The R User Conference, useR! 2011
August 16-18 2011
University of Warwick, Coventry, UK

Book of Contributed Abstracts

Compiled 2011-07-28

Contents

Spectroscopic Data in *R* and Validation of Soft Classifiers: Classifying Cells and Tissues by Raman Spectroscopy

Claudia Beleites[1,2,*], Christoph Krafft[2], Jürgen Popp[2,3], and Valter Sergo[1]

1. CENMAT and Dept. of Industrial and Information Engineering, University of Trieste, Trieste/Italy
2. Institute of Photonic Technology, Jena/Germany
3. Institute of Physical Chemistry and Abbe Center of Photonics, University Jena/Germany

*Contact author: cbeleites@units.it

Keywords: spectroscopy, soft classification, validation, brain tumour diagnosis

Medical diagnosis of cells and tissues is an important aim in biospectroscopy. The data analytical task involved frequently is classification. Classification traditionally assumes both reference and prediction to be *hard*, i. e. stating exactly one of the defined classes. In reality, the reference diagnoses may suffer from substantial uncertainty, or the sample can comprise a mixture of the underlying classes, e.g. if sample heterogeneity is not resolved or if the sample is actually undergoing a transition from one class to another (e. g. rather continuous de-differentiation of tumour tissues). Such samples may be labelled with *partial* or *soft* class memberships.

Many classification methods produce soft output, e. g. posterior probabilities. Methods like logistic regression can also use soft training data. Yet, for medical diagnostic applications it is even more important to include soft samples into the model validation. Excluding ambiguous samples means retaining only clear (i. e. easy) cases. Such a test set is not representative of the original unfiltered population, and creates a risk of obtaining overly optimistic estimates of the model performance.

With **softclassval** (softclassval.r-forge.r-project.org), we introduce a framework to calculate commonly used classifier performance measures like sensitivity and specificity also for samples with soft reference and prediction. Briefly, if the soft class labels are interpreted as uncertainty, best and worst case as well as expected performance are obtained via the weak, strong and product conjunction (and-operators, see e. g. Gottwald, 2010). For the mixture interpretation, weighted versions of well-known regression performance measures like mean absolute and root mean squared errors are derived.

As real world example, we classify 37 015 Raman (thereof 55 % soft) spectra of 80 brain tumour patients into "normal", "low grade", and "high grade" tissue morphologies in order to delineate excision borders during surgical treatment of the tumours. Thus, borderline cases are our actual target samples. We demonstrate spectroscopy-related functionality supplied by **hyperSpec** (hyperspec.r-forge.r-project.org) and its conjoint use with other packages.

Financial support by the Associazione per i Bambini Chirurgici del Burlo (IRCCS Burlo Garofolo Trieste) and of the European Union via the Europäischer Fonds für Regionale Entwicklung (EFRE) and the "Thüringer Ministerium für Bildung, Wissenschaft und Kultur" (Project: B714-07037) is highly acknowledged.

References

Gottwald, S. (2010). Many-valued logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2010 ed.).

Revisiting Multi-Subject Random Effects in fMRI

Jonathan Rosenblatt[1]

1. Raymond and Beverly Sackler Faculty of Exact Sciences. Tel Aviv University, Tel-Aviv, Israel
*Contact author: rosenbla@post.tau.ac.il

Keywords: fMRI, Finite Mixture, EM

Introduction Random Effects analysis has been introduced into fRMI research in order to generalize findings from the study group to the whole population (see Friston et al.2002). Generalizing findings is obviously harder then detecting activation in the study group since in order to be significant, an activation has to be larger then the inter-subject variability which is assessed and controlled for within the study group. Indeed, detected regions are smaller when using random effect analysis versus fixed effects. The statistical assumptions behind the classic random effects model are that the effect is normally distributed over and "activation" refers to a non-null mean effect. We argue this model is unrealistic and conservative compared to the true population variability.

Method Inspired by the Two-Populations model widely used in genetics (Efron2008) and in fMRI (Hartvig and Jensen20 we develop a model for the inter-subject voxel-wise effect as a mixture of two populations: One in which the voxel was activated by the paradigm with a normally distributed effect and a second, in which the voxel is inactive, thus- a null effect. We suggest two justifications for this model: (a) Ill registration might map active and non-active voxel to the same locations. (b) Brain plasticity permits the same anatomical location to serve different functions.

The implementation of the method was done in *R* and included: (a) ML estimation of the parameters of the proposed finite mixture using an EM algorithm over a non-convex constrained parameter space (to solve identifiability issues). (b) Bootstrapping of the estimates' distribution for inference on the number of mixed populations, known to be an analytically unsolved problem (Garel2007) . (c) Optional parallelization of the Bootstrapping using *Condor* and/or the **snow** package.

Results We demonstrate our method on real fMRI dataset of 67 subjects at 60,000 brain locations. We construct estimate maps and p-value maps of the voxel-wise proportion of population responding to an experimental paradigm. Once these have been constructed, we define "activation" as locations where more then a given percentage of the population has been found active. We revisit the activation maps created under the classical definition of activation and compare them to activation found under this new definition to show the power gained using the finite Gaussian mixture.

References

[Efron2008] Efron, B. 2008. Microarrays, empirical Bayes and the two-groups model. Statistical science 23 (1): 1–22.

[Friston et al.2002] Friston, K. J., D. E. Glaser, R. N. A. Henson, S. Kiebel, C. Phillips, and J. Ashburner. 2002. Classical and Bayesian Inference in Neuroimaging: Applications. NeuroImage 16 (2): 484–512 (June).

[Garel2007] Garel, Bernard. 2007. Recent asymptotic results in testing for mixtures. Computational Statistics & Data Analysis 51 (11): 5295–5304 (July).

[Hartvig and Jensen2000] Hartvig, N.V., and J. L. Jensen. 2000. Spatial mixture modeling of fMRI data. Human Brain Mapping 11 (4): 233–248.

Putting the R into Randomisation

Zoe Hoare [1,⋆]

1. NWORTH, Bangor trials unit, Bangor University, Y Wern, LL57 2PZ http://www.bangor.ac.uk/imscar/nworth/
⋆Contact author: z.hoare@bangor.ac.uk

Keywords: Randomisation, Clinical trials, Simulation, Validation

When introducing a new dynamic adaptive randomisation method to the clinical trials field the functionality and flexibility of *R* allowed the programming of the algorithm without any restrictions. Whilst using *R* for implementing randomisation procedures is not a new idea, expanding its use to include a modelling stage for optimisation of individual trials was a particularly useful addition.

The setting up of the novel randomisation procedure in *R*, linking it to Excel using RExcel for simulation purposes, validation of the processes and the possibilities for linking to the web for web based centralized randomisation processes will all be addressed.

The randomisation algorithm used is fully tuneable by setting model parameters, reproducing anything between deterministic allocation (minimisation) and simple randomisation. Simulation with *R* allowed investigation the robustness of the algorithm, the statistical properties of bias and loss in terms of a clinical trial and the plausible outcomes of a wide variety of possible situations that may arise. The algorithm can now be tuned before the trial starts to give confidence intervals around the split of the final allocation.

Inputs to the simulations allow trials to be customised for differing number of participants, number of treatment groups, ratios of allocation to treatment groups, number of stratification variables and their defined levels and sets of parameters to be tested. The algorithm is currently used in two ways: for sequential and for complete list randomisation. While the base code is the same the wrapper for the implementation varies slightly.

The use of *R* in clinical trials is slowly becoming more accepted and has certainly been discussed at length on the *R* help mailing lists. Validation of any system is integral to ensuring the system is functioning correctly. There have been instances where systems that have not been properly validated have resulted in significant costs within the clinical trial world. However, mistakes like this are not a reason to avoid using more complex methods altogether. All the code written within *R* was easily tested and validated within an expected operating range ensuring that the system is doing exactly what is expected.

Future work entails linking into web based system, developing reporting functionality of the simulation tool and improve the robustness of the tool for multiple users.

References

Baier, T. and E. Neuwirth (2007, April). Excel :: Com :: R. *Computational Statistics 22*(1), 91–108.

Group, C. S. (2001). Effect of low dose mobile versus traditional epidural techniques on mode of delivery : a randomised controlled trial. *The Lancet 358*, 19–23.

Henry, S., D. Wood, and B. Narasimhan (2009). Subject Randomization System. In *useR! 2009, The R User Conference (Rennes France)*, pp. 85.

Hewitt, C. and D. Torgerson (2006). Is restricted randomisation necessary? *BMJ 332*, 1506–1508.

Russell, D., Z. Hoare, R. Whitaker, C. W. C, and I. Russell (2011). Generalized method for adaptive randomisation in clinical trials. *Statistics in Medicine*, Early view published online DOI 10.1002/sim.4175.

