Modeling

Although several R packages for fMRI are available, the spmR package is unique in that it is capable to mimic the results of the widely used SPM package (http://www.fil.ion.ucl.ac.uk/spm). For standard fMRI analyses as well as for studying brain connectivity networks (Dynamic Causal Modeling), the spmR package can be used as a plugin replacement for SPM, yielding exactly the same results. This is important if the Matlab environment is not available (for example in high-performance computing environments), yet SPM comparable results are highly desirable. The R environment is ideal to run large-scale simulation studies. This in contrast to SPM, which is mainly GUI based.

If the fMRI analysis is just a part of a larger pipeline, access to a dedicated R package for fMRI is enormously convenient. spmR is more than just a port of SPM to R. Instead of copying and translating the original Matlab code of SPM, we tried to exploit the R language as much as possible to obtain elegant, clean, and maintainable code. While a large portion of the SPM source code was written to implement statistical routines, we could often use the built-in functionality of R, leading to a huge reduction in code size. At the same time, the names of the functions and the structure of the central SPM.mat file have been largely retained, so that SPM users should have little difficulty using the spmR package instead.

During the presentation, we will demonstrate how spmR can be used for analyzing typical fMRI datasets. Using a single-subject dataset, we show how easy it is to define an experiment, and to run a standard mass-univariate analysis with spmR. Perhaps most important for SPM users, we compare the results with the output of SPM. For example, we compare the estimated F-values in a SPM{F} map for a random selection of voxels. The values computed by spmR are identical to the ones computed by SPM. Next, we illustrate how convenient spmR is when it is used as part of a larger simulation study. We show an example of a real-world simulation script, where a complete pipeline (including data generation, pre-processing, activation detection, and finally estimating a Dynamic Causal Model) is repeated a large number of times.

R for Labour Market Policies

Gloria Ronzoni[1,*], **Ettore Colombo**[1], **Matteo Fontana**[1]

1. CRISP (Interuniversity research centre on public services), University of Milan Bicocca
*Contact author: gloria.ronzoni@crisp-org.it

Keywords: labour market, Talend integration, RMySQL , 3 millions workers

The project "Labour Market Observatory of Lombardia" started in 2005 with a collaboration between CRISP research centre and Regione Lombardia. The Observatory is aimed at gathering, updating and analyzing data and useful information that effectively investigates the efficacy of employment policies, educational system, professional training, further education and the regional labour market trends.

Here we propose an application of labor market monitoring at regional level dealing with longitudinal data: it deals with a classification of Lombardia workers' careers with the aim to identify the trend of contractual profiles (stable, improve or worse condition). The period of interest is since January, 2000 to June, 2009 and 3 millions workers are involved.

To achieve this goal we used a set of 'open source' programs like R, MySQL (Data Storage DBMS) and Talend Open Studio (ETL instrument). The RMySQL package allowed R to communicate with MySQL using the SQL language. The data set used contained 7 millions of rows with a set of 20 quantitative and qualitative variables, so the large amount of data required the implementation of a cyclic R algorithm that allowed to work with events relative to 100.000 workers each time.

Concerning the statistical methodology, the complexity of information didn't permit to use employment status approaches based on panels tipically used in the literature, so our purpose was to use an optimal scaling approach (**smacof** package, MDS approach) that determined, for each contractual typology, a weight of stability based on the working duration time. This quantification allowed to compute a career stability index for each worker and define his career condition.

Time of elaboration (less than one day) was widely reduced throught the joint use of R and MySQL. At the end of the algorithms run the resulting data set contained 3 millions of workers' careers clustered.

In the final part of this work we propose some analysis of the resulting data set and in particular an application of multiple correspondence analysis (**ca** package) that shows how workers' profiles are related to the associated qualitative variables, such as demographic and social information.

References

Lovaglio P.G. (2008). Analisi classificativa longitudinale dei percorsi lavorativi della provincia di Milano, in M. Mezzanzanica e P.G. Lovaglio (Eds), Numeri al lavoro, il sistema statistico del mercato del lavoro: metodologie e modelli di analisi, Quaderni dellosservatorio del mercato del lavoro, 3, Franco Angeli, Milano, pp59-81, ISBN 13: 9788846496195.

Teaching Time Series analysis course using RcmdrPlugin.Econometrics

Dedi Rosadi
Department of Mathematics
Gadjah Mada University, INDONESIA
email: dedirosadi@ugm.ac.id

For time series analysis, there are various packages of R, available under the taskviews Econometrics, Finance and Time Series in CRAN, with the main user interaction via Command Line Interface (CLI). See, for instance, Cryer and Chan (2008), Kleiber and Zeilis (2008), Pfaff (2008), Racine and Hyndman (2002) and Cribari-Neto and Zarkos (1999) for a comprehensive discussion of R application in Time Series and Econometrics modeling. Unfortunately, for teaching purpose, R-CLI seems to be less user friendly and relatively difficult to use, especially if we compare it with the commercial softwares which have an extensive GUI capabilities, such as Eviews. For solving this problems, Hodgess and Vobach (2008) introduced RcmdrPlugin.epack, a R-GUI package for doing time series analysis. In this talk, we introduce a new GUI package for time series analysis, called as RcmdrPlugin.Econometrics (Rosadi et al., 2009). In the current version of the plug-in, the menu contains all models that have been introduced in the course of "Introduction to Forecasting" and "Introduction to Time Series Analysis", offered in our Department. We provide empirical examples showing some unique features of RcmdrPlugin.Econometrics. For the future development, we plan to add more menus for various models that are popular for Econometric society and studied in various econometrics related courses in our Department.

Keywords: R Commander Plug-ins, Open Source, Time Series Analysis

Modified segmentation methods of quasi-stationary time series

Irina Roslyakova

Department of Scale Bridging Thermodynamic and Kinetic Simulation, ICAMS (Interdisciplinary Centre for Advanced Materials Simulation), Ruhr-Universität Bochum, UHW 10/1022, Stiepeler Str. 129, 44801 Bochum

Tel.: +49 234 32 22612, E-mail: irina.roslyakova@rub.de

Abstract

The statistical analysis of quasi-stationary processes, like chemical production processes driving this development, requires a division of the measurements into different stationary time segments. Without such segmentation it is not possible to analyze the influence of parameters and setting to the output. This segmentation problem does not only occur in chemical processing. A completely different application is described in [2001GAL] with the segmentation of human heart rate data sets. The segmentation of quasi-stationary time series is a tedious computational problem. One effective method for time series segmentation was proposed by Pedro Bernaola-Galván et al. [2001GAL] and analyzed later by Kensuke Fakuda et al. [2004FAK]. This method can provide exact segmentation in short computational time, but it is applicable only for data sets following normal distributions. Analysis of segmentation algorithm from [2001GAL] with different sets of data indicates that the proposed method is sensitive and not robust to significant outliers near bounds. In this work, segmentation algorithm, presented in [2001GAL] was modified and improved. Both, original and modified algorithms are compared and implemented with R. Additionally a modified segmentation method was compared with functions breakpoints from R-package *strucchange*. The comparison of these two methods shows that the modified segmentation method can provide better segmentation in significant less computational time as function breakpoints from existing R-package *strucchange.*

Literature

[2001GAL] Bernaola-Galván, Pedro; Ivanov, Plamen Ch.; Amaral, Luís A. Nunes; Stanley, H. Eugene: Scale Invariance in the Nonstationarity of Human Heart Rate. In: Physical review letters (2001), Volume 87, number 16 (abgerufen am 13. August 2009). http://polymer.bu.edu/hes/articles/bias01.pdf

[2004FAK] Fukuda, Kensuke; Stanley, H. Eugene; Amaral, Luı´s A. Nunes: Heuristic segmentation of a nonstationary time series. In: Physical review letters 69 (2004). http://polymer.bu.edu/hes/articles/fsa04.pdf (abgerufen am 13. August 2009)

lavaan: an R package for structural equation modeling and more

Yves Rosseel[1,⋆]

1. Department of Data Analysis, Ghent University, Belgium
⋆Contact author: Yves.Rosseel@UGent.be

Keywords: structural equation modeling, growth curves, latent variables

In the social sciences, latent variables are ubiquitous, and many software packages have been developed that implement multivariate latent variable techniques such as confirmatory factor analysis (CFA), structural equation modeling (SEM) and growth curve modeling. However, perhaps the best state-of-the-art software packages in this field are still closed-source and/or commerical.

The **lavaan** package is developed to provide useRs, researchers and teachers a free, open-source, but commercial-quality package for latent variable modeling. The long-term goal of **lavaan** is to implement all the state-of-the-art capabilities that are currently available in commercial packages, including support for various data types, discrete latent variables (aka mixture models) and multilevel datasets.

During the presentation, I will discuss the design of **lavaan**, its current features, and our plans for the near future. Using several examples, I will illustrate perhaps the most prominent feature of **lavaan**: the formula-based 'model syntax', which is designed to be as compact, elegant, and natural to R users as possible. Finally, a comparison with other related R packages will be made, and I will briefly touch on various ways to interface with other software packages.

Teaching Statistics in Quality Science using the R-Package qualityTools

Thomas Roth, Joachim Herrmann

The Department of Quality Science - Technical University of Berlin
*Contact author: thomas.roth@tu-berlin.de

Keywords: Process Capability, Distribution Fitting, Gage R&R, Design of Experiments, Desirability

A key role in the education of engineers is the teaching of statistical methodology used in Quality Science. Among the many requirements of an engineer is the competence to plan and conduct data collections in a technical environment as well as to analyze the obtained data with respect to quality.

Over the last two years the Department of Quality Science of the TU Berlin successfully held an obligatory course for undergraduates with the title ”Data Analysis and Problem Solving” using the statistical software R and the **qualityTools** package. In the following is a brief illustration of the topics covered in this R-Course, a presentation of the developed R-Package **qualityTools** its concept and methods as well as a resume of the challenge to teach the Six Sigma Quality Management methodology to undergraduates with no statistical background using the **qualityTools** package and the statistical software R in general.

References

Herrmann, Roth(2010). *Qualitaetsmanagement als Pflichtfach fuer Bachelor an der TU-Berlin*, GQW 2010, Conference of the Gesellschaft fuer Qualitaetswissenschaft, (Aachen, Germany), February 2010.

R Development Core Team(2009). *R: A Language and Environment for Statistical Computing.*

John M. Chambers(2008). *Software for Data Analysis.* Springer Verlag.

Uwe Ligges(2007). *Programmieren mit R.* Second Edition. Springer Verlag.

Box, Hunter, Hunter(2005). *Statistics for Experimenters.* Second Edition. Wiley-Interscience.

Montgomery(2005). *Introduction to Statistical Quality Control.* Fifth Edition. Wiley.

Montgomery, Runger(2006). *Applied Statistics and Probability for Engineers.* Wiley.

DIN EN ISO 9000:2005

DIN EN ISO 9001:2008

CXXR and Add-on Packages

Andrew Runnalls[1]

1. School of Computing, University of Kent, UK, A.R.Runnalls@kent.ac.uk

Keywords: R, CXXR, C++, packages, CRAN

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++. 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 design of CXXR endeavours to reconcile three objectives:

- Above all, to be functionally consistent with standard R, both at the R language level, and at the C/Fortran package interface level.
- For the core of the interpreter to be written in idiomatic, standards-conforming C++, making best use of the C++ standard library, and providing a well documented C++ API on which C++ package writers can build.
- To provide a reasonably simple mechanism for CXXR to be upgraded to parallel the continuing evolution of standard R.