Bringing the power of complex data analysis methods into *R*

Teh Amouh[1,⋆], Benot Macq[2], Monique Noirhomme-Fraiture[1]

1. School of Computer Science, University of Namur, Belgium
2. School of Engineering, University of Louvain-la-neuve, Belgium
⋆Contact author: tam@info.fundp.ac.be

Keywords: Set-valued data, Modal data, Histogram data, Multidimensional data

In classical statistical data analysis, individuals are described by single-valued scalar type variables. A rectangular data matrix defines the relation between a set of individuals and a series of variables such that each cell of the data matrix contains one single scalar value (quantitative or categorical).

Sometimes, however, individuals require to be described by set-valued type variables. For example, in a time budget study, a variable that records the daily time spent watching television would not allow single-valued answer (such as 3 hours), because this value normally varies from day to day for each individual. An interval-valued answer like "between 2.5 and 3.5 hours", reflecting an internal variability, would be more appropriate. When considering a stock market, information on an unpredictable share price would include the degree of uncertainty. A possible statement would be: "the price of share S varies between 130 and 140, with probability 30%", leading to data expressed as an histogram for variable *price* and individual *S*. The classical single-valued cell data table can hardly cope with such complex descriptions. There is a need on a data table model in which each cell could contain a (weighted) listing of values taken from a predefined set, and not just a single quantitative or categorical value. The `data.frame` model provided by *R* does not apply.

A research field named *symbolic data analysis* (Bock and Diday (2000)) and defined as the extension of standard data analysis to complex data, proposes a great deal of methods for set-valued data analysis (Diday and Noirhomme-Fraiture (2008)). These data are called *symbolic data* and encompass interval type data (which means subsets of the set of real values), multi-valued categorical type data (which means listings of ordinal or nominal categories) and multi-valued quantitative type data (which means listings of numerical values). A listing of categories can be summarized as a weighted set of distinct categories or as a discrete probability distribution. In either case we talk about modal data. A listing of numerical values can be summarized as an interval with the lower and upper bounds being respectively the minimum and maximum values in the listing. A listing of numerical values can also be summarized or as a histogram (if the range of values in the listing is segmented into intervals) or as a cumulative density function. A *symbolic data table* is a rectangular data matrix which allows such complex data values in each of its cells. Powerful data analysis methods are available for these types of data.

In order to bring the power of set-valued data analysis methods into *R*, we develop appropriate data structures using both *S3* and *S4* object approaches available in *R* (Chambers (2008)). Our data structures include table objects that extend the `data.frame` object and allow complex data values in each cell. This talk is about the design and implementation of these data structures. An *R* package containing these data structures will be available as a basic building bloc for the *R* implementation of complex data analysis methods.

References

Bock, H.-H. and E. Diday (Eds.) (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin: Springer-Verlag.

Chambers, J. (2008). *Software for Data Analysis: Programming with R*. Springer.

Diday, E. and M. Noirhomme-Fraiture (Eds.) (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.

Using the Google Visualisation API with R

Markus Gesman, Diego de Castillo

Contact authors: rvisualisation@gmail.com
Project site: http://code.google.com/p/google-motion-charts-with-r/

Keywords: Google Visualisation API, R.rsp, RApache, JSON

In 2006 Hans Rosling gave an inspiring talk at TED (Rosling, 2006) about social and economic developments in the world over the last 50 years, which challenged the views and perceptions of many listeners. Rosling had used extensive data analysis to reach his conclusions. To visualise his talk, he and his team at Gapminder (Gapminder Foundation, 2010) had developed animated bubble charts, aka motion charts.

Rosling's presentation popularised the idea and use of interactive charts, and as a result the software behind Gapminder was bought by Google and integrated as motion charts into their Visualisation API (Google Inc., 2010h) one year later.

In 2010 Sebastián Pérez Saaibi (Saaibi, 2010) presented at the R/Rmetrics Workshop on Computational Finance and Financial Engineering the idea to link Google motion charts with R using the **R.rsp** package (Bengtsson, 2009).

Inspired by those talks and the desire to use interactive data visualisation tools to foster the dialogue between data analysts and others the authors of this article started the development of the **googleVis** package (Gesmann and de Castillo, 2011).

The **googleVis** package provides an interface between *R* and the Google Visualisation API. The Google Visualisation API offers interactive charts which can be embedded into web pages. With the **googleVis** package users can create easily web pages with interactive charts based on *R* data frames and display them either via the *R* HTTP help server locally or within their own sites. The current version 0.2.4 of the package provides interfaces to Motion Charts, Annotated Time Lines, Maps, Geo Maps, Tables and Tree Maps.

This session will provide an overview of the package functionality and the authors will share examples and experiences with the audience.

References

H. Bengtsson. R.rsp: R server pages. http://CRAN.R-project.org/package=R.rsp, 2009. R package version 0.4.0.

Gapminder Foundation. Gapminder. http://www.gapminder.org, 2010.

M. Gesmann and D. de Castillo. googleVis: Using the Google Visualisation API with R. http://code.google.com/p/google-motions-chart-with-r/, 2011. R package version 0.2.4.

Google Inc. Google Visualization API. http://code.google.com/apis/visualization/documentation/gallery.html, 2010h.

H. Rosling. TED Talk: Hans Rosling shows the best stats you've ever seen. http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html, 2006.

S. P. Saaibi. *R/RMETRICS Generator Tool for Google Motion Charts*. https://www.rmetrics.org/, 2010. Meielisalp, Lake Thune Switzerland, June 27 - July 1, 2010.

Experimenting with a `tty` connection for *R*

Matthew S Shotwell

Department of Biostatistics, Vanderbilt University
Contact author: Matt.Shotwell@Vanderbilt.edu

Keywords: terminal, devices, hardware, user interface

The Portable Operating System Interface (POSIX; IEEE, 2004) offers a standard concept of the *computer terminal* and an associated Application Programming Interface (API). Operating systems that natively support the POSIX concept include GNU Linux, BSD, and Mac OS X. Support for Microsoft Windows is provided by third-party software, such as Cygwin (Red Hat, 2011). The POSIX terminal serves as a common mechanism for asynchronous data transfer, including keyboard input, USB and other hardware communications, and software-to-software connections. We present a parsimonious extension to *R* that implements the POSIX terminal API as an *R connection*. The new feature is styled as a '`tty` connection'.

Applications of the `tty` connection are broad in scope. This presentation highlights applications to the use*R* interface, and in hardware communications. We demonstrate how the `tty` connection facilitates common user interactions, such as 'press any key to continue' functionality, and password entry without displaying the password text. A live demonstration is prepared using *R* to control an external USB temperature sensor, as well as a GPS navigation device. We conclude with a discussion of integrated data collection and signal processing in medical devices.

References

Institute of Electrical & Electronics Engineers, Inc. (2004). IEEE Std 1003.1 (POSIX), 2004 Edition, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1309816.

Red Hat, Inc. (2011). Cygwin (version 1.7) [Software], http://www.cygwin.com

The R Ecosystem

David M Smith[1,*]

1. VP Marketing and Community, Revolution Analytics
*Contact author: david@revolutionanalytics

Keywords: community, users, ecosystem

In the last couple of years, and especially since the New York Times feature article about *R*[1], I have observed a dramatic shift in the perceptions of R. Long gone are the days when it was regarded as a "niche" statistical computing tool mainly used for research and teaching. Today, *R* is viewed as a critical component of any data scientist's toolbox, and has been adopted as a layer of the analytics infrastructure at many organizations. As a result, a broad ecosystem has sprung up around open-source R. In this talk, I'll share my observations on this evolving ecosystem, including:

- Estimates of the number of R users
- The R Community: people sharing R software and experiences
- How R is being used in academia and in industry
- R's role in the Data Science movement
- How R is bringing modern Statistics to business applications

The talk will conclude with some thoughts on what the future holds for R's role in the worlds of statistics, data science, and business.

References

Vance, Ashlee (2009). "Data Analysts Captivated by R's Power". The New York Times, January 6 2009.

Rc^2: R collaboration in the cloud

E. James Harner[1⋆] **and Mark Lilback**[1]

1. Department of Statistics, West Virginia University
⋆Contact author: jharner@stat.wvu.edu

Keywords: R, cloud computing, collaboration, web 2.0

Rc^2 is a cloud-based, collaborative, web 2.0 interface to R that works with WebKit-based browsers (e.g., Safari and Chrome). In addition to desktop browsers, client-specific style sheets and scripting provide a touch-optimized interface for mobile devices such as the iPad. With Rc^2, R sessions are no longer tied to a specific computer or user.

Rc^2 is designed for simplicity and ease-of-use from the start, allowing students to learn R without solely imposing a command-line interface. At the same time, power users will find the system flexible enough to meet most of their needs. Users can have per-project workspaces, and instructors can predefine workspaces containing the required data and R packages for each assignment.

The groundbreaking features of Rc^2 arise from the collaborative and instructional benefits that cloud-based computing brings. Instructors can schedule classroom sessions where students can watch the instructor interact with R in real-time. The instructor can turn control of R over to individual students as a virtual blackboard to pose questions and work through problems while still communicating via voice chat. Integration with the social web allows instructors to provide notifications to students over the messaging platform of their choice, be it email, SMS, Twitter, or Facebook.

The same features allows researchers to collaborate over the Internet without concern for data becoming out of sync. Users can start long-running computations and Rc^2 will notify the user(s) when the process is complete. Full support for Sweave allows users to easily include, update, and format R output within LaTeX documents for both classroom assignments and publishable papers.

Rc^2 is designed to be a fast, scalable, distributed system with features like load-balancing, multiple authentication mechanisms (password, LDAP, Kerberos), dynamic auto-configuration using mDNS and DNS-SD, fast and persistent client-server communications using WebSockets, and custom R packages for database interaction and graphics generation. Rc^2 can start on a single server and expand to a cluster of computers optimized for each tier of the system (web server, application server, R instances, and database).

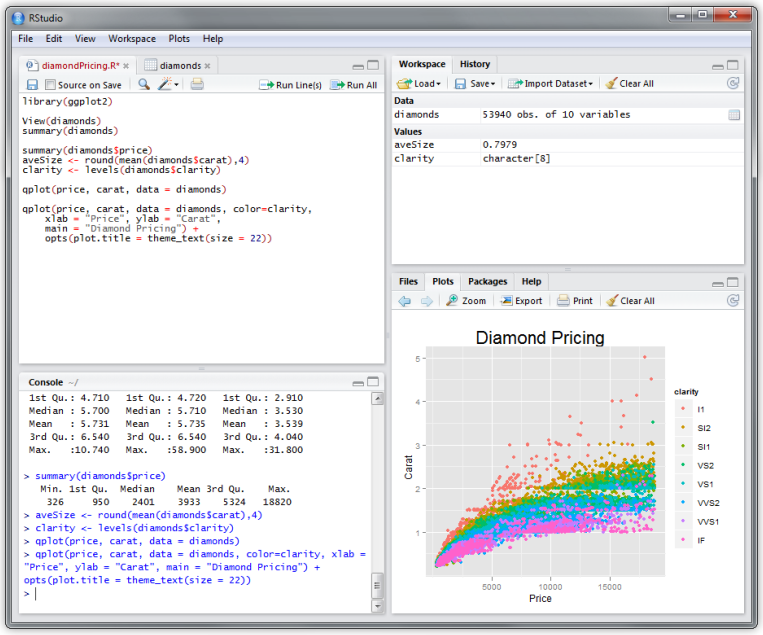
RStudio: Integrated Development Environment for R

J.J. Allaire[1*]

1. CEO of RStudio
*Contact author: jj@rstudio.org

Keywords: GUI, Tools, Sweave, Web

RStudio is a new Integrated Development Environment (IDE) for the R programming language. RStudio is an open-source project intended to combine the various components of R (console, source editing, graphics, history, help, etc.) into one seamless and productive workbench. It is designed to both ease the learning curve for new R users as well as provide high productivity tools for more advanced users. RStudio runs on all major platforms including Windows, Mac OS X, and Linux. In addition to the desktop application, RStudio can be deployed as a server to enable web access to R sessions running on remote systems.



RStudio running on Windows

For the useR! 2011 conference, we would like to introduce RStudio as a new open-source IDE for the R community. We will cover its various features in depth as well as tips on how to optimize workflow. We will cover the basics (console, editing, plots, etc) and also discuss more advanced features including code completion, searchable history and authoring LaTeX and Sweave documents.

The presentation will also cover the details of setting up an RStudio server. We will show a hands on demonstration of installing RStudio on an Ubuntu Server as well as describe the various configuration options and administrative utilities.

NOTE: Currently, RStudio is not publicly available (it is in beta) however will be released well in advance of the conference. Our website http://www.rstudio.org/ will also be live at this time. For the purposes of considering this presentation, please contact us to receive access to the website.

R in the Practice of Risk Management Today

Jagrata Minardi[1,*]

1. TIBCO Software, Inc.
*Contact author: jminardi@tibco.com

Keywords: financial risk management, portfolio management, simulation

This talk explores several concrete computational tasks in risk management and portfolio management, in the true scale at which they occur in large financial institutions. Using *R* and well-known packages, there is strong coverage of analytical methods, but challenges remain as size of problems grow and answers are required more quickly.

How does today's software improve our ability to manage risk in the time needed by business? When models change, how quickly can they be brought to market? How has recent developments in both regulatory bodies and the competitive landscape changed the business problems themselves? We'll identify the state of the art against several key challenges in financial risk management, and indicate how software---and the industry itself---are changing the playing field.

Stress Testing with R-Adamant

Rocco Claudio Cannizzaro[1,⋆], **Fausto Molinari**[2]

1. R-Adamant
2. R-Adamant
⋆Contact author: rc.cannizzaro@r-adamant.org

Keywords: R-Adamant, Stress Testing, VaR, Expected-Shortfall

Value-at-Risk (VaR) and Expected-Shortfall are standard and widely used risk-management tools.
The calculation of such metrics is generally based either on some a-priori assumption on the statistical distribution of the time series or on historical data analysis and simulations.

Even when assumptions are made on the form of the distribution, the estimation of the related parameters is typically based on historical data and the choice of how much history should be used is a non trivial problem: short data history produces noisy estimates while long historical analysis results in biased estimations which may not be reflective of the changing nature of the markets and their volatilities.

An alternative approach is to exploit the correlation between the performance of the managed portfolios and the macro economic conditions in which the underlying sectors of the investment are operating.
Once the link is established, it is possible to evaluate the impact of macro economical changes on the performance of the given investments and generate a multitude of possible forecast scenario outcomes from which the risk metrics (VaR, ES) can be computed.

Stress scenarios/events can also be analysed and the associated risks quantified, leading to a more dynamic and consistent assessment of the financial position of the investments.

During the presentation we will demonstrate how to exploit **R-Adamant** to simulate several possible scenarios for the economy and the market, and attempt to forecast financial trend and returns of a portfolio in each of those scenarios.
We will use the statistical and graphical tools contained in **R-Adamant**, like: Efficient Portfolio estimation, VaR and ES estimation, Vector Autoregressive models, and Monte-Carlo simulations.

The tutorial scope is to show how a valuable tool as **R-Adamant** can be a powerful ally for researchers and university students for their thesis and for companies.

References

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

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

Markowitz, Harry M. (1991), *Portfolio Selection*, second edition, Blackwell.

Steven M. Kay (1993), *Fundamentals of Statistical Signal Processing*, Volume 2: Detection Theory.

Handling Multiple Time Scales in Hedge Fund Index Products

Jovan Njegic[1,*], Drago Indjic[2,3]

1. Business School of Novi Sad, Serbia
2. Sunningdale Capital LLP, UK
3. London Business School, UK
*Contact author: jovan.nj@uns.ac.rs

Keywords: hedge fund data sources, time series, risk management, R functions

The analysis of portfolio of hedge funds suffers from numerous data quality problems. In particular case of hybrid fund structure investing across the full universe of hedge fund products in marketplace with contractual liquidity ranging from intra-daily to annual, we have to handle time series with varying periodicity and possibly multiple price status (estimated, executed, audited etc).

We have been using virtually all commonly known *R* time series packages libraries (such as **xts**, **zoo**) to clean and 'synchronize' the time series in order to provide standard risk reporting and risk limit management functionality, exposing several design flaws and fixing minor bugs (e.g. `to.weekly`, `Return.portfolio`) and attempting to develop a new wrapper function specifically addressing hedge fund data sources and features.

This work is based on "*R* in Modern Portfolio of Hedge Fund Analysis" presentation given at London *R* group meeting in December 2010.

References

Indjic, D., (2010), *R* in Modern Portfolio of Hedge Fund Analysis, http://www.londonr.org/LondonR-20090331/R%20in%20modern%20FoHF.pdf

AFLP: generating objective and repeatable genetic data.

Thierry Onkelinx [1,⋆], **Pieter Verschelde**[1], **Sabrina Neyrinck**[1], **Gerrit Genouw**[1], **Paul Quataert**[1]

1. Research Institute for Nature and Forest (INBO), Scientific Services, Gaverstraat 4, B-9500 Geraardsbergen, Belgium
⋆Contact author: Thierry.Onkelinx@inbo.be

Keywords: Genetics, normalisation, classification, repeatability, randomisation

Amplified fragment length polymorphisms (AFLP) is a technique to get a DNA fingerprint from an organism. The total genomic DNA is split into fragments at fixed combinations of nucleotides, resulting in fragments of different lengths. After electropherese we get a DNA profile ("bands" or "peaks") of intensity versus fragment length. High intensity indicates the presence of fragment(s), low intensity the absence.