Development of CXXR started in May 2007, then shadowing R 2.5.1; at the time of this abstract it reflects the functionality of R 2.10.1. At useR! 2009 Chris Silles described an offshoot project to introduce provenance-tracking facilities into CXXR, 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.

In principle any R add-on package should work without alteration under CXXR, provided it conforms to the `R.h` or `S.h` APIs. (Code using `Rinternals.h` may need alterations, usually minor, as explained in the CXXR documentation.) The primary purpose of this paper—after giving a general update on the CXXR project—is to gauge to what extent this is true in practice, by describing the author's experiences in installing and testing under CXXR a number of the most widely used packages from CRAN. A particular observation will be how CXXR can quickly bring to light memory-protection errors (i.e. incorrect use of `PROTECT()`, `UNPROTECT()` etc.) that may long lie dormant under the standard R interpreter.

The paper will go on to explain how CXXR offers the prospect of making life simpler for package writers incorporating native C/C++ code, and allowing—in a controlled way—closer interaction between package code and the underlying interpreter. For example, the following are already feasible:

- Direct access to the underlying garbage collection system *via* a well-documented and well-encapsulated API.
- In CXXR the `SEXPREC` union is replaced by a C++ class hierarchy. Package writers can extend this class hierarchy as they see fit, rather than needing to use external pointers and finalizers. In particular, new R classes can be wrapped around new C++ classes within the hierarchy.
- Instead of using `PROTECT()` and kindred functions, package writers can use C++ smart pointers which afford memory protection to whatever they point to. This is much simpler and less error-prone than the `PROTECT()` mechanism.

These points will be illustrated by showing how the **ff** package can be reengineered under CXXR. (Admittedly, these facilities come at the expense of compatibility with the standard R interpreter.)

References

Chris Silles and Andrew Runnalls (2009). Provenance Tracking in CXXR, `http://www.agrocampus-ouest.fr/math/useR-2009/slides/Silles+Runnalls.pdf`.

Daniel Adler et al. (2007). ff: memory-efficient storage of large data on disk and fast access functions, `http://cran.r-project.org/web/packages/ff/index.html`.

R as a statistical engine for a water quality trend analysis web-service

Paul Rustomji[1,*], **Brent Henderson**[2], **Katie Mills**[1], **Bai Qifeng**[1], **Peter Fitch**[1]

1. CSIRO Land and Water, GPO Box 1666, Canberra, A.C.T., 2601, AUSTRALIA
2. CSIRO Mathematical and Information Sciences, ANU Campus, Acton, A.C.T., 2601, AUSTRALIA

*Contact author: paul.rustomji@csiro.au

Keywords: Water quality, web service, trend analysis, Sweave

The range of environmental characteristics (climatic, hydrologic etc.) being routinely monitored is growing rapidly with time and increasingly these observations are being made available on-line. One technology that has an important role to play in analysing this growing data volume (not least in the field of hydrology) is web-service technology (1; 3). A web service is an interface that describes a collection of operations that are network-accessible through standardised XML messaging (4). Here, a web service that performs trend analyses of water quality data, using R as the statistical engine for analysis and visualisation, is presented.

Microsoft .NET is used to construct the web service. For each analysis a text file containing user choices regarding method selection etc. and a data file (`*.csv` format, based on data retrieved either from on-line data bases or stored locally), are generated by the client application. R is invoked with an initial `trend_analysis.r` script using the `Rscript.exe` command with the data and parameter files supplied as arguments. This script in turn calls individual sub-routines (depending on parameters specified via the parameter file). Each sub-routine comprises a **Sweave** file which, when `Sweave`'d produces both a LaTeX output file and PDF graphs. The LaTeX output files (`*.tex` format) and graphs are compiled using PDFLaTeX to produce a downloadable PDF report. Current trend analysis methods supported include generalised additive models incorporating a non-linear spline trend term (5), linear models and a non-parametric Seasonal Kendall's Tau slope estimate (2). Finally, a multi-site wrapper routine using GoogleMaps tiles downloaded on the fly (based on **RGoogleMaps**) provides a spatial map of water quality trends.

The advantages of this approach include:

1. Making the statistical power, flexibility and graphing capabilities of R available to a larger audience. Essentially a web-service approach provides access to high level statistical analyses without any requirements for direct R programming capacity on behalf of the user.
2. The ability to reference R objects in the report using the Sweave `Sexpr{}` function.
3. The ability to include interpretative statements in the report tailored to the statistical results (using the LaTeX **ifthenelse** package in conjunction with `Sexpr{}` statements).
4. Harnessing the typesetting capabilities of the TeX engine to produce a high quality PDF report.
5. Accessing the mapping capabilities of GoogleMaps (an extant web-service).
6. Internet-wide accessibility and the ability for this web-service to be called by other users and/or applications (e.g. from Microsoft Excel).

The web service can be accessed at `http://wron.net.au/WebApps/WQSARPortal/Home.aspx` through either a browser based web application or a downloadable Add-In for Microsoft Excel.

References

[1] J. L. Goodall, J. S. Horsburgh, T. L. Whiteaker, D. R. Maidment, and I. Zaslavsky, *A first approach to web services for the National Water Information System*, Environmental Modelling and Software **23** (2008), 404–411.

[2] B. Henderson and R. Morton, *Strategies for the estimation and detection of trends in water quality*, Tech. report, CSIRO Mathematical and Information Sciences Technical Report 07/70, Canberra, 2008.

[3] J. S. Horsburgh, D. G. Tarboton, M. Piasecki, D. R. Maidment, I. Zaslavsky, D. Valentine, and T. Whitenack, *An integrated system for publishing environmental observations data*, Environmental Modelling & Software **24** (2009), 879–888.

[4] H. Kreger, *Web Services Conceptual Architecture (WCSA 1.0)*, Tech. report, IBM Software Group, Somers, New York, 2001.

[5] R. Morton and B. Henderson, *Estimation of non-linear trends in water quality: An improved approach using generalized additive models*, Water Resources Research **44** (2008), 7420.

Column Databases Made Easy with R

Using mmap and indexing for high-performance data management and queries

Jeffrey A. Ryan
insight algorithmics, inc.
jeffrey.ryan@insightalgo.com
Chicago, Illinois USA

Keywords: column database, very large data, searching, queries, indexed searches

As data sets grow ever larger, the difficulty of accessing even subsets of data using R increases. By design, R stores all objects in memory and performs full table searches to extract relevant matches. Without an external database, users are limited to objects that must be a few times smaller than available memory. Often the only practical solution is to use an external database system interfaced with R.

The goal of the **indexing** package is to provide native R search semantics to very large data residing on-disk, organized by column, while reducing memory usage and dramatically speeding up search times. In effect, this alleviates the need to manage a separate external database. **indexing** accomplishes this by:

- Providing advanced index and search functionality to enable very fast and memory efficient boolean searches — including sparse, dense, and bitmap indexes. These tools work on most data objects in R, offering the ability to index vectors, matrix columns, entire data.frames, or atomic components of complex objects with nearly identical semantics. These can be used with resident memory objects, or in combination with disk-based data.

- Using optional memory-mapped files, data resides on disk until needed. Only relevant subsets are loaded into memory when needed for a search or extraction. This enables multi-gigabyte databases to be accessed as easily as if they resided in memory. It also allows for simple binary data files to be used, facilitating cross-application usage.

- Finally, and most critically, the design requires *no additional* R or database language skills, as it makes use of the standard R sub-setting and boolean search semantics.

The talk will examine the design, architecture, and implementation of the **indexing** and related **mmap** packages. Examples from a proprietary database of equity derivative option data covering over one million contracts, across sixty-seven million observations, and nineteen variables will be used to illustrate performance and functionality.

References

Jeffrey A. Ryan (2010). indexing and mmap packages
http://indexing.r-forge.r-project.org.

Trading in Real Time with R and IBrokers

Jeffrey A. Ryan
insight algorithmics, inc.
jeffrey.ryan@insightalgo.com
Chicago, Illinois USA

Keywords: event processing, real-time trading, streaming data, sockets, API-programming

As financial trading strategies become more complex, the integration of software and strategy becomes intertwined. So-called "black-box" trading involves computer-driven strategy and execution in the absence of direct human action. These strategies can be applied in time-scales ranging from end-of-day decision and trade execution, all the way to decisions at the microsecond level.

Most automated trading in the sub-second range requires a fast compiled language as well as high-performance hardware and dedicated connections. Lower frequency trading, such as one second intervals and greater, lends itself to lower performance requirements. R offers facilities to handle socket-based connections on commodity hardware that can easily exceed the more modest performance requirements of this style of trading. This talk will explore the latter approach using the R package **IBrokers** to interface with the popular Interactive Brokers brokerage platform via their proprietary and free API.

The **IBrokers** package allows for event-driven programming using R. The package combines interactive access to historical market data, real-time streaming of sub-second market data on cash, equity, future and option products from multiple markets, as well as the ability to seamlessly manage order execution and account functionality.

This talk will explore the design aspects of recreating an external API within the context of R — from managing socket connections and streaming data, to handling asynchronous path-dependent events using functional closures. The talk will include examples of many of the features that may be of interest to those looking to implement event-processing code in R, as well as a look at specific functionality related to automating a real strategy with a personal or professional account with Interactive Brokers.

Interactive Brokers is a registered trademark of Interactive Brokers LLC. **IBrokers***, insight algorithmics inc. and the author are in no way affiliated or endorsed by Interactive Brokers.*

References

Jeffrey A. Ryan (2010). IBrokers package
http://cran.r-project.org/web/packages/IBrokers/index.html.

Drug Supply Modeling Software

Sourish Saha, Vladimir Anisimov, Valerii Fedorov, Richard Heiberger

The design of multicentre clinical studies consists of several interconnected stages including patient recruitment prediction, choosing a randomization scheme and a statistical model for analyzing patient responses, and drug supply planning. The Research Statistics Unit (RSU) at GlaxoSmithKline (GSK) has developed a supply modeling tool to predict drug supply needed to cover patient's demand in a single study with a given risk of running out of stock for a patient. The tool allows for central and centre-stratified randomization of the patients, equal and different treatment proportions within the randomization block, single and multiple dispense types of studies, and other factors. All algorithms are based on closed-form analytic expressions so no Monte Carlo simulation is necessary. The primary tool is built as an R package. In order to support Clinical Trials Supply and Global Supplies Operations teams at GSK, the RSU created a user-friendly RExcel interface embedding the risk-based supply modeling tool into the Excel environment.

Generalized Significance in Scale Space: The GS3 Package

Daniel V. Samarov

In traditional approaches to multivariate nonparametric regression the focus is on the estimation of a single optimal bandwidth matrix. An alternative approach which has gained increased attention in the statistics community is based on scale-space theory from computer vision, which takes into consideration a family of smooths. The intuition behind the latter approach is that different levels of smoothness reveal different, potentially significant features about the data. While in the multivariate case much work has been done in determining feature significance (such as local extrema), little work has been done on how to select a sensible family of smooths. In this talk we present the GS3 R package which provides an implementation of a novel approach for finding such a family in the multivariate nonparametric regression setting based on the rodeo algorithm. This methodology, which we will refer to as Generalized Significance in Scale Space (or GS3) is motivated by and applied to the estimation of aerosol extinction in the atmosphere and hyperspectral medical imaging.

References

Chaudhuri, P., Marron, J.S., Scale-space view of curve estimation, Annals of Statistics 28 (2000) 408-428.