Our package AFLP, available on R-Forge[1], supports the analysis of AFLP data on several points: i) randomise the samples and add replicates, ii) normalise the intensity for differences due to laboratory handlings and equipment, iii) classify the normalised intensity into presence/absence data, iv) calculate the repeatability of the analysis and v) apply multivariate analysis on the classified data.

Replicates are added during the randomisation of the samples. This allows to measure the repeatability and helps the normalisation between batches. Therefore both within batch as between batch replicates are added. Bonin et al. (2004) suggest a "technical difference rate" (TDR) to estimate the repeatablity of two replicates from the same sample. We adapted this TDR formula to work with more than two replicates per sample. Furthermore we use this TDR not only to measure the repeatability of the samples, but also the fragments and the overall repeatability.

The normalisation is based on linear mixed models (**lme4**). The idea is to model the average intensity in function of the different sources of variability (DNA extraction, lab batch, capillary, ...). The raw residuals of the model are used as the normalised intensity. Several diagnostic plots are available to highlight potential problems. The normalised intensity is classified based on their distribution per fragment. The distribution is bimodal for a polymorph fragment (both presences and absences) and unimodal for a monomorph fragment (only presences or only absences). The lowest density between the modi is used as threshold for the classification. The user can get graphs displaying the distribution and the selected threshold. The build-in analysis capabilities are currently limited to methods for `hclust` and `princomp`. Both the normalised and classified data can be exported to other packages when needed.

Advantages of this package: i) The package starts from the design in the lab, ensuring the necessary replicates and randomisation. ii) The normalisation considers a much wider variety of sources than other software. iii) The intensity, typically lognormal distributed, can be transformed during the normalisation. iv) The classification threshold depends not an arbitrary threshold, but on the distribution of the normalised intensity. v) TDR is used as a measure of overall quality, quality per sample and quality per fragment.

References

Bonin, A., E. Bellemain, P. Bronken Eidessen, F. Pompanon, C. Borchmann, and P. Taberlet (2004). How to track and assess genotyping errors in population genetic studies. *Molecular Ecology 13*, 3261–3273.

[1] http://r-forge.r-project.org/projects/aflp/

The nz.seq package for handling genetic sequences in the Netezza Performance Server

Przemysław Biecek[1,2,⋆], **Paweł Chudzian**[1,3], **Peter Gee**[1], **Jeff Foran**[1], **Justin Lindsey**[1]

1. NzLabs, Netezza, an IBM company,
2. Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw,
3. Faculty of Electronics and Information Technology, Warsaw University of Technology

⋆Contact author: przemyslaw.biecek@gmail.com

Keywords: genetic sequences, parallel processing, relational database, next generation sequencing.

In the first part of this talk we will show why it is advantageous to store genetic sequences in a relational database instead of flat files. In the second part we will present the *R* interface supporting massively parallel operations on genetic sequences which are stored in the relational database. The presented implementation of **nz.seq** package is specific to the Netezza Performance Server (NPS) database but the presented approach is general and might be easily applied to any other database with a similar architecture, i.e. the parallel share nothing environment.

DNA, RNA and amino acid sequences are usually stored in flat files instead of relational databases. The three main reasons for that are: it is easier to copy or download flat files, it is easier to load flat files into a statistical packages like *R* and it is easier to use command line tools like SAMtools, bowtie or bwa. On the other hand storing data in a database allows one to incorporate the data structure, and allows for fast access to required sequences based on their properties or metadata information. Note that the volume of genetic data may be huge, dozens or thousands of terabytes, thus the access time to a required sequence may be significant. Moreover, databases support efficient merging of different kinds of data, e.g. genetic sequences might be combined with other features like sex, age, medical history, geographical coordinates etc.

In the NPS one can take advantages of both approaches. Sequences are stored in the relational database, but some operations might be performed in all computing nodes transparently to the internal database storage model.

The package **nz.seq** leverages the communication with the database in two ways:

- From the *R* client one can easily download any sequence or set of sequences from the remote database using their names, IDs or some matching criteria. Afterwards in the local *R* client one can operate on such sequences and store results from local operations in the remote database.
- Instead of downloading the data from the remote database to local *R* client, one can upload the *R* code from the local *R* client to the remote database. The submitted *R* script is then applied to all selected sequences in a fully parallel way. That decreases the overall computation time by reducing the data download time, which can be a significant improvement due to the large volume of data. Most of the *R* functions from the *R* packages available on CRAN or Bioconductor might be used in such remote operations. One can easily apply a variety of *R* operations to a large number of sequences stored in the remote database.

At the end of this talk we will present a real life application using data from The 1000 Genomes Project.

Classification of Coverage Patterns

Stefanie Tauber[1,⋆], Fritz Sedlazeck[1], Lanay Tierney[2], Karl Kuchler[2], Arndt von Haeseler[1]

1. Center for Integrative Bioinformatics Vienna, Max F Perutz Laboratories, University of Vienna, Medical University of Vienna, University of Veterinary Medicine, Dr.-Bohr-Gasse 9, A-1030 Vienna, Austria
2. Christian Doppler Laboratory Infection Biology, Max F Perutz Laboratories, Medical University of Vienna, Campus Vienna Biocenter, Dr.-Bohr-Gasse 9, A-1030 Vienna, Austria
⋆Contact author: stefanie.tauber@univie.ac.at

Keywords: Next Generation Sequencing Data, Coverage, Fractals

The advent of DNA sequencing technologies [Metzker (2009)] has brought along an enormous amount of data that still poses a fundamental data-analysis challenge for bioinformaticians and biostatisticians.

When speaking of sequencing data the term 'coverage' is widely used but, at the same time, not well defined. It has to be distinguished between theoretical ('sequencing depth') and observed ('local') coverage. The local coverage can be defined as an integer vector counting per nucleotide the number of reads mapping to the respective nucleotide. In the following the term 'coverage' always refers to the observed local per nucleotide coverage.

In genome resequencing we expect and aim for uniform coverage whereas technologies like RNA-Seq [Ozsolak and Milos (2010)] or ChIP-Seq [Park (2009)] are especially interested in coverage jumps. However, any kind of differential expression analysis relies on a count table containing the number of mapped reads per gene model. This summarization step is not well investigated and its implications on the downstream analysis are not fully understood yet. It is obvious that a summarization value like the sum of reads per gene model is not able to exhaustively capture the underlying coverage information.

Therefore we introduce the fractal dimension (FD) [Kaplan and Glass (1995)] and the Hurst exponent (H) [Peitgen et al. (1992)] in order to distinguish between more or less 'reliable' coverage patterns. FD, as well as H do not make use of any user-defined parameters and are hence free of any ad-hoc heuristics. We propose a re-weighting of the read counts with both FD and H yielding a more reliable count table.

Additionally we show the influence of different mapping strategies on the observed coverage patterns and read counts. This is of course of special interest as any mapping peculiarity propagate to all downstream analysis. We discuss our results on a large Illumina RNA-Seq data set. The host-pathogen interaction of Candida albicans and dendritic mouse cells are investigated by a time course design with three replicates per time point.

We illustrate the entire analysis as well as all up-mentioned methods by means of a *R* package we are developing.

References

Kaplan, D. and L. Glass (1995). *Understanding Nonlinear Dynamics*. New York: Springer.

Metzker, M. L. (2009, December). Sequencing technologies the next generation. *Nature Reviews Genetics 11*(1), 31–46.

Ozsolak, F. and P. M. Milos (2010, December). RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics 12*(February).

Park, P. J. (2009, October). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics 10*(10), 669–80.

Peitgen, H.-O., H. Jürgens, and D. Saupe (1992). *Chaos and Fractals*. New York: Springer.

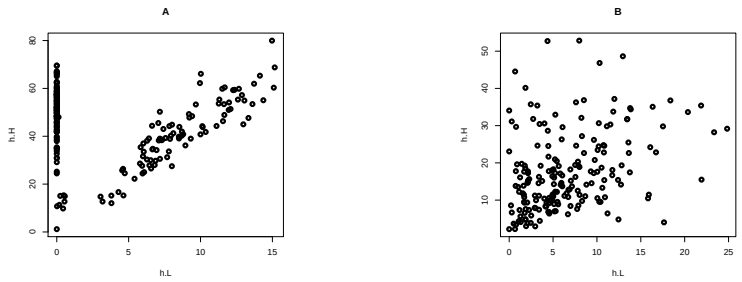
Finite Mixture Model Clustering of SNP Data

Norma Coffey[1,⋆], **John Hinde**[1], **Augusto Franco Garcia**[2]

1.National University of Ireland, Galway
2. Department of Genetics, ESALQ/USP, Piracicaba, Brazil
⋆Contact author: norma.coffey@nuigalway.ie

Keywords: Clustering, finite mixture models, orthogonal regression, SNPs.