Duong, T., Cowling, A., Koch, I., Wand, M.P., Feature significance for multivariate kernel density estimation, Computational Statistics and Data Analysis 52 (2008) 4225-4242.

Godtliebsen, F., Marron, J.S., Chaudhuri, P., Statistical Significance of features in digital images, Image and Vision Computing 22 (2004) 1093-1104.

Forecast Monitoring via Multivariate Statistical Process Control with R

Robert Wayne Samohyl [1], Elisa Henning [2]

1. Federal University of Santa Catarina
2. State University of Santa Catarina
* Contact author: samohyl@deps.ufsc.br

Keywords: Control Charts, MCUSUM, forecast monitoring, multivariate

In order to identify and analyze the major causes of variability that affect the accuracy of the forecast object simultaneously monitoring two or more forecasts depends on the development of specific statistical tools including graphics that support monitoring in real time. Procedures for calculating forecasts may depend upon purely subjective insight or upon statistical methods like Box-Jenkins or other computational methods such as neural networks and exponential smoothing models (R package **forecast**). In other words, the forecast model may be based upon either subjective or scientific principles. An important question involves the measurement of forecast accuracy and the determination of the relevance of forecasting model. The researcher must determine when a forecast error is disturbingly large, and how should the model be corrected to improve forecast accuracy. Statistical process control offers a series of monitoring tools for the analysis of quality characteristics, in this case the accuracy of forecasts. Multivariate control charts represent one of these emerging statistical techniques already successfully used to monitor simultaneously several correlated characteristics relevant to the production process. The use of graphics in the industrial environment has increased in recent years due to many resources of information technology now available to reduce the complexity of modern industrial processes and as argued in this paper including the forecasting process. This article presents some computational routines developed in the GNU R package for the application of statistical control for multivariate processes of various simultaneous forecasts based on the cumulative sum (MCUSUM) control chart. The routines were developed in the R programing language in order to facilitate information entry to produce a clear graphics interface and to return the maximum amount of information needed for forecast monitoring. The routines were applied successfully to artificially simulated data and to real life examples. We can conclude that the R environment is an important alternative for the diagnosis and monitoring of multivariate forecasts.

References

Hawkins, D.M. and Olwell, D.H. (1998) *Cumulative Sum Charts and Charting for Quality Improvement,* Engineering and Physical Science, Springer**.**

Henning, E., Alves, C. C; Samohyl, R. W (2008). The development of graphics and control MCUSUM environment MEWMA in R as an alternative procedure for statistical analysis of multivariate processes. ENEGEP 2008 (Rio de Janeiro, Brazil), pp. 1-10.

Hyndman, R.J. and Khandakar, Y. (2008) Automatic time series forecasting: The forecast package for R, Journal *of Statistical Software,* 26, (3).

Samohyl, R.W., Souza, G.P. e Miranda, R.G. de, (2008) *Métodos Simplificados de Previsão Empresarial***.** Rio de Janeiro: Ciência Moderna,.

Bivariate Analyses

Héctor Sanz[1,2*], **Isaac Subirana**[3,1,4], **Joan Vila**[1,3]

1. Institut Municipal d'Investigació Mèdica (IMIM), Barcelona, Spain
2. Institut d'Investigació Biomèdica Girona (IDIBGI), Girona, Spain
3. CIBER Epidemilogía y Salud Pública
4. Statistics Department, University of Barcelona, Barcelona, Spain

*Contact author: hsanz@imim.es

Keywords: Software Design, Bivariate Table, LaTeX, Descriptive Analysis

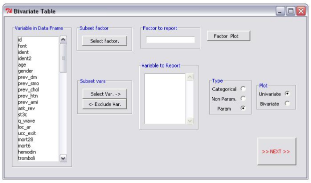
In many studies, especially epidemiological studies, it is important to compare characteristics between independent groups of individuals. Usually these comparisons are presented in the form of tables of descriptive statistics where rows are characteristics, and each column is a group. Tables of this form are usually called bivariate. For example, the comparison between treated and untreated patients (column variable), in terms of age, history of hypertension, triglyceride levels, etc (row variables). Usually the number of characteristics is large, and thus construction of the table is laborious. And if, as often happens, the results must be presented stratified by sex, the process is even more laborious.

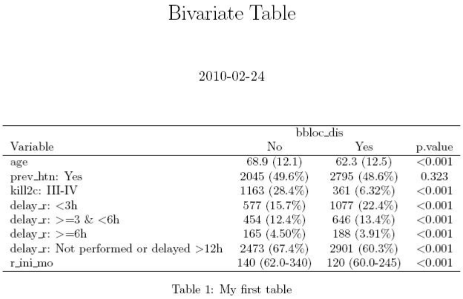
For these reasons, we have implemented a function which quickly and efficiently generates bivariate tables in plain text, or as LaTeX instructions. Depending on the nature of the row variable, different statistics can be calculated:

- for numeric, normally distributed variables: mean and standard deviation are calculated, and a Students t-test/ANOVA test is performed.
- for non-normally distributed numeric variables: median, and first and third quartiles are calculated, and a Mann-Whitney U or Kruskal-Wallis test is performed.
- for categorical variables, frequencies are presented, and a Chi-square or Fishers Exact test is performed, as appropriate.

In order to facilitate construction of the bivariate table, we have implemented a graphical interface which allows the user to easily modify a variety of options: for example, specify how many decimals, select individuals to include and variables to analyse, choose whether to display absolute or relative frequencies, etc. All this functionality is also available using R syntax.

To illustrate its way of working, our function uses a data set resulting from a cross-sectional study[1].




Bivariate Table

2010-02-24

	bbloc_dis		
Variable	No	Yes	p.value
age	68.9 (12.1)	62.3 (12.5)	<0.001
prev_htn: Yes	2045 (49.6%)	2795 (48.6%)	0.323
kill2c: III-IV	1163 (28.4%)	361 (6.32%)	<0.001
delay_r: <3h	577 (15.7%)	1077 (22.4%)	<0.001
delay_r: >=3 & <6h	454 (12.4%)	646 (13.4%)	<0.001
delay_r: >=6h	165 (4.50%)	188 (3.91%)	<0.001
delay_r: Not performed or delayed >12h	2473 (67.4%)	2901 (60.3%)	<0.001
r_ini_mo	140 (62.0-340)	120 (60.0-245)	<0.001

Table 1: My first table

References

1. Tomas M, Vazquez E, Fernandez-Fernandez JM, Subirana I, Plata C, Heras M, Vila J, Marrugat J, Valverde MA, Senti M. (2008). Genetic variation in the KCNMA1 potassium channel alpha subunit as risk factor for severe essential hypertension and myocardial infarction. *J Hypertens*, 26(11):2147-53.

Adaptive Middleware and High Performance Software For Multi-core Deployments Across Cloud Configurations

Doug Schmidt[1*]

1. Ziron Computing LLC
*Contact author: doug.schimdt@zircomp.com

Keywords: High-performance computing, Commercial applications of R, Cloud computing, Financial applications

Statisticians, analysts, scientists, and engineers require massive processing power to conduct data analysis, predictive modeling, visualization, and other complex tasks. Although these groups could use specialized super computers, the custom development time and the hardware costs are prohibitive. This paper describes how Zircon applied the Zircon adaptive ultra high-performance computing software platform and tools with the R programming language and environment for cloud enablement to substantially improve the performance of a representative complex computational finance application via distribution and parallelization, thereby reducing the total computation time from 3093.1 minutes to 40.8 minutes (*) on an off-the-shelf commodity multi-processing platform.

pR: Enabling Automatic Parallelization of Data-Parallel Tasks and Interfacing Parallel Computing Libraries in R with Application to Fusion Reaction Simulations[†]

Neil Shah[1, 2], **Guruprasad Kora**[2], **Paul Breimyer**[1, 2], **Yekaterina Shpanskaya**[1, 2], **Nagiza F. Samatova**[1, 2, *]

[1]North Carolina State University; [2]Oak Ridge National Laboratory; *Contact author: samatovan@ornl.gov

The ever-growing size and complexity of modern scientific data sets constantly challenge the capabilities of existing statistical computing methods. High-Performance Statistical Parallel Computing is a promising strategy to address these challenges, especially with the advent of powerful multi-core computing architectures. However, parallel statistical computing techniques introduce many implementation complexities, resulting in a need for more efficient and streamlined processes. In response to this need, we introduce **pR**, a lightweight, easy-to-use middleware for the statistically-rich *R* engine that further enables parallelization in statistical computing. **pR** branches into two main approaches that have strong potential to simplify and improve parallel-computing capabilities: 1) interfacing existing third-party parallel codes with the flexible scripting statistical environment of *R* and 2) supporting the automatic parallelization of data-parallel tasks in hybrid multi-node and multi-core environments.

R's inherent extensibility and flexibility make it an ideal platform for statistical computing. However, *R* has limited native support for parallel computing. To enable parallelization, researchers have previously produced add-on packages such as **Rmpi** and **rpvm** to provide a low-level base for writing parallel codes. Other packages, such as **snow** use these packages as a foundation for handling embarrassingly-parallel statistical computations. However, these packages burden the end-user with the responsibilities of implementing such parallel codes. Additionally, they can also result in slower execution times as a result of the interpreted nature of *R*. By bridging the *R* environment with existing parallel-computing libraries, **pR** allows developers to leverage existing parallel codes written in compiled languages without modifying the package itself. Additionally, **pR** shifts the responsibility of handling parallel programming details away from end-users [1].

Although bridging parallel computing libraries to *R* is a promising approach, an ideal system would automatically execute researchers' serial codes in parallel. However, the Holy Grail of statistical computing is elusive. A simpler, yet powerful approach involves the automatic execution of a single task on multiple sets of data in parallel, thereby avoiding inter-process dependency issues. *R* supports a family of *apply* methods that serially execute a given function to each element in a collection, making it an ideal candidate for automatic parallelization. For example, the *lapply* method in *R* accepts a list and executes a function against each element in that list. Previous parallel implementations of *lapply* have utilized approaches involving assigning the elements of the list to different processes, computing, and gathering results. Current projects implementing the parallel *lapply* function include **snow** and **multicore**, for multi-node and multi-core environments, respectively. However, neither provides support for hybrid multi-node, multi-core environments. To assuage this issue, we extend **pR** to support automatic parallelization of statistical computations in such settings. **pR** implicitly migrates the R environment to each node and distributes data equally among the nodes. The work per node is further divided among cores and, thereby striving for a 'best of both worlds' approach. Using *R*'s *lapply* method, we demonstrate **pR**'s benefits, particularly in improved overall performance and transparent parallelization [2].

In application, **pR** can serve to expedite the end-to-end pipelines of knowledge-discovery workflows and process extreme-scale data in realistic time using hybrid multi-node, multi-core architectures. One tested application of **pR** is the discovery and analysis of turbulent-fronts in simulation data produced by the XGC particle-in-cell gyrokinetic fusion simulation code. In practice, the discovery of fronts in fusion simulations could provide insight to engineering a solution for viable fusion-energy production.

References

1) P. Breimyer, W. Hendrix, G. Kora, N.F. Samatova, 2009. pR: Lightweight, Easy-to-Use Middleware to Plugin Parallel Analytical Computing with R. In *The International Conference on Information and Knowledge Engineering (IKE)*.
2) P. Breimyer, G. Kora, W. Hendrix, N. Shah, N.F. Samatova, 2009. pR: Automatic Parallelization of Data-Parallel Statistical Computing Codes for *R* in Hybrid Multi-node and Multi-core Environments. In *International Association for Development of the Information Society Applied Computing Conference* (*IADIS ACC*).