Sugarcane is polypoid, i.e. has 8 to 14 copies of every chromosome, with individual alleles in varying numbers. It is therefore important to develop methods that identify the many different alleles and associated genotypes. One way of doing this is through the analysis of single nucleotide polymorphisms (SNPs). The frequency of a SNP base at a gene locus will vary depending on the number of alleles of the gene containing the SNP locus. Capturing this information accurately across several SNPs can give an indication as to the number of allele haplotypes present for a gene. Such information could have implications for sugarcane breeding since high yield potential may be due to the presence of and/or different number of copies of, a specific allele(s) present at a gene locus. The figures below display the data collected for two SNPs of the sugarcane plant. Each point represents the intensity of two SNP bases; h.L is the intensity of the A base, h.H is the intensity of the T base. The data in Figure A can clearly be clustered into two groups - the group along the y-axis and the group along the line with a particular (unknown) angle. These groups correspond to two genotypes and thus clustering is essential for genotyping. In Figure B it is not clear how many clusters (genotypes) are present and therefore it is necessary to develop a technique that can determine the number of clusters present, determine the angles between the clusters to identify different genotypes, and provide a probabilistic clustering to identify points that have high probability of belonging to a particular cluster (have a particular genotype) and those that are regarded as an unclear genotype.



The above criteria indicate that model-based cluster analysis techniques could be useful for analysing these data. However standard model-based cluster analysis techniques such as those implemented in the *R* package **mclust** (Fraley and Raftery, 2002) attempt to fit sperical/ellipsoildal components thus failing to cluster these data in an appropriate way and do not provide estimates of the angles between the clusters. To determine these angles it is necessary to fit a regression line to the data in each cluster. Using finite mixtures of linear regression lines is also inappropriate since it is not clear for these data which is the response variable and which is the explanatory variable. Problems are also encountered when attempting to fit a regression line to the group parallel to the y-axis in Figure A since this line has infinite slope in the usual least squares setting. As a result we propose to use finite mixtures of *orthogonal* regression lines to cluster the data, which ensures that using either variable as the response variable yields the same clustering results and that a regression line can be fitted to the group parallel to the y-axis. We implement this technique in *R* and show its usefulness in clustering these data.

References

Fraley, C. and A. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*, 611–631.

GPU computing and R

Willem Ligtenberg[1,⋆]

1. OpenAnalytics BVBA
⋆Contact author: willem.ligtenberg@openanalytics.eu

Keywords: R, GPU, OpenCL, High-performance computing

Modern CPUs currently have 4 to 8 cores. In contrast an affordable GPU (200 euro) has 384 cores. This makes them ideally suited for parallel processing. Examples of problems that can benefit from processing on the GPU are Fast Fourier Transforms, the k-nearest neighbour algorithm or sequence alignment. If the problem can be cut into small pieces, it might be interesting to port it to the GPU.

In order to make this compute power more accessible there a few options. One is to make some of the basic *R* functionality use the GPU directly, a second option is to wrap *C(++)* applications that make use of the GPU with *R*. Finally, we can build an interface between *R* and *OpenCL*, which leaves all possibilities open for the user and therefore provides maximum flexibility.

In the first part of our presentation we will provide an overview of current *R* packages that use the GPU. In the second part, we will present our progress on the **ROpenCL** package. **ROpenCL** provides functions to transfer data frames to and from the GPU and to push kernel functions to the GPU, thereby unlocking the full potential of the GPU. These kernel functions still need to be written in *OpenCL*, however the interfacing with the GPU can all be done with *R*. The **ROpenCL** package is similar in spirit and based on the PyOpenCL *Python* library Klöckner (Klöckner).

References

Klöckner, A. PyOpenCL. http://mathema.tician.de/software/pyopencl.

Deploying and Benchmarking R on a Large Shared Memory System

Pragneshkumar Patel[1] , George Ostrouchov[1,2], Luke Tierney[3]

1. University of Tennessee
2. Oak Ridge National Laboratory
3. University of Iowa

*Contact author: pragnesh@utk.edu

Keywords: high performance computing, parallel computing, multicore architecture, multithreading

We describe our experience in deploying *R* on a large shared memory system, *Nautilus* (SGI UltraViolet), which has 1024 cores (Intel Nehalem EX processors), 4 terabytes of global shared memory, and 8 NVIDIA Tesla GPUs in a single system image. This system is part of RDAV, the University of Tennessee's Center for Remote Data Analysis and Visualization sponsored by the National Science Foundation as part of TeraGrid XD.

One of our goals on *Nautilus* is to provide implicit parallel computing for serial *R* codes. Implicit parallelism automatically exploits multiple cores without changes to user code. Another goal is to provide an environment where users can effectively explore explicit parallelism through the use of many *R* packages that have been developed recently for parallel computing.

For implicit parallelism, we report on Intel's Math Kernel Library (*MKL*) and on the **pnmath** package. *MKL* includes a high performance multithreaded *BLAS* implementation. The **pnmath** package provides multithreading to many *R* math functions for operating on large vectors. We report on benchmark runs for many core count and vector size combinations, which we use for optimal speedup calibration. We observe speedups in excess of 600 for some compute intensive **pnmath** functions when operating on large vectors.

In addition to reporting on specific *BLAS* and **pnmath** functions, we include **R Benchmark 2.5** and some of its modifications. If time permits, we also mention experiences with RDAV center customers and some packages for exploiting explicit parallelism.

References

Luke Tierney (2008). Implicit and Explicit Parallel Computing in R, *COMPSTAT* 2008, Part II, 43-51.

Markus Schmidberger, Martin Morgan, Dirk Eddelbuettel, Hao Yu, Luke Tierney, Ulrich Mansmann (2009). State of the Art in Parallel Computing with R. *Journal of Statistical Software*, August 2009, Volume 31, Issue 1.

Luke Tierney (2010). pnmath, http://www.stat.uiowa.edu/~luke/R/experimental/.

Intel Math Kernel Library, http://software.intel.com/sites/products/documentation/hpc/composerxe/en-us/mklxe/mkl_userguide_lnx/index.htm.

The "CUtil" package which enables GPU computation in R

Kazutaka Doi[1,*†] **, Kei Sakabe**[†]

1. Regulatory Science Research Program, National Institute of Radiological Sciences, Japan
[†] Both authors contributed equally to this work
*Contact author: kztkdi@gmail.com

Keywords: CUDA, GPU computation, operator overriding, Windows

There are some *R* packages which uses GPU for computation, and the better performance is provided by GPU-equipped computers. However, the current number of implemented functions in their packages is limited because special procedures are required for utilization of GPU, such as memory allocation in video memory, and data transfer between main memory and video memory. Furthermore, these packages are not easy to use for Windows users because binary packages are not provided, and it is not easy for Windows users to build a developing environment. Since there seems to be a large number of potential Windows users, it is beneficial to develop a package which enables GPU computing easily on Windows. Therefore, we began to develop a new package which utilize GPU in computation with minimum modification of *R* source code, and also can be used easily for Windows users. The package development is ongoing, and the detail specifications are not fixed, except that the package is developed with NVIDIA's CUDA toolkit. Our package **CUtil** (**C**UDA **Util**ity package) will be equipped with the following features.

As described above, the first feature is that Windows users will be able to use this package easily. This package will be available from CRAN, and we will also try to make Windows binary packages available. The second feature is that after loading the package, the frequently used operators and functions can be overridden. With this overriding, we can enjoy GPU computation performance with minimum modification of *R* source code. The third feature is that the package minimize the data transfer cost between main memory and video memory. The data on video memory is treated as external pointer in *R* objects, and once the *R* object on main memory is transferred to video memory for GPU computation, the data is kept on video memory as *R* objects. The next time of the GPU computation, the data can be used directly by GPU without the data transfer. Because it is known that the proportion of time needed for the data transfer is relatively high, this advantage will be beneficial especially for a long series of computations, such as Markov chain Monte Carlo methods in Bayesian statistics. Other than these features, we are going to implement double precision floating-point and complex number computation in addition to single floating-point, and garbage-collection, which enables us to use a small amount of video memory effectively. This package will require computers with NVIDIA's GPU (compute capability is equal to or greater than 2.0).

References

Buckner J, Wilson J, Seligman M, Athey B, Watson S, Meng F (2010). The gputools package enables GPU computing in R. *Bioinformatics 26*, 134-135.

Adelino Ferreira da Silva (2010). cudaBayesreg: Bayesian computation in CUDA. *The R Journal 2*, 48-55.