[†] We would like to acknowledge Prof. CS Chang for suggesting the application problem. This work was funded by the Scientific Data Management Center under the Department of Energy's SciDAC program. Oak Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. D.O.E. under contract no. DEAC05-00OR22725.

Visualization of titrated dose and recurring events using R/ggplot2

Yue Shentu

In a clinical trial with titratable dosing scheme, it is often crucial to understand the relationship between recurring Adverse events, longitudinal efficacy profile and the associated titration path of the medication. visualization of individual data as well as population summary is an effective tool to identify underlying trend and correlations. Using the ggplot2 package and R base graphic, we generated series of graphs that provided important insight to the characteristics of the experimental drug, and helped clinicians to monitor patient safety in on-going studies. This presentation is a summary of what we have done for a real-life clinical study, with graphics generated using mocked data.

DIAGNOSTICS IN COUNT DATA MODELS

Sibnarayan Guria; *West Bengal State University, Kolkata, India.*

Abstract

There are primarily two reasons for the paucity of diagnostic studies beyond the linear model. First, most diagnostic techniques exploit the nature of the relationship between the response and the explanatory variables thus making it easier to deal with the linear form of relationship. Second reason is the closed form solution that the normal equations of a linear model provides. Non-linearity requires some form of iteration and hence makes the diagnostic measures difficult to derive algebraically.

Unlike the least-squares estimation method used in the classical linear model, the generalized linear models use the maximum likelihood method of parameter estimation which, as Pregiborn (1981) points out, is extremely sensitive to outlying responses and extreme points in the design space. However, since the likelihood equations are not in a closed form and some iterative techniques are required to obtain the estimates it is difficult to derive diagnostics to identify such observations.

In this paper a log-linear linear model is considered and show that the deletion technique can be used to study the diagnostics of such a model. As is well-known in the literature, the convergence of the likelihood solution is extremely rapid for such models and hence we base our study on a one-step approximation of the likelihood estimates. The model is fitted using the maximum likelihood method and the deletion of observation technique is used to identify outliers. Expressions for the change in the estimates of the parameters after re-fitting are obtained.

In this paper all the computations are done in R, Using simplicity and beautiful nature of Java programming language, input data files and the output are manipulated and represented in a lucid and meaningful way. Using the Java port to R - “Java2R“ library (http://www.sngforge.co.nr/projects/java2r) any R-function can be accessed from Java.

Tools : Java2R

Key Words : DFBETA, DFFIT, Log-Linear model.

Automating biostatistics workflows for bench scientists using R-based web-tools

Jeff Skinner, Vivek Gopalan, Jason Barnett, Yentram Huyen*

Bioinformatics and Computational Biosciences Branch (BCBB)
Office of Cyber Infrastructure and Computational Biology (OCICB)
National Institute of Allergy and Infectious Diseases (NIAID)
National Institutes of Health (NIH)
*Contact author: huyeny@niaid.nih.gov

Keywords: Web-tools, Computational Biology, Curve-fitting, Structural Biology, NIH

Biological data can be complex, so bench scientists often develop complicated workflows to process, analyze and present their data. Often biological data will be output from a laboratory instrument as a text data file, then a researcher will spend hours processing the data in MS Excel®, performing analyses in a commercial statistics software package or a custom built script written in FORTRAN or perl, before processing the results further back in Excel and finally producing a report in MS Word® or PowerPoint®. These complicated workflows are tedious and time consuming; they introduce multiple opportunities for error and they can be difficult for future researchers to reproduce. It may be easy to replicate these workflows in R, but its steep learning curve prevents many bench scientists from using R scripts that might simplify their analyses. Statisticians and R programmers need to provide more intuitive user interfaces before their R scripts can be widely adopted by biologists. We present results from two web-based tools, which use R to reproduce critical analysis workflows while providing bench scientists with a simple point-and-click GUI front-end. The Dose-Response Analysis Pipeline (DRAP) allows immunologists to apply curve-fitting analyses to multiple dose-response experiments conducted on one or more 96-well plates. Logistic curve-fit results can be compared among several groups or factors distributed within or among the 96-well plates. Final results are presented in an interactive PDF report with high-resolution images. The DRAP workflow was able to analyze approximately 2000 plates in 30 minutes, which had taken more than 200 hours to analyze by hand (`http://exon.niaid.nih.gov/DRAP`). The Hydrogen Exchange with Normalized Assessment of Maximum Entropy (HDXNAME) workflow allows structural biologists to compute protein flexibility estimates called protection factors from hydrogen exchange experiment data using Maximum Entropy Methods (MEM). This single workflow replaces several cut-and-paste procedures between hard-to-find Excel templates and a custom software application written in MS BASIC. The final result includes a statistical summary comparing two conformational states of the protein and an image of the protein with protection factors mapped on the protein surface (`http://exon.niaid.nih.gov/HDX_NAME`). These two examples show how R programming and web-site design can be used to create meaningful custom applications that will be widely used and shared among biology researchers.

References

JM Sa, O Twu, K Hayton, S Reyes, MP Fay, P Ringwald and TE Wellems (2009). Geographical patterns of *Plasmodium falciparum* drug resistance distinguished by differential responses to amodiaquine and chloroquine. *PNAS*, 106(45): 18883–18889.

L Kong, C Huang, SJ Coales, KS Molnar, J Skinner, Y Hamuro and PD Kwong (2010). Local conformational stability of HIV-1 gp120 in unliganded and CD4-bound states as defined by amide hydrogen/deuterium exchange. *in preparation*

Evolving R for Use in Commercial Environments

David Smith[1,*]

1. REvolution Computing
* Contact author: david@revolution-computing.com

Keywords: Commercial applications of R, REvolution Computing

Use of R has grown dramatically over the past decade for statistical analysis in academia and research, but has yet to make significant inroads in the commercial world – at least compared to incumbent, proprietary tools. Over the past year, I've been documenting (at the *Revolutions* blog) examples where R is used for commercial applications, usually by technologically-advanced companies benefiting from R's flexibility and access to the latest statistical tools from CRAN. But what does the more mainstream commercial user need from their R environment to make it a compelling alternative to the proprietary tools?

As a company that sells extensions to the R environment to businesses, REvolution Computing has performed extensive research into the requirements of data analysts, statisticians and predictive modelers in commercial environments. Some of these requirements have been addressed in features of REvolution R Enterprise released to date, but there's much more to come. In this talk, I'll share some results of our research and describe how we intend to further evolve our extensions to R – particularly in the areas of scalability and usability – to make REvolution R Enterprise a mainstream alternative to the likes of SAS and SPSS.

References

Smith, David (ed.) (2010). *Revolutions blog: News about R, statistics and the world of open source from the staff of REvolution Computing*
http://blog.revolution-computing.com/.

REvolution Computing (2010). *REvolution R Enterprise*
http://www.revolution-computing.com/products/revolution-enterprise.php

Plotting Advanced Mathematical Functions in Excel using RExcel

Christopher Snyder, Keith Halbert

Keywords: Microsoft Excel, R through Excel, RExcel, plot function

Students first learning about functions benefit immensely from being able to visualize them on a graph. ExcelRGraph is an Excel workbook that uses the power of RExcel to give the user the ability to interactively plot functions via R and view them within the familiar Excel interface. Whereas the plotting window of many graphing calculators is small and has unlabled axes, ExcelRGraph is able to take full advantage of all the advanced graphing features available in R. Advanced features include: full-color plotting, user-specified fully-labeled axes, plots of derivatives and integrals of functions, the ability to view multiple functions on the same plot window, and much more.

References

[1] Erich Neuwirth, with contributions by Richard Heiberger, Christian Ritter, Jan Karel Pieterse, , and Jurgen Volkering, *Rexcelinstaller: Integration of r and excel, (use r in excel, read/write xls files)*, 2009, R package version 3.0-18.

SHOGUN - A Large Scale Machine Learning Toolbox

Sören Sonnenburg[1,2,*], Gunnar Rätsch[2], Sebastian Henschel[2], Christian Widmer[2], Jonas Behr[2], Alexander Zien[2,5], Fabio de Bona[2], Christian Gehl[3], Alexander Binder[3], Vojtech Franc[4]

1. Machine Learning Group, TU-Berlin, Franklinstr. 28/29, 10587 Berlin, Germany
2. Friedrich Miescher Laboratory, Spemannstr. 35, 72076 Tübingen, Germany
3. Fraunhofer Institut FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany
4. Center for Machine Perception, Technicka 2, 166 27 Praha 6, Czech Republic
5. LIFE Biosystems GmbH, Poststr. 34, 69115 Heidelberg, Germany

*Contact author: Soeren.Sonnenburg@tu-berlin.de

Keywords: Machine Learning, Large Scale, Support Vector Machines, Kernels, Open-Source

We have developed R-bindings for our machine learning toolbox SHOGUN, which features algorithms for hidden markov models, regression and classification problems. SHOGUN's focus is on Support Vector Machines, but also implements a number of linear methods like Linear Discriminant Analysis, Linear Programming Machines and Perceptrons. It provides a generic SVM interface enabling the choice between *fifteen* different SVM optimizers as back-ends, among them the state of the art LibSVM[1] and SVM^{light}[4], SVMOcas [3] or Liblinear [2]. The SVMs can be easily combined with more than 35 different kernel functions (see http://www.shogun-toolbox.org/doc). Moreover, it offers options for using precomputed kernels and easily allows the integration of custom kernels. One of SHOGUN's key features is the *combined kernel* to construct weighted linear combinations of multiple kernels that may even be defined on different input domains and learned using Multiple Kernel Learning (MKL) algorithms, e.g., [5, 7]. Input feature objects can be dense or sparse vectors of strings, integers (8, 16, 32 or 64 bit; signed or unsigned), or floating point numbers (32 or 64 bit), and can be converted into different feature types. Chains of "pre-processors" (e.g., subtracting the mean) can be attached to each feature object allowing on-the-fly pre-processing. Finally, several commonly used performance measures for evaluation (e.g., area under ROC) are implemented.

A central aspect in the design of SHOGUN was to enable large-scale learning. We implemented auxiliary routines that allow faster computation of combinations of kernel elements that lead to significant speedups during training and evaluation [6] enabling us to solve several large-scale learning problems in biological sequence analysis [3, 6, 8] involving millions of sequences. Furthermore, linear SVMs can be efficiently trained by computing feature spaces on-the-fly, even allowing to mix sparse, dense and other data types.

All of SHOGUN's core functions are encapsulated in a library (libshogun) and are easily accessible and extendible by C++ application developers. Built around SHOGUN's core we provide two types of R interfaces: A modular interface created with SWIG[1], and a static interface. However, note that interfaces are available from C++ to other languages, such as *Python*, *Octave* and $\textit{Matlab}^{TM}$. SHOGUN's source code is freely available under the GNU General Public License at http://www.shogun-toolbox.org.

References

[1] C.-C. Chang and C.-J. Lin, *Libsvm: Introduction and benchmarks*, Tech. report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 2000.

[2] R. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, *LIBLINEAR: A library for large linear classification*, Journal of Machine Learning Research **9** (2008), 1871–1874.

[3] V. Franc and S. Sonnenburg, *Optimized cutting plane algorithm for large-scale risk minimization*, Journal of Machine Learning Research (2009).

[4] T. Joachims, *Making large–scale SVM learning practical*, Advances in Kernel Methods — Support Vector Learning (Cambridge, MA) (B. Schölkopf, C.J.C. Burges, and A.J. Smola, eds.), MIT Press, 1999, pp. 169–184.

[5] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien, *Efficient and accurate l_p-norm multiple kernel learning*, Advances in Neural Information Processing Systems 21, MIT Press, Cambridge, MA, 2010.

[6] S. Sonnenburg, G. Rätsch, and K. Rieck, *Large-scale learning with string kernels*, Large-Scale Kernel Machines (L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, eds.), MIT Press, 2007, pp. 73–103.

[7] S. Sonnenburg, G. Rätsch, S. Schäfer, and B. Schölkopf, *Large scale multiple kernel learning*, Journal of Machine Learning Research (2006), accepted.

[8] S. Sonnenburg, A. Zien, and G. Rätsch, *ARTS: Accurate recognition of transcription starts in human*, Bioinformatics **22** (2006), no. 14, e472–480.

[1] The Simplified Wrapper and Interface Generator, cf. http://www.swig.org.

A high-performance compiler for a subset of R

Ansgar Steland[⋆]

RWTH Aachen University
[⋆]Contact author: steland@stochastik.rwth-aachen.de

Keywords: Concurrency, Compiler, Parallel Computing, Statistical Computing, Simulation, Threads

As the de-facto standard for statistical computing and programming, R is heavily used for computationally demanding tasks ranging from the analysis of huge and high-dimensional data sets arising in genetics, econometrics or environmetrics to large-scale stochastic simulation. Despite the progress made, the performance of R can be rather poor when it comes to statistical computations with numbers, i.e. *number crunching*, which is a consequence of the general concept behind R. This particularly applies when computations involve new algorithms which do not reduce to some calls to build-in functions. Another issue is that R has no simple concept for parallel computing which would allow us to exploit the present day multi-core computers effectively.

The common approach to overcome this drawback is to write the time-critical parts (functions) in C and to invoce them from R. This approach often provide optimal speedups, but it has several drawbacks. First, it is no solution for R users not trained in a language such as C. Second, developing C implementations of complex algorithms arising in present day statistical research is often difficult, time consuming, thus expensive, and also error-prone. Third, one has to waive all the elegant and powerful concepts R offers, particularly matrix calculus and lists, or map them to rather cryptic C function calls and pointer arithmetics. The project P allows for an intermediate way where crucial language features are available and the code runs substantially faster, since it is compiled. Indeed, the compiler results in tremendous speedups in many cases.

The compiler implements a dialect of a subset of the S/R language enlarged by various extensions yielding a language called P. The available subset covers the most important data types for mathematical computing (double precision reals, vectors, matrices and lists) and basic operations for them. Loops, if-then-else, while, do-while and functions (with recursion) are available as well. P runs on common 32bit as well as 64bit platforms including OS X, Linux and Windows, without the need to have installed additional programs or libraries. Since the compiler can be embedded in R, source code of P and R can be mixed in one source file. Alternatively, a portable binary file can be generated which is then executed by the runtime system. Although P differs to some extent from R, porting R functions dealing with mathematical computations usually requires only a few trivial changes. Further, porting C code is supported by a `cstyle` environment which, among other issues, changes the indexation of vectors and matrices from $1..n$ to $0 \ldots n-1$.

P also implements various extensions of the R/S language, e.g. pointers allowing for call by reference, which can speed up programs substantially and is important when handling large data structures. More importantly, we implemented a powerful and easy to use approach for parallel computing. Indeed, many computations in statistics can easily be parallelized. P has a `thread` statement which allows us to create concurrent shared memory threads. Local threads initiated by a function can share local variables or can use them as private ones, e.g. as loop variables, and can even recursively call the function which created them. Hence the powerful divide-and-conquer approach for parallelization is at our disposal, which can greatly simplify the parallelization of existing programs. Additional extensions (e.g. `inf`) simplify coding certain statistical algorithms and mathematical formulae as arising, e.g., in sequential analysis.

The project is work in progress of the author (during his spare time), but is now rather settled and no longer what is called a *proof-of-concept* in computer science.

References

Steland, A (2010). P: A compiled language, *under preparation.*

tsX: An R package for the exploratory analysis of a large collection of time-series

G. Subramaniam, [1] R. Varadhan,[2] S. Urbanek, [1] and S. Epstein [1]

[1] AT&T Labs - Research, [2] Division of Geriatric Medicine and Gerontology,
School of Medicine, Johns Hopkins University,

1 Abstract

We discuss **tsX**, an R package for exploring a large collection of time series. This was motivated by a time series data mining problem in telecommunication network data, where given a large number of time series the objective was to identify time series that have potentially "interesting behavior" (Subramaniam and Varadhan 2007 and 2008). Smooth representation of each time-series (which can be unequally spaced) is first obtained using automatic smoothing parameter selection procedures. Various features of the time-series are then derived based on the smoothed functions. Some of the useful features include mean value, scale, first and second derivatives, critical points, wiggliness, signal/noise ration, and potential outliers. A key feature of this package is that it provides a choice of different smoothing techniques and automatic smoothing parameter estimation procedures. A comprehensive simulation study of these smoothing techniques was performed to evaluate the performance of the smoothing techniques in terms of their ability to estimate the smooth underlying function and the first and second derivatives (Varadhan and Subramaniam 2009). This provided validation of the smoothing techniques available in **tsX** for exploratory use in time series data mining setting. **tsX** provides useful visualization techniques that provide the capability to easily identify and collect curves exhibiting interesting or anomalous behavior using interactive graphics.

References

Subramaniam. G and Varadhan. R (2007). *Feature Extraction Using FDA*. JSM 2007, (Salt Lake City, UT,USA), Aug. 2007

Subramaniam. G and Varadhan. R (2008). *Borrowing Strength in Time Series Data Mining.* JSM 2008, (Denver, CO,USA), Aug. 2008

Varadhan. R and Subramaniam. G (2009). *Automatic Numerical Differentiation of Noisy, Time-Ordered Data in R.* UseR!2009, (Renne, France), Jul. 2009

Criss-Crossing the Org Chart: Predicting Colleague Interactions with R

Eric Sun[1,*]

1. Facebook, 1601 S California Ave. Palo Alto, CA 94304
*Contact author: esun@facebook.com

Keywords: Machine Learning, Organizational Behavior

In this talk, we present a novel application of R for machine learning in the workplace: automatically predicting future levels of interaction between co-workers.

In mid-to-large sized organizations such as Facebook, employees typically work on projects with colleagues across many teams. In a fast-paced work environment, it quickly becomes difficult to remember the colleagues an employee has worked with and to identify the co-workers with whom she is most likely to interact with in the future. Such a system would have many useful applications:

- Suggesting peer reviewers during performance review season
- Optimizing seating charts for maximum productivity
- Setting up optimally-constructed teams within a company
- Automatically filtering internal feeds of employee content (such as commit logs) to deliver personalized content to each employee
- Suggesting new colleague interactions (based on second-degree connections) that may be useful to one's work
- Giving managers more insight into their employees' interactions

To accomplish this task, we generate a dataset where each row consists of a pair of employees. For each pair we calculate many features in several different categories: direct communication (such as the number of code reviews requested from one member of the pair to the other), implicit interaction (such as the number of meetings co-attended), and implicit communication (such as the number of common mailing list threads). As controls, we also include dummy variables for manager, direct report, and peer relationships, and also control for physical proximity (from seating charts). For each event, we weight each interaction by 1 / (number of people involved). Thus, a meeting with 5 people would count less than a one-on-one meeting.

Using these features, we create a model using the **randomForest** package in R that predicts the total number of weighted interactions between a pair of employees in the next 28 days using data from the previous 28 days. New predictions are generated automatically every night and are displayed in a dashboard for each employee to view. With the current model, mean square error on a held-out test set is 0.1049, and anecdotally, most employees have reported that predictions from the system are very accurate.

While the model is set up to predict future interactions, it is also interesting to examine the coefficients of the features to find out which of the independent variables leads to long-term, persistent interactions.

In addition to presenting results from the random forest model, we compare the performance and accuracy of various other algorithms including boosted trees (using the **gbm** package) and ordinary linear regression.

ChinaMap: Maps of China for analysing spatial data

Qiushan Tao[1*]

[1] Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center
*Contact author: qtao@bjmu.edu.cn

Keywords: map, spatial data, spatial statistics, data mining

Spatial data, also known as geospatial data or geographic information, is the data or information that identifies the geographic location of features and boundaries on Earth. Spatial data exists in many scientific fields and applications. Spatial statistics was known as one of powerful tools of data mining for all scientific fields related with spatial data. Due to vast improvements in spatial statistics, R has had an increasing number of contributed packages for handling and analyzing spatial data. There were a few R packages, **mapdata** for example, did have a map of China with provincial boundaries. However, it demands detailed China map for spatial analysis applications. This study aimed to build a detailed China map, a novel R packages named as **ChinaMap**, based on the public map databases in China national fundamental geographic information system. This study use a R packages named **MapTools** to read China map shape files and build map objects according to different geographical regions and points. The map scale was 1:4000000 for public usages. A few novel functions was built for better map visualization in **ChinaMap** packages.

References

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`.

Richard A. Becker, and Allan R. Wilks, "Maps in S", *AT&T Bell Laboratories Statistics Research Report [93.2], 1993.*

Richard A. Becker, and Allan R. Wilks, "Constructing a Geographical Database", *AT&T Bell Laboratories Statistics Research Report [95.2], 1995.*

Ray Brownrigg (2010). mapdata: Extra Map Databases, `http://cran.r-project.org/web/views/Spatial.html`.

Roger Bivand (2010). CRAN Task View: Analysis of Spatial Data, `http://cran.r-project.org/web/packages/mapdata/index.html`.

Nicholas J. Lewin-Koh and Roger Bivand, et al. (2010). maptools: Tools for reading and handling spatial objects. `http://cran.r-project.org/web/packages/maptools/index.html`.

China national fundamental geographic information system. `http://ngcc.sbsm.gov.cn/`.

RGtk2Extras and DanteR: rapid GUI development for an "omics" R package

Tom Taverner[1], Ashoka Polpitiya[2], Gordon A Anderson1[1], Richard D Smith[1]

1. Biological Sciences Division, K8-98, Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352
2. Translational Genomics Research Institute, Phoenix, AZ 85004

Abstract: We describe two recently developed R packages. RGtk2Extras is based on the RGtk2 package and gWidgets and allows rapid development of GUI interfaces, dialogs and tables for existing R functions. Building on the tools developed in RGtk2Extras, the DanteR package gives researchers the ability to perform common tasks on data sets in bioinformatics, such as file import/export, rolling up peptide data into protein data, performing statistical tests, and producing graphs and plots. A number of algorithms for data mining, statistics and plotting are available.

Poster Session: Analytics for Trading Financial Spreads

Paul Teetor[1,⋆]

1. Private trader; Adjunct Instructor, DePaul University
⋆Contact: paulteetor@yahoo.com

Keywords: Financial spreads, securities trading, cointegration, seasonality

In the financial markets, a spread is the difference between two securities. These spreads often have more predictable characteristics than the underlying securities, making them fertile ground for trading.

The author implemented a suite of applications written in R for analyzing spreads. These applications include analysis of mean-reverting spreads; analysis of seasonal spreads; logistic regression models; local polynomial smoothing; and some analysis of relative value trades. The applications are intended for "grey box" trading, where the trader and the computer work together to discover trading opportunities.

The poster session will display some of the analytics and graphics available in the applications. It will also outline the software architecture behind the applications.