OBANSoft: integrated software for Bayesian statistics and high performance computing with R

Manuel Quesada[1,*], **Domingo Giménez**[1], **Asunción Martínez**[2]

1. Universidad de Murcia
2. Universidad Miguel Hernández
*Contact author: manuel.quesada@um.es

Keywords: Bayesian statistics, parallelism, OpenMP, MPI, CUDA

Among all the applications for Bayesian analysis, there is none that integrates the whole process of Objective Bayesian Analysis. The aim of the "Software for Objective Bayesian Analysis" (*OBANSoft*) is to cover this gap. The first version includes the easiest models (Yang and Berger (1996)) but its design will allow for more complex algorithms in time.

The statistics engine has been implemented using *R*. Implementing the algorithms with this language lets us take advantage of the increasing number of new routines that the *R* community is developing, so we can focus the effort on other purposes such as high performance computing. For this reason, *OBANSoft* has been designed taking into account that it is going to incorporate a variety of parallelism levels, and it will be the first *R* parallel in a shared memory architecture. Moreover, the heterogeneity of parallel architectures will obstruct non advance users in parallelism optimizing the performance of their programs without the help of an expert. Concerning parallel performance, *OBANSoft* is ready to include an auto tuning module that configures the best parameters to execute the parallel algorithms (Katagiri et al. (2004)).

Although the statistics engine is developed in *R*, this is envolved in a top layer defined like a *Java* interface (using ***JRI*** to link them). This main layer coordinates the lower layers where the different engines based on several architectures are implemented. Indeed, this top layer will implement the auto-tuning algorithms. The final user will only use a desktop application that will solve a problem in the most efficient way for each platform.

The first version is a complete and integrated *Java* Desktop Application (*OBANSoft*) implementing the first models used to teach Bayesian analysis (Quesada (2010)). All the models are implemented using *R* libraries that have been reorganized to compose our *R* engine. The ***SnowFall*** library (SnowFall (2011)) allows exploitation of parallelism in multicore systems. However, the performance of this library is not satisfactory enough and other parallelism strategies need to be considered.

In conclusion, we have developed the base application where we are going to include more complex models with higher computational needs. Our concern is to implement and link the *R* engine with other parallel languages and high performance libraries (BLAS, OpenMP, MPI, CUDA, etc.). *OBANSoft* will assume all those new routines, including auto-tuning decisions. Our goal is to distance the final user from the problems of parallel computation.

Referencias

Katagiri, T., K. Kise, H. Honda, and T. Yuba (2004). Effect of auto-tuning with user's knowledge for numerical software. In *Proceedings of the 1st conference on Computing frontiers*, pp. 12–25. ACM.

Quesada, M. (2010, Julio). Obansoft: aplicación para el análisis bayesiano objetivo y subjetivo. estudio de su optimización y paralelización. Master's thesis, Universidad de Murcia.

SnowFall (2011). Url. `http://cran.r-project.org/web/packages/snowfall/`.

Yang, R. and J. O. Berger (1996). A catalog on noninformative priors. *Discussion Paper, 97-42, ISDS, Duke University, Durham, NC*.

R2wd: Writing MS-Word Documents from R

Christian Ritter[1,2] , Nathan Uyttendaele[1]

1. Université Catholique de Louvain, Louvain-la-Neuve, Belgium.
2. Ritter and Danielson Consulting, Brussels, Belgium.
*Contact author: christian.ritter@ridaco.be

Keywords: COM services, automatic reporting.

R2wd is a package for creating new and modifying existing *MS-Word* documents. It allows to outline the document structure (title, headings, sections), insert text content, *R* graphs in metafile and bitmap formats, *R* dataframes as *Word* tables, mathematical formulas written in a *LaTeX* like style, and to apply *Word* themes to the document. Bookmarks of all inserted elements are created automatically which can be useful in building cross references and in creating tagged PDF documents from the original *Word* documents. **R2wd** helps creating documents automatically from *R* scripts and thereby supports automatic and traceable reporting. In addition, the code behind the **R2wd** functions serves as a tutorial on how to use *COM* objects under Windows. In this talk, we shall show how **R2wd** works and point out which tasks already work well and where there are still challenges. Although **R2wd** has already been used in real automatic reporting situation, it is still work-in-progress. We encourage the community of *R* users to help us in making **R2wd** more reliable and powerful.

By default **R2wd** is based on the **rcom** package which, in turn, makes use of the *statconnDCOM* connector. It can also be adapted to work with the legacy **RDCOM** package.

References

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

FRAD - Fast Results Analysis and Display

Dmitry Bliznyk[1,*]

1. Researcher and PhD student at Riga Technical University
*Contact author: dmitrijs.bliznuks@rtu.lv

Keywords: R, fast analysis, universal usage, template

Paper presents approach for statistical data visualization and analysis. For this purposes the template based on *R* package were created. The FRAD - Fast Results Analysis and Display template provides possibility to make fast data analysis even for inexperienced user without the need of programming. It is ready for processing ASCII input files (e.g. CSV) and producing timeline, histogram and summary plots as well as summary table in .csv format. Possible application areas: data quick overview, visual comparison of data (timeline, histogram), big datasets analysis and results publishing.

FRAD template is not limited to specific task. It's designed to provide the power of *R* language for broad range of users, which are dealing with data analysis. The template contains many features, which are aimed to reduce time that is needed to get the first data analysis results. Some examples of features: automatic range selection accordingly to minimal and maximal values in all analyzed files, skipping corrupted data, automatic axis labeling. One more valuable feature is that proposed approach prevents "human made mistakes" in analysis, specifically in result files naming. In FRAD result files are named automatically using input file name and analyzed data type.

User of FRAD template is able to choose from three levels of control according to his needs and available time. After analyzing big group of user needs, the default parameters where chosen for the first level of control, which is fully automatic. If user needs more freedom, there is the parameters file that gives ability to select analysis features and tune visual look of the output plots. And finally user is able to change template's *R* code, to adjust anything.

FRAD template proves proved itself as valuable tool in research work of the "ifak"[1] institute, specifically in ZESAN[2] project. Previously used Origin[3] software could not satisfy demands in fast test case data analysis. Since Origin does not support automated analysis of multiple files, it could be effectively used for test case data analysis, where number of files per test case is more than ten. For one test case analysis (consisting of four files) user previously should spend around an hour, with FRAD template it takes just two minutes (at 1 million records per file). Advantage is huge!

References

[1] Ifak - Institut für Automation und Kommunikation e.V. Magdeburg, www.ifak.eu

[2] ZESAN project - Reliable, energy-efficient wireless sensor/actor networks for building automation, supervision and process control, http://www.ifak.eu/index.php?id=651&L=3

[3] Origin software, http://www.originlab.com/

The Emacs Org-mode: Reproducible Research and Beyond

Andreas Leha[1,*], **Tim Beißbarth**[1]

1. Department of Medical Statistics, University Medical Center Göttingen
*Contact author: andreas.leha@med.uni-goettingen.de

Keywords: Literate Programming, Reproducible Research, emacs, Org-mode

Reproducible research and literate programming are widespread techniques in the R community using the power of the well established Sweave package [Leisch, 2002] and other approaches (e.g. R2HTML [Lecoutre, 2003], odfWeave [Kuhn et al., 2010], via docutils [Dasgupta, 2010]). The main advantage of all literate programming efforts is, that (R-)code and results, mainly tables and plots, are kept together. This ensures that the reported results originate from the current version of the code, making the report reproducible and adding other benefits, like simplifying version control for results and code in conjunction.

Org-mode (http://orgmode.org/) can be used to achieve similar results as Sweave. But while in contrast to Sweave Org-mode is bound to the editor emacs (although there are different ports to vi(m)), it is a far more flexible approach in other aspects:

The most prominent feature is the variability of output formats: the output can be chosen from LaTeX (PDF), HTML, and soon also ODF.

The language that is woven into the text is not limited to R, but includes Perl, Python, Octave, shell, Matlab, Lisp, SQL, and many more. This is a key feature, as quite often data analysis requires the use of different tools on the same data: For example, Perl is frequently used during the data pre-processing step. Likewise, external programs are often integrated for specialized tasks (e.g. short read alignment) and might best be called from command line. Source code blocks in different languages can easily interact within Org-mode by passing the output from one block as input to another block. Results from a code block can be cached to be re-calculated only when the code changed, avoiding repeated execution of lengthy calculations.

It is even possible to do metaprogramming in Org-mode, meaning that one block of source code generates source code again, which will be handled correctly by Org-mode, too.

And finally, Org-mode is much more than a tool for literate programming. Being initially developed for outlining, (TODO) lists, and project and task management, all this functionality is also available to the R programmer.

We present workflows of how to perform data analyses, create reports, and publish results on your website from within one tool.

References

Abhijit Dasgupta. Flexible report generation and literate programming using r and python's docutils module. In *useR!*, 2010. URL http://user2010.org/slides/Dasgupta.pdf.

Max Kuhn, Steve Weston, Nathan Coulter, Patrick Lenon, and Zekai Otles. *odfWeave: Sweave processing of Open Document Format (ODF) files*, 2010. URL http://CRAN.R-project.org/package=odfWeave. R package version 0.7.17.

Eric Lecoutre. The R2HTML package. *R News*, 3(3):33–36, December 2003. URL http://cran.r-project.org/doc/Rnews/Rnews_2003-3.pdf.

Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. URL http://www.stat.uni-muenchen.de/~leisch/Sweave. ISBN 3-7908-1517-9.

Efficient data analysis workflow in *R*

Peter Baker[1]

1. Statistical Consultant & Senior Lecturer, School of Population Health, University of Queensland, Herston, QLD. Australia
*Contact author: p.baker1@uq.edu.au

Keywords: workflow, unix tools, reproducible research, project templates, data handling

After the fifth large data set came through my door in as many months, I thought it would be more efficient to automate workflow rather than start afresh each time. For instance, I knew my collaborators and clients were likely to *tweak* their data a number of times because even though they had provided data with hundreds or even thousands of variables, they'd undoubtedly left out a few important variables or discovered that some were coded wrongly or recorded inconsistently. And of course, I also thought that my collaborators were likely to contact me a couple of days before a grant application or final report was due with a revised data set and new questions in the hope that I could just *press the button* to instantly extract some final answers.

About the same time, I noticed a few interesting posts on R–bloggers and stackoverflow. I particularly liked the software engineering term **DRY** (don't repeat yourself) with suggestions about automating processing with *R* functions or packages. Another post referred to Long (2009) which provides a useful guide to managing workflow for data analysis in large projects. Long's ideas revolve around using *stata* in a `Windows` environment in order to efficiently facilitate replication of work by following a cycle of Planning, Organising, Computing and Documenting. *stata* has some useful features like using variable "labels" in plots and tables (unlike standard *R*), datasignatures and Long provides good strategies for using codebooks for data handling and checking. However, the approach concentrates on manual methods rather than programming tools like *make*, automatic initial data processing, regular expressions for text processing or version control.

Many tools are available for efficiently managing projects and carrying out routine programming tasks. One such tool is *GNU make*. It is standard on `linux` and `MacOSX` and available via `Rtools` or `cygwin` for `Windows`. Originally developed for programming languages like *C* it is well suited to statistical analysis. Since the late 80's I've used *make* to project manage data analysis using *GENSTAT*, *SAS*, *R* and other statistical packages. It is very efficient in only re–running analyses or producing reports when dependencies such as program files (*R*, *Rnw*, ...) or data files change. *R* is used for all data analysis steps described below.

Using the *R* **ProjectTemplate** as a starting point, the following will be outlined:

- the set up of project directories, Makefiles, R program files to read and check data, initial documentation and log files for use with *emacs* `Org-mode` or other editor;
- using codebook(s) to label variables, set up factors with suitable labels;
- producing initial data summaries and plots for data checking and exploration;
- setting up an initial *git* repository for version control; and
- producing initial summaries using literate programming via `Sweave`.

An *R* package which automates the steps above is under development.

References

Kuhn, M. (2011). ReproducibleResearch cran task view. `http://www.r-project.org/`.

Long, J. S. (2009, February). *The Workflow of Data Analysis Using Stata* (1 ed.). Stata Press.

White, J. M. (2010). *ProjectTemplate: Automates the creation of new statistical analysis projects.* R package version 0.1-3.

Teaching Statistics to Psychology Students using Reproducible Computing package RC and supporting Peer Review Framework

Ian E. Holliday[1*] and Patrick Wessa[2]
1. Aston University, School of Life and Health Sciences, Aston Brain Centre, Birmingham B4 7ET UK
2. Leuven Institute for Research on Information Systems, University of Leuven, Belgium
*Contact author: i.e.holliday@aston.ac.uk
Keywords: statistics education, peer review, reproducible computing, package **RC**

There is increasing impetus towards a reform of statistics teaching in the face of a widespread recognition that traditional approaches to statistics education focussed on computational and analytical skills does not provide students with the ability to apply statistical thinking in real world situations (Garfield et al 2002). Moore (1997) argues for a revised approach requiring *"the abandonment of an 'information transfer' model in favour of a 'constructivist' view of learning:.."* (Moore, 1997). Furthermore, recent research suggests that there are clear benefits of collaborative working involving opportunities for peer review and feedback, particularly for female students who tend to use such opportunities most effectively (Wessa, 2008).

We have developed an undergraduate second year statistics course for psychology students based on principles of constructivist education, whose design is modelled on an established course in a business education setting, aiming to encourage "statistical thinking and literacy". The key software components applied were developed and introduced by Patrick Wessa: 1) the reproducible computing framework available at http://www.freesatistics.org via package **RC**; and, 2) a supporting online peer review (PR) management system (Wessa, 2009). Our presentation illustrates these components from the point of view of the student and the instructor.

Students receive instruction via traditional lectures and supporting workshops, however the workshop material is presented as a 'compendium', an enhanced document form containing all the data and computations necessary to fully reproduce and communicate the results of a statistical analysis. The students' task is to complete workshop assignments and create a new compendium (report) based on their own analysis and interpretation that contain links to their 'blogged' (i.e. archived in the reproducible computations repository) computations; their documents are then uploaded to the PR system and circulated anonymously to students' peers for review. The students' assessed assignment is to provide peer review feedback on up to 5 workshop compendiums they receive each week. The course design provides a social constructivist framework within which independent learning can flourish.

The instructor is able to monitor, evaluate, and control the learning process through a series of statistical reports and accompanying query tools which are available through the RC package and the PR system. The data that is retrieved through these systems can be used for the purpose of educational research based on objectively measured observations. Our analysis of students' performance over two years provides evidence that the combination of RC and PR leads to deep learning of statistical concepts. We briefly present the latest findings to support this conclusion, based on a structural equation model (PLS-PM) and a randomized experiment.

Garfield J; Hogg B, Schau C & Whittinghill D. (2002) "*First Courses in Statistical Science: The Status of Educational Reform Efforts*." Journal of Statistics Education Vol. 10(2)
Moore, DS (1997) "*New Pedagogy and New Content: The Case of Statistics"* International Statistical Review vol. 65 (2), pp123-165
Wessa, P. (2008). "*How Reproducible Research Leads to non-Rote Learning Within a Socially Constructivist e-Learning Environment.*" 7th. Europ. Conf. on E-Learning, Vol 2, 651-658.
Wessa, P. (2009) "*A framework for statistical software development, maintenance, and publishing within an open-access business model."* Computational Statistics, 24(2), 183-193

Teaching applied statistics to students in natural sciences using the R Commander

Kristian Hovde Liland [1,*], Solve Sæbø[1], Thore Egeland[1], Ellen Sandberg[1], Lars Snipen[1], Trygve Almøy[1]

1. The Norwegian University of Life Sciences, Dept. of Chemistry, Biotechnology and Food Science, P.O. Box 5003, N-1432 Ås, Norway
*Contact author: kristian.liland@umb.no

Keywords: R Commander, natural sciences, applied statistics

Statistical software has become an important part of teaching applied statistics on all levels of higher education. The assortment of available applications and packages is formidable, and choosing correctly can be a key to successful learning. At many universities different software has been used for the students that focus on statistics than those that have statistics as an obligatory subject in a study of biology, chemistry or other experimental sciences. Typically software with a graphical user interface and a limited set of statistical methods, e.g. *Minitab* or *SPSS*, is chosen for those who need easy access to basic analysis, while more advanced software with a programming based interface, e.g. *R*, is chosen for those in need of more flexible and adaptable analyses.

With the advent and continual expansion and refinement of *the R Commander* universities do not have to make this choice anymore as everything is available through *R*. At the Norwegian University of Life Sciences we are phasing out the other rigid and expensive programs. Through a local repository, http://repository.umb.no/R, and automatic updates of our own plug-in to *the R Commander*, **RcmdrPlugin.UMB**, we achieve instant distribution of the improvements and additions we make to our statistical toolbox. We continually adapt the plug-in to the courses we teach, tweaking the built in functions of *the R Commander* and adding what we need.

Utilizing the above set-up, students of natural sciences get a soft introduction to computer aided statistical analysis, they get used to seeing and interpreting code, and they use the same base as the students aiming for a statistical education. Teachers and researchers only have to relate to one software package, and they can use their expertise when adapting it to their needs and research. Finally, the institutes save money on expensive licenses.

In addition to this we have a dedicated course for teaching basic use of *R* for master and PhD students in need of customized analysis in their research. This has become very popular.

Automatic generation of exams in *R*

Bettina Grün[1], **Achim Zeileis**[2,⋆]

1. Department of Applied Statistics, Johannes Kepler Universität Linz
2. Department of Statistics, Universität Innsbruck
⋆Contact author: Achim.Zeileis@R-project.org

Keywords: Exams, Multiple Choice, Arithmetic Problems

Package **exams** provides a framework for automatic generation of standardized statistical exams which is especially useful for large-scale exams. To employ the tools, users just need to supply a pool of exercises and a master file controlling the layout of the final PDF document. The exercises are specified in separate Sweave files (containing *R* code for data generation and LaTeX code for problem and solution description) and the master file is a LaTeX document with some additional control commands. An overview is given of the main design aims and principles as well as strategies for adaptation and extension. Hands-on illustrations – based on example exercises and control files provided in the package – are presented to get new users started easily.

References

Grün, B. and A. Zeileis (2009). Automatic generation of exams in R. *Journal of Statistical Software 29*(10), 1–14. http://www.jstatsoft.org/v29/i10/

An S4 Object structure for emulation - the approximation of complex functions

Rachel Oxlade[1], **Peter Craig**[1]

1. Durham University, UK
*Contact author: r.h.oxlade@durham.ac.uk

Keywords: Bayesian, computer model, emulation, nested models, S4 objects

We aim to facilitate the *R* user who has data from a complex function which is costly to evaluate directly (for example a deterministic computer model, see for example Craig et al. (1997), Kennedy and O'Hagan (2001)). We approximate the function by a gaussian process, and thereby produce a posterior distribution for the function's behaviour throughout its domain. Where the function's value is already known, that is at the input points belonging to the data mentioned above, the posterior will predict the correct value with certainty. Elsewhere, the distribution will give an approximation which represents our uncertainty. Typically the domain of the function may cover twenty or more dimensions, and there may be several thousand data points. An object oriented structure is presented for dealing with this problem, enabling the user to interact with the posterior distribution of the function across the domain. This is similar in its goals to the package **BACCO**, but with a different approach.

A particular scenario of interest to us is where we are given a second model (or function), whose inputs extend those of the first. It may contain an additional process, and it is useful for us to be able to compare the two models, to assess the value of including this new process, and to see how the two functions differ across the input domain. We deal with this problem by creating a hierarchical structure, which we again use S4 objects to capture. Joint emulation of models with differing input spaces has, to our knowledge, not yet been addressed, and so this approach is a novel one. The object oriented structure will be presented, along with some methods for dealing with such objects.

In constructing these emulators there are many choices to make, for example whether or not to include a regression surface to capture large scale variation, and if so, how it is to be formed. Incorporated into the *R* code described above are functions which automate such choices given the user's stipulations. This means that, given some initial data showing the function's behaviour and some directions as to how *R* is to proceed at certain points, the functions' behaviour at any set of previously unexplored points can be predicted, and a measure of uncertainty given. The S4 object structure allows the user to interact with the objects, and in particular provides methods to adjust them by making alterations to the original choices made.

References

Craig, P. S., M. Goldstein, A. H. Seheult, and J. A. Smith (1997). Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. *Case Studies in Bayesian Statistics 3*, 37–93.

Kennedy, M. C. and A. O'Hagan (2001). Bayesian Calibration of computer models. *Journal Of The Royal Statistical Society Series B 63*, 425–464.

The structmcmc package: Structural inference of Bayesian networks using MCMC

Robert J. B. Goudie[1]

1. University of Warwick, UK

*Contact author: r.j.b.goudie@warwick.ac.uk

Keywords: Bayesian networks, Graphical models, MCMC, MC^3

I will describe the **structmcmc** package, which implements the widely-used MC^3 algorithm (Madigan et al., 1994), as well as a number of variants of the algorithm. The MC^3 algorithm is a Metropolis-Hastings sampler for which the target distribution is the posterior distribution of Bayesian networks.

The implementation allows the local conditional distributions to be multinomial or Gaussian, using standard priors. Arbitrary structural priors for the Bayesian network can be specified. The main difficulty in sampling Bayesian networks efficiently is ensuring the acyclicity constraint is not violated. The package implements the cycle-checking methods introduced by King and Sagert (2002), which is an alternative to the method introduced by Giudici and Castelo (2003). To enable convergence to be assessed, a number of tools for creating diagnostic plots are included.

Interfaces to a number of other *R* packages for Bayesian networks are available, including **deal** (hill-climbing and heuristic search), **bnlearn** (a number of constraint-based and score-based algorithms) and **pcalg** (PC-algorithm). An interface to **gRain** is also included to allow its probability propagation routines to be used easily.

References

Giudici, P. and R. Castelo (2003). Improving Markov Chain Monte Carlo Model Search for Data Mining . *Machine Learning 50*, 127–158.

King, V. and G. Sagert (2002). A Fully Dynamic Algorithm for Maintaining the Transitive Closure. *Journal of Computer and System Sciences 65*(1), 150–167.

Madigan, D., A. E. Raftery, J. C. York, J. M. Bradshaw, and R. G. Almond (1994). Strategies for Graphical Model Selection. In P. Cheeseman and R. W. Oldford (Eds.), *Selecting Models from Data: AI and Statistics IV*, pp. 91–100. New York: Springer-Verlag.

Computation of (generalized) Nash equilibria

Christophe Dutang[1,⋆]

1. Université de Lyon, Université Lyon 1, Institut de Sciences Financière et d'Assurance, 50 avenue Tony Garnier, 69007 Lyon, France
⋆Contact author: dutangc@gmail.com

Keywords: Generalized Nash equilibrium, Fixed-point method, Regularized gap function, Optimization

When John Nash introduces an equilibrium concept in his seminar paper Nash (1950) for n-player games. Soon after, Debreu (1952) generalized the Nash equilibrium concept for his abstract economy by allowing the strategy space of players to depend on others player actions. From that time, the concept was known as generalized Nash equilibrium (GNE).

Applications of GNE can be found in many fields, e.g. economics, engineering, mathematics, computer science, operational research. However, the computational side of the GNE has not been very studied since his introduction in the 50's, because of its mathematical-economic origin according to Facchinei and Kanzow (2009). Only recently, papers such as Facchinei and Kanzow (2009), von Heusinger and Kanzow (2009) and Nabetani et al. (2009) focus on this topic.

The presentation will focus on computational methods when working with continuous strategy space, i.e. excluding matrix games. In the current litterature, there are three main approaches to solve GNE problem: (i) fixed-point algorithms, (ii) gap function minimization and (iii) extended KKT system solving methods. These approaches are implemented in the *R* package **GNE**[1] by three distinct functions: `fixedpoint`, `minGap` and `NewtonKKT`.

References

Debreu, G. (1952). A social equilibrium existence theorem. *Proc. Nat. Acad. Sci. U.S.A.*.

Facchinei, F. and C. Kanzow (2009). Generalized nash equilibrium problems. Updated version of the 'quaterly journal of operations research' version.

Nabetani, K., P. Tseng, and M. Fukushima (2009). Parametrized variational inequality approaches to generalized nash equilibrium problems with shared constraints. *Computational Optimization and Applications*.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proc. Nat. Acad. Sci. U.S.A. 36*, 48–49.

von Heusinger, A. and C. Kanzow (2009). Optimization reformulations of the generalized nash equilibrium problem using the nikaido-isoda type functions. *Computational Optimization and Applications 43*(3).

[1] available on R-forge.

Segmented regression in thermo-physics modeling

Irina Roslyakova[1*], Holger Dette[2], Mauro Palumbo[1]

1 - Department of Scale Bridging Thermodynamic and Kinetic Simulation, ICAMS, Ruhr-Universität Bochum, UHW 10/1022, Stiepeler Str. 129, 44801 Bochum

2 - Ruhr-Universität Bochum, Fakultät für Mathematik, Mathematik III, Universitätsstraße 150, 44780 Bochum

[*]Contact author: irina.roslyakova@rub.de

Keywords: segmented regression, bootstrap, confidence intervals, thermophysical properties of materials, heat capacity

Traditionally in applied thermo-physics the temperature dependence of the heat capacity is described by polynomials in temperature (from 298K to the melting point) [1], with adjustable parameters fitted to experimental data. The effort to extend that description to low temperature demands more physical modeling which takes into account the recently available theoretical data.

The more physical approach requires the modeling of several contributions (e.g. electronic, vibrational, etc.) that appear in different temperature ranges [2].

In this work we propose the development of statistical segmented models, that can be used as a support tool for CALPHAD[a] modeling [3] and we implement them in *R* [4, 5]. Several segmented regression functions were considered and analyzed, with the corresponding