References

Teetor, Paul (2007). "Predicting the Direction of Swap Spreads," `http://quanttrader.info/public/predictingSwapSpreads.pdf`.

Teetor, Paul (2009). "Finding Seasonal Spreads," `http://quanttrader.info/public/findingSeasonalSpreads.html`.

Teetor, Paul (2009). "Using R to Test Pairs of Securities for Cointegration," `http://quanttrader.info/public/testForCoint.html`.

robCompositions: An R-package for robust statistical analysis of compositional data

Matthias Templ[1,2,⋆], **Karel Hron**[3], **Peter Filzmoser**[1]

1. Department of Statistics and Probability Theory, Vienna University of Technology
2. Methods Unit, Statistics Austria
3. Palackỳ University Olomouc, Czech Republic

⋆Contact author: templ@tuwien.ac.at

Keywords: Multivariate Methods for Compositional Data, Robustness, R-package **robCompositions**

Compositional data are data that contain only relative information (see, e.g. Aitchison 1986)). Typical examples are data describing expenditures of persons on certain goods, or environmental data like the concentration of chemical elements in the soil. If all the compositional parts would be available, they would sum up to a total, like 100case of geochemical concentrations. Frequently, practical data sets include outliers, and thus a robust analysis is desirable. The R-package **robCompositions** (Templ et al., 2009) contains functions for robust statistical methods designed for compositional data, like principal component analysis (Filzmoser et al., 2009a) (including the robust compositional biplot), factor analysis (Filzmoser et al., 2009b), and discriminant analysis (Filzmoser et al., 2009c). Furthermore, methods to improve the quality of compositional data sets are implemented, like outlier detection (Filzmoser et al., 2008), and imputation of missing values (Hron et al, 2010). The latter one, based on a modified k-nearest neighbor algorithm and a model-based imputation, is also supported with measures of quality of imputation and diagnostic plots. The usage of the package will be illustrated on practical examples.

References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. *Chapman & Hall*, London.

K. Hron, M. Templ, P. Filzmoser (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics and Data Analysis*, to appear.

P. Filzmoser and K. Hron (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40(3), 233-248.

P. Filzmoser, K. Hron, and C. Reimann (2009a). Principal component analysis for compositional data with outliers. *Environmetrics*, 20, 621–632.

P. Filzmoser, K. Hron, C. Reimann, and R.G. Garrett (2009b). Robust factor analysis for compositional data. *Computers and Geosciences*, 35, 1854–1861.

P. Filzmoser, K. Hron, and M. Templ (2009c). Discriminant analysis for compositional data and robust parameter estimation. Research Report SM-2009-3, Department of Statistics and Probability Theory. Vienna University of Technology, 28 pages.

M. Templ, K. Horn, P. Filzmoser (2009). robCompositions: Robust Estimation for Compositional Data. Manual and Package, R package version 1.3.3.

Many Solvers, One Interface — ROI, the R Optimization Infrastructure Package

Stefan Theußl Kurt Hornik David Meyer

Currently, R and a wide variety of contributed packages on CRAN and other package repositories offer tools to solve many different optimization problems (Theußl, 2009). However, the user interfaces to available optimizers and their output, i.e., the format of the returned solution, often differ considerably. It is not only the users interested in R as an optimization tool, but also the developers who need to handle different problem classes transparently and who are facing this lack of standardization. Therefore, an integrative multi-purpose optimization framework for R seems to be desirable.

In this talk we present the state of the art of general purpose continuous and mathematical programming solvers available in R. We discuss different classes of optimization problems encountered in statistical computing and related fields in order to derive the necessary building blocks for a general optimization infrastructure package. Based on the gained knowledge we propose an object oriented approach to optimization: The main function in the R Optimization Infrastructure package **ROI** takes only three arguments, namely the problem object containing information about the optimization problem which is to be solved, the solver to use, and a list of additional control arguments like e.g., solver parameters. This makes the process of optimizing a given problem very transparent for the user as it is completely abstracted from the solver. Furthermore, this approach allows for easy switching between solvers, given that the solver supports the corresponding problem class and thus enhances comparability.

Finally, we show that the complexity of creating large problem instances can be significantly reduced via appropriate constructor functions and methods.

References

Stefan Theußl. CRAN task view: Optimization and mathematical programming, 2009. URL `http://CRAN.R-project.org/view=Optimization`.

Integration of R to VTK, Adding Statistical Computation to a Visualization Toolkit

Jeff Baumes[1], Andinet Enquobahrie[1], Thomas O'Connell[2], Tom Otahal[3], Philippe Pébay[4], Wesley Turner[1*], Michelle Williams[5]

1. Kitware Inc., Clifton Park, NY, U.S.A.
2. The Hamner Institutes for Health Sciences, 6 Davis Dr. P.O. Box 12137, Research Triangle Park, NC 12137, U.S.A.
3. Sandia National Laboratories, Albuquerque NM, U.S.A.
4. Sandia National Laboratories, MS 9159, PO Box 969, Livermore CA 94551, CA, U.S.A.
5. University of Washington, 1959 NE Pacific St., Seattle WA 98195 U.S.A.

*Contact author: wes.turner@kitware.com

Keywords: R, VTK, Information Visualization, Informatics

Conveying the sense of complex data to the human mind requires sophisticated visualization methods. The Titan [4] informatics toolkit, a Sandia funded collaboration between Sandia National Laboratory and Kitware, represents an effort to add graphical, tabular, and geospatial visualization algorithms to the Visualization Toolkit (VTK) [3]. VTK is an open-source, freely available software system for 3D computer graphics, image processing and visualization. The Infovis additions to VTK expand the the toolkit to include visualization of spatially ambiguous entities. However, simply displaying relationships among entities is not sufficient. Statistical analysis such as that provided by R is a powerful tool for suppressing noise in the data and enhancing real relationships. This abstract describes the addition of an R interface to the VTK toolkit and introduces the use of the R engine in several VTK Infovis application areas.

The interface to R primarily consists of three VTK classes: vtkRAdapter, vtkRInterface, and vtkRCalculatorFilter that provide a separation between the VTK and R code. vtkRAdapter handles the conversions between R and VTK data structures allowing VTK tables and arrays to be converted to and from R SEXP data structures. vtkRInterface launches an instance of the R interpreter and manages access. vtkRCalculatorFilter uses instances of the vtkRAdapter and vtkRInterface classes to specify the R command to be executed. It also provides the programmatic interface to specify VTK data structures as inputs to the R analysis and R variables as outputs from the analysis.

We are currently working on two applications that leverage this interface. The first application is statistical hypothesis testing. More specifically, an option to perform statistical tests has been added to the descriptive [1] and contingency [2] VTK statistics engines: respectively, the Jarque-Bera normality and χ^2 independence tests. Although the test statistics themselves can and are directly computed in VTK, the calculation of the corresponding p-values requires that one-tailed probability values of the χ^2 distribution be available. We integrated this feature with the `EvalScript()` method of the VTK/R interface. Examples that demonstrate this functionality are available in the toolkit. A second application is an 'omics analysis viewer targeting the exploitation of genomic, metabolomic and proteomic data in the discovery and analysis of disease processes. Investigators at the University of North Carolina and the University of Washington, are collaborating with Kitware on this application to analyze genes, metabolites and proteins from biological assays; discover the statistically important components; and explore biological pathways modeling the disease process. This application links the visualization capabilities of VTK with R; allowing for statistical analysis including PCA, K-Means clustering, Student T tests, correlations, and pathway analysis. The combination of R with VTK allows for rapid prototyping of algorithms and for intuitive visualizations of the results.

References

[1] P. Pébay and D. Thompson, *Scalable descriptive and correlative statistics with Titan*, Sandia Report SAND2008-8260, Sandia National Laboratories, December 2008.

[2] ———, *Parallel contingency statistics with Titan*, Sandia Report SAND2009-6006, Sandia National Laboratories, September 2009.

[3] Bill Lorensen Will Schroeder, Ken Martin, *The visualization toolkit, an object oriented approach to 3d graphics*, third ed., Kitware, Inc., 2004.

[4] Brian Wylie and Jeffrey Baumes, *A unified toolkit for information and scientific visualization*, Visualization and Data Analysis 2009 (Katy Borner and Jinah Park, eds.), SPIE, 2009, p. 72430H.

BradleyTerry2: Flexible Models for Paired Comparisons

Heather Turner[1*], **David Firth**[1]

1. University of Warwick
*Contact author: Heather.Turner@warwick.ac.uk

Keywords: statistical modeling, social statistics

There are many situations where a binary choice is made between two objects—two sports teams compete against each other to win a match, for example, or survey participants are asked to indicate their preference given two options, say. In such situations, the Bradley-Terry model may be used to model the odds of one object, or 'player', beating the other. The model assumes that each player has some 'ability' and that the odds of one player beating another are given by the ratio of the corresponding abilities.

The standard Bradley-Terry model estimates a separate ability parameter for each player, however the model can be modified in several ways. For example, it may not be reasonable to assume constant ability in situations where players gain in experience or the outcome of a contest can depend on the conditions under which the contest occurs. In such cases, contest-specific variables should be included in the model. Another possibility is that the substantive interest lies in the dependence of player ability on covariates. In this case, the player abilities may be modeled by a linear predictor rather than individual parameters, however random effects should then be added to allow for variability between players with the same covariate values.

This talk presents the **BradleyTerry2** package, an extended version of the **BradleyTerry** package (Firth, 2005) with a more flexible interface. The facilities for model specification allow variables that vary by contest, player, judge, or any other relevant index and also allow specification of random effects. An implementation of the Penalised Quasi-Likelihood algorithm (Breslow and Clayton, 1993) is provided for mixed models. In the talk, a number of applications will be presented that illustrate the main features of the new package and future developments will also be discussed.

References

Breslow, N. E. and Clayton, D. G. (1993). *JASA*, 88(421), 9–25.

Firth, D (2005). Bradley-Terry Models in R. *JSS*, 12(1).

Front propagation using fast marching in R

Daniela Ushizima[1,2,⋆]

1. Lawrence Berkeley National Laboratory
2. National Energy Research Scientific Computing Center

⋆Contact author: daniela@hprcd.lbl.gov

Keywords: fast marching, computer vision, EBImage

Image segmentation is one of the most important challenges in computer vision since all the other steps in pattern recognition depend on the definition of regions of interest. Among the algorithms for image segmentation, level sets and fast marching have been used frequently in problems in shape recovery. A drawback of level sets is its computational expense, particularly when considering scripting languages. We implement the Fast Marching method in R, a numerical technique for solving the Eikonal equation as a boundary value problem without iteration over the whole data. In this scheme, a front advances monotonically with a speed function that never changes sign, such that the marching method efficiently computes the arrival time at each grid point by modeling a moving interface under an inward or outward motion. We illustrate bellow the method using a single seed and multiple seeds in a constant velocity field.

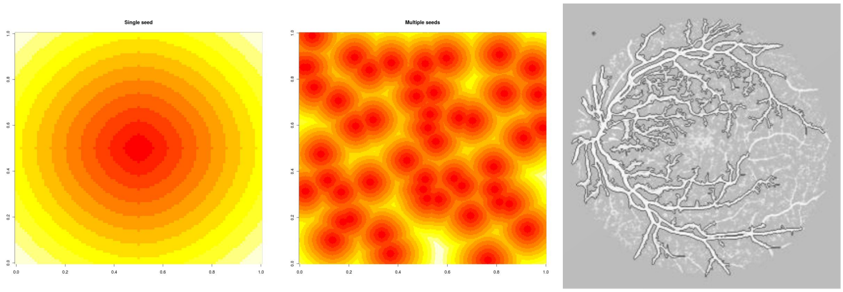


Figure 1: Front propagation given a single seed (left) and multiple seeds (center) in a constant velocity field and application of front propagation to biomedical images (right).

This implementation requires the following R packages: **PolynomF** for solving the quadratic polynomial in the fast marching method. Extensions could consider **biOps** for image filtering and mathematical morphology to provide more sophisticate speed functions. Future work include the use a C++ heap data structure, which characterizes FM fast capability of solving the equation in O(nlogn) and encapsulate the code into a package to be available.

References

Sethian JA. Level Set Methods and Fast Marching Methods, Cambridge Press, 2005. `http://math.berkeley.edu/~sethian`.

Martins CIO, Veras RMS, Ramalho GLB, Medeiros FNS, Ushizima DM. Automatic microaneurysm detection and characterization through digital color fundus images. *IEEE SBRN'08*, 2008.

Ushizima DM, Cuadros J. Image analysis of ocular fundus for retinopathy characterization, Bay Area Vision Meeting, Feb 2010.

Distribution based outlier detection with the extremevalues package

Mark P.J. van der Loo[1]

1. Statistics Netherlands, PO box 24500, 1490 HA The Hague, the Netherlands
Contact: m.vanderloo@cbs.nl

Keywords: Economic data, outliers, QQ-plot, distribusion-based outlier detection

Outlier detection is performed by statistical agencies, such as Statistics Netherlands, to identify observations that either contain errors or have to be treated differently in the estimation process.

The **extremevalues** package is an implementation of the outlier detection methods described in van der Loo (2010), which were recently developed to detect outliers in economic data. The method can be applied when an approximate data distribution is known. For example, in the reference it is shown that certain types of economic data distributions (Value Added Tax turnover values) often resemble a lognormal distribution.

Distribution based outlier detection takes advantage of this knowledge in two steps: first, the distribution's parameters are determined robustly, by fitting a (possibly transformed) subset of the data to the QQ-plot positions for the model distribution. Second, a test is performed to deceide whether the smallest or largest values are outliers. Our method involves two test options: the first option computes a value above (below) which less than ρ (say 1.0) observations are expected, given the sample size N. The second option tests the hypothesis that a small (large) value can be drawn from the model distribution using its fit residual as a test statistic.

The extremevalues package currently supports outlier detection, assuming the normal, lognormal, Pareto, exponential or Weibull distribution as a model. It also includes a number of plotting facilities which can be used to graphically analyze the outlier detection results. See Figure 1 for an example.

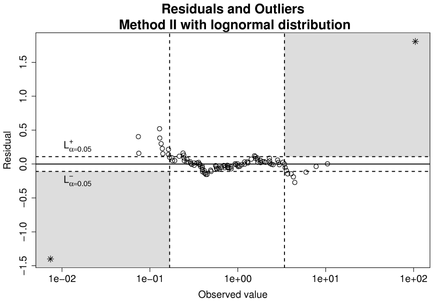


Figure 1: Results of outlier detection on a simulated dataset, using the second test method. Outliers (indicated with a $\ast$) are detected using the fit residuals as a test statistic. The horizontal lines indicate $\alpha = 0.05$ probability levels, assuming normally distributed residuals. Points between the vertical dotted lines were used in the fit.

References

M.P.J. van der Loo (2010). Distribution based outlier detection for univariate data, Discussion paper 10xxxx *Statistics Netherlands, the Hague, (in press)* `http://www.cbs.nl/en-GB/menu/methoden/onderzoek-methoden/discussionpapers/archief/2010/default.htm`.

M.P.J. van der Loo (2010) **extremevalues**: a package for outlier detection in unvariate data, R package version 2.0
`http://cran.r-project.org`,`http://www.markvanderloo.eu`

SQUAREM: An R package for Accelerating Slowly Convergent Fixed-Point Iterations Including the EM and MM algorithms

Ravi Varadhan[1*]

1. School of Medicine, Johns Hopkins University
*Contact author: rvaradhan@jhmi.edu

Keywords: acceleration, global convergence, polynomial extrapolation, vector extrapolation

Fixed-point iterations are extremely common in applied mathematics. Well known examples include the Jacobi and Gauss-Seidel iterations for solving a linear system of equations, Newton's method for solving a nonlinear system of equations, power method for finding the dominant eigenvector of a matrix, and the expectation maximization (EM) algorithm for finding the maximum likelihood estimate. Fixed-point iterations that are contraction mappings are particularly attractive because of their global convergence property, i.e. they find the fixed-point from any starting point. Many fixed-point iterations are only linearly convergent, and their convergence can be very slow especially when the linear rate constant is close to 1. Over the past several years, we have been working on a class of numerical schemes called squared extrapolation methods (SQUAREM) for accelerating the convergence of smooth, linearly convergent fixed-point iterations (Varadhan 2004; Roland and Varadhan 2005; Roland, Varadhan Frangakis 2007; Varadhan and Roland 2008; and Varadhan 2010 (under preparation)). These numerical acceleration schemes have many attractive properties: (1) they are easy to implement; (2) they are widely applicable to "any" smooth, linearly convergent fixed-point iteration; (3) they utilize minimal memory and computational effort, and therefore, are ideal for high-dimensional problems; and (4) they provide good trade-off between speed of convergence and global convergence.

Here we discuss an R package called SQUAREM that represents the culmination of our research over the past 5 years. This package can accelerate fixed-point iterations under two different situations: (A) when there is an underlying objective function (i.e. a Liapunov function) that is minimized (or maximized) at the fixed-point, or (B) when there is no underlying objective function. Situation A is common in many statistical problems exemplified by the EM and MM (majorize and minimize) algorithms. Situation B is exemplified by fixed-point iterations for solving nonlinear systems and the power method for finding the dominant eigenvector of a matrix. . The package contains 4 main functions which implement different first-order and higher-order SQUAREM schemes for each of the two scenarios. An essential feature of these algorithms is that the user can explicitly choose the degree of trade-off between speed of convergence and global convergence. For example, when reasonable starting values are available or when the fixed-point mapping is globally contractive, the user can opt for more speed by choosing a large value of the non-monotonicity parameter; conversely, the user can opt for more stability at the cost of lesser acceleration by choosing a small value of the non-monotonicity parameter. In our experience, SQUAREM algorithms generally provide good accelerations (at least 2-3 fold) without sacrificing global convergence. In this talk, we will demonstrate the power and utility of SQUAREM on a number of well-known fixed-point iterations.

Points, Curves and Haystacks: Datavis and Metabolomics

Marie Vendettuoli[1*], Dianne Cook[1], Heike Hofmann[1]

1. Bioinformatics and Computational Biology Program
   Human Computer Interaction Program
   Department of Statistics
   Iowa State University

* Contact author: mariev@iastate.edu

Keywords: GUI, metabolomics, data visualization, self-directed learning, CLI

A challenge to researchers using *R* for analysis of large datasets is the lengthy computation time associated with visualization and the static nature of such images generated using base graphics. This limits opportunities to rapidly gain insights into data structures. One solution is to use a new *R* package, **qtpaint**; rendering graphics much more quickly, especially for large datasets. However, **qtpaint** offers only methods for drawing low-level graphical elements and requires an investment of time and effort on the part of the researcher, both in skill acquisition and programming, to implement. `qtpaintgui()` is a tool created to support portable graphics development, accessible via command-line interaction and a point-and-click GUI. We examine how graphics generated using `qtpaintgui()` allow data visualization approaches to support efforts of the Metabolomics community to encourage transparency and automation in data processing.

References

A. Buja, D. Cook, and D.F. Swayne, (1996) *Interactive High-Dimensional Data Visualization*, *Journal* of Computational and Graphical Statistics, vol. 5, pp. 78-99.

M. Guzdial (2004). Programming environments for novices. *Computer Science Education Research*, S. Fincher and M. Petre, Eds. Taylor & Francis, Abingdon, U.K., 2004: 127–154.

E. Lahtinen, K. Ala-Mutka, H. Järvinen, (2005) A *Study of the Difficulties of Novice Programmers*, Proceedings of the 10th annual SIGCSE conference on Innovation and technology in computer science education, 2005: pp. 14-18

M. Lawrence (2010). *Interfaces to the Qt framework from R.* http://r-forge.r-project.org/projects/qtinterfaces/

J. Yi, Y. Kang, J. Stasko, J. Jacko, (2007). *Toward a Deeper Understanding of the Role of Interaction in Information Visualization*, IEEE transactions on visualization and computer graphics, 2007: 13(6), pp. 1224-1231

The traitr package

Abstract

The traitr package provides an interface to writing GUIs within R as an alternative to using RGtk2 directly or gWidgets. The basic design is inspired by the traitsUI python package, and uses the model-view-controller paradigm. The simplest usage requires the user to simply specify the types of data for the model behind the GUI, leaving the choice of control and layout to the package. More advanced uses include specifying a layout or specifying controllers to be called when changes to the underlying model occur. This talk will introduce the package through several examples.

Teaching Statistics: An example of "How to" improve the students' statistical skills using individualized assignments

Joan Vila[1,2,3*], **Montse Rue**[4], **Nuria Codern**[2], **Albert Sorribas,**[4] **Cristina Rodríguez**[2], **Anna Foraster**[2]

1. Institut Municipal d'Investigació Mèdica (IMIM), Barcelona, Spain
2. Escola Universitària d'Infermeria Creu Roja, Terrassa, Spain
3. CIBER Salud Pública, Barcelona, Spain
4. Universitat de Lleida, Lleida, Spain

*Contact author: jvila@imim.es

Keywords: Teaching Materials, Software Design, LaTeX , Sweave.

Sweave is a well known tool that makes it possible to embed R code in LaTeX documents[1]. When the code is run and the resulting output, figures and tables are automatically inserted into the final document.

We present the results of using Sweave for designing individualized assignments for undergraduate and graduate students of:

- Human Nutrition (HN). University of LLeida, Spain.
- Medicine (ME). University of LLeida, Spain.
- Nursing (NU). Autonomous University of Barcelona, Spain.
- Occupational Therapy (OT). Autonomous University of Barcelona, Spain.

in two different universities:

For each course we chose a clinical trial (CT). The main results of the CT were simulated to create a dataset for each student: 50 HN, 120 ME, 176 NU and 79 OT. Data were *similar* but different for each student in the course. Each student received a spreadsheet with his/her particular data and a PDF with his/her name and id. The assignment consisted of 50 exercises that covered most of the concepts taught during the course.

The exercises and the methods needed to solve them were exactly the same in each discipline, but answers could be the same, similar or absolutely different depending on each dataset.

The exercices consisted of open questions that required calculating a figure (i.e. a Student-t statistic) as well as multiple choice questions for which the answers were randomly ordered. So, not only could answers be different for different students, but even if the answers were the same, the choice could be a different item.

The students provide their answers in a free and open-source e-learning software platform (Moodle or Sakai, depending on the university).

By the assignment deadline, each student received a detailed and personalized answer, showing the correct and detailed explanation of how to solve the problem. In the multiple choice questions we explained why the remaining answers were not correct.

At the end of the semester the students had an exam. We compared the exam grades after applying this new system with historic exams performed two years before with the same teachers and similar students. Results showed dramatic improvement. For example, in the NU program, 52.7% of 186 students in the academic year 2007-08 and 70.6% of the 177 in 2009-10 passed the final exam (p-value <0.001). Furthemore, although the percentage of students who took the exam was similar, 146 (78.5%) in 2007-08 versus 136 (76.8%) in 2009-10, the students that took the exam received higher grades: $\bar{\mathrm{X}}$ (SD): 5.99 (2.12) out of 10 versus 7.43 (1.54) (p-value < 0.001) and the percentage of students that passed the exam was higher: 67.1% versus 91.9% (p-value < 0.001).

Students were surveyed about about strengths and weaknesses of the teaching method. The results of this analysis, obtained using open questions of the type "positive / negative" with qualitative research methodology, provided descriptions and interpretations of the experience of the student in learning statistics on an innovative context. It also identified some room for improvement in future courses.

In conclusion, we present instruction on **how to** perform:

- Individualized assignments
- Simulated data from a Clinical Trial
- Open or multiple random choice questions
- Individualized, detailed, explanatory answers
- Embedding R code in LaTeX documents with `Sweave`

as well as the main quantitative and qualitative results of applying the new method.

References

1. Leisch F (2003). Sweave and Beyond:: Computations on text documents. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria, 2003. `http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/Leisch.pdf/`.

Time Series Inference Applications of R in Finance and Econometrics

H. D. Vinod[1*]

1. FORDHAM UNIVERSITY, New York
*Contact author: vinod@fordham.edu

This paper reviews some recent applications of R software for handling difficult statistical inference issues when the available time series data are state-dependent and evolutionary, rather than stationary. Traditional methods call for first converting the data to stationary to obtain reliable inference when the underlying series are integrated of order d, I(d). We propose an alternative which uses the R package meboot for maximum entropy bootstrap We report some interesting new simulations and applications to econometrics and Finance. For example, we show how the problem of spurious regression can be solved. The proposal has the potential of bypassing a great deal of unit root testing literature.

Bayesian Monitoring of A Longitudinal Clinical Trial Using R2WinBUGS

Yue Wang[1,*] , Narinder Nangia[1]

1. Clinical Statistics, Abbott Laboratories, Abbott Park, IL, USA
* Contact author: yue.wang@abbott.com

Keywords: Bayesian, Clinical Trial, Longitudinal Model, Probability of Success, R2WinBUGS

Learning stage study (a proof of concept or a dose-ranging study) is crucial in the drug development process as decisions to continue or halt development of the drug candidate meeting desired target product profile must be made with incomplete information. It is important to characterize the dose-response curve and estimate probability of success at an interim stage in a clinical trial to facilitate decision-making about further development of the experimental drug. When interim analyses are conducted, some subjects will have complete data, but others will have incomplete or partial information. We handle the partial data using a longitudinal model and Bayesian imputation. An algorithm that characterizes relationship between the early responses of subjects and their final responses in a longitudinal hierarchical model is developed to account for the uncertainty associated with having missing observations in estimating the dose-response curve. This algorithm facilitates remote execution of WinBUGS from within R using **R2WinBUGS**. Implementation of the Bayesian approach for monitoring a longitudinal clinical trial will be presented using a normal dynamic linear model for the dose-response curve.

Analyzing Direct Marketing Data with R

Liang Wei*, Brendan Kitts

Lucid Commerce LTD Inc.
*Contact author: lwei@lucidcommerce.com

Keywords: Direct Marketing, Data Warehousing, ROI, Data Visualization, Business Intelligence

Lucid Commerce LTD Inc. (Lucid), headquartered in Seattle, WA, provides analytic solutions for direct marketers to maximize their revenue of investment (ROI) from targeted advertising. Lucid has built several advanced statistical models using R and have them integrated into our real-time media optimization and planning system, those models have significantly boosted Lucid's media planning performance and also simplified the analytics process. This talk will demonstrate how to create an R tool chain which combines *Microsoft SQL Server 2008* for data storage and data warehousing, *Microsoft SQL Server Reporting Services (SSRS)* for report rendering. By utilizing R packages such as **MASS**, **RPART** and also many of our own R scripts, we extract large amount of data through **RODBC** library, have them analyzed and send the outputs back to the database. We also utilizes R's powerful graphics components together with several libraries such as **Lattice** to generate reports and visualize high dimensional data. This talk will discuss both the practical challenges with operating R on a large scale as well as the areas where the language excels.

References

R Development Core Team (2005). R: A Language and Environment for Statistical Computing, `http://www.R-project.org`.

Lucid Commerce LTD Inc. (2009). Lucid Commerce's website, `http://www.lucidcommerce.com`.

W.N. Venables, B.D. Ripley (2002). Modern Applied Statistics with S.

John Chambers (2008). Software for Data Analysis: Programming with R.

Deepayan Sarkar (2008). Lattice: Multivariate Data Visualization with R

Elizabeth J. Atkinson, Terry M. Therneau (2000). An Introduction to Recursive Partitioning Using the RPART Routines, `http://www.mayo.edu/hsr/techrpt/61.pdf`

Are there latent decision rules in expert occupational exposure assessments?

David Wheeler[1*], Kai Yu[1], Melissa Friesen[1]

1. National Cancer Institute, Bethesda, MD, USA
* Contact author: wheelerdc@mail.nih.gov

Keywords: CART, data mining, decision trees, epidemiology

The expert assessment approach to determine occupational exposure risk factors based on questionnaire responses in population-based epidemiology studies is often criticized because it occurs in a ‘black box’ and does not provide any mechanism for applying the expert’s decision rules to other studies that used the same questionnaires. However, there are likely latent rules used by the experts while determining the exposure assignments. In this analysis, we use data mining methods implemented in *R*, including classification and regression trees (CART) and tree ensembles, to determine if latent rules can be uncovered from questionnaire responses and an expert’s assigned exposure metrics in a study of diesel exhaust exposure. Uncovering the latent decision rules provides a mechanism for replicating these decision rules in other subjects within or across studies, making the often lengthy exposure assessment process in epidemiologic studies more efficient.

The REQS package for linking the SEM software EQS to R

Eric Wu, Patrick Mair, Peter Bentler

In this talk we present the REQS package which connects EQS, a software for structural equation modeling, and R. The package consists of three main functions that read EQS script files and import the results into R, call EQS script files from R, and, finally, run EQS script files from R and, again, import the results after EQS computation.We explain the functionalities of the package and show how to use it by means of several examples with special emphasis on SEM simulations

Creating Animations with R

Yihui Xie[1,*]

1. Department of Statistics, Iowa State University
*Contact author: xie@yihui.name

Keywords: Animation, Demonstration, Simulation, R

The R package **animation** (Xie, 2010) was designed to demonstrate and explain statistical ideas in an interesting way. Xie and Cheng (2008) is a short introduction to this package. In this talk, we will give a few examples in statistics first, ranging from the classical probability theory and mathematical statistics (e.g. Buffon's needle, CLT) to applications in model diagnostics (e.g. outlier detection). Then we will introduce four approaches in the **animation** package to create animations with R, i.e. via HTML, GIF, Flash and LaTeX. Finally we will take a look at other possibilites to create animations, such as **swfDevice** (Bracken, 2009) and **SVGAnnotation** (Temple Lang, 2009).

References

Bracken C (2009). *swfDevice: R graphics device for swf (flash) output.* R package version 0.1, URL `http://swfdevice.r-forge.r-project.org/`.

Temple Lang D (2009). *SVGAnnotation: Tools for post-processing SVG plots created in R.* R package version 0.5-0.

Xie Y (2010). *animation: Demonstrate Animations in Statistics.* R package version 1.1-0, URL `http://animation.yihui.name`.

Xie Y, Cheng X (2008). "animation: A Package for Statistical Animations." *R News*, **8**(2), 23–27. URL `http://CRAN.R-project.org/doc/Rnews/`.

Analysis of interlaboratory studies using R and the metRology package

James Yen, Stephen Ellison

Interlaboratory studies are vital to measurement science. The R package metRology contains a number of functions for exploring interlaboratory data, including both graphical representations and statistical tests. Motivating examples include comparison studies that test the measurement capabilities of measurement laboratories.

Social network analysis with R sna package

George Zhang
iResearch Consulting Group
bird@iresearch.com.cn
birdzhangxiang@gmail.com

February 25, 2010

This speech is mainly a share of learning experience about using sna package in R. For beginners it can also be used as a handbook for social network analysis. We hope to promote the use of sna in China and looking for more cases to practice the sna method.

Content catalog is listed here:

1. Social network definition
   Actual graph: scale free, small world
   Sample graph
2. Network description, GLIs
   Vertex edges distribution-example: epidemiology
      Exponential random graphs (ERGs)
   Edge strength distribution
   Basic measurement
   Path and cycle census
   Measure of structure
      Connectedness, Hierarchy, efficiency, lubness
   Graph centrality
      degree, betweenness, closeness
3. Relation between GLIs
4. Graph distance and clustering:
   1) Graphs distance: hdist(Hamming Distances), sdmat, structdist
   2) Vertices distance: equiv.clust(structural equivalence), sedist, geodist
5. Graph cov based function
   Canonical correlation
   Prime component analysis
   Linear/logit regression
   Linear autocorrelation model
      Combined theory-example: telecom client trend
6. Random graph models
   Network evolution
      Random
      Biased
   Statistic test
   Bayesian Network

Use of R in Genetic Epidemiology Designs

Jing Hua Zhao[1*]

1. MRC Epidemiology Unit, Institute of Metabolic Science, Box 285, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK

*Contact author: jinghua.zhao@mrc-epid.cam.ac.uk

Keywords: Genetic Epidemiology, Study Design, Power Calculation

Statistical genetics or genetic epidemiology is essential in unravelling the genetic and environmental influences on common diseases and/or quantitative measurements in human population. The major strategies for study of these complex traits have been analyses of familial aggregation, segregation, linkage and association, with the latter two relying on the availability of genetic markers. While advance of high throughput genotyping technologies has contributed significantly to the recent success in such studies, adequate study designs remain to be critical.

With a brief introduction of genetic epidemiology and statistics, I will elaborate R implementations for several scenarios including family, case-control, case-cohort, staged designs and Mendelian randomisation studies. I will provide updated results for those as reported in an earlier paper, as well as practical use and empirical results which has clearly demonstrated the need of more analytical work, Connection will be further made to other available software, which encapsulates extensions to be made in the near future.

The presentation has been motivated from our work on design and analysis of several large epidemiological cohorts and also complementary to a pre-conference tutorial on genetic analysis of complex traits. The examples should add to the list of successful stories of the R environment in research work.

References

Armitage P, Colton T. (2005). Encyclopedia of Biostatistics, Second Edition. *Wiley.*

Zhao JH (2007). **gap**: genetic analysis package. *J Stat Soft*, 23(8), 1–18.

Generalized Linear Mixed Model with Spatial Covariates

Alex Zolot[1*]

[1]. StatVis Consulting
*Contact author: alex.zolot@statvis.com

Keywords: kriging, GLMM, spatial analysis, mixed model, tcltk

We analized measured numerical characteristics $z = z(i, x, y)$ of subjects i that belong to a number of groups $G : i \in G$ and are in different spatial locations characterized by 2D coordinates (x_i, y_i) . Our task was to separate the group and spatial components in the measured characteristics:

$$g(z) = f_1(G) + f_2(x, y) + noise$$

where g is a linking function and $f_1(G)$ is considered as a fixed or random effect.

We solved the problem in R by iteration with sequence of general linear mixed model (glmm) and kriging cross-validation `krige.cv` (package **gstat**), using exponential fitting of variogram with four parameters: (nugget, range, sill and anisotropy). GUI was done in package **tcltk**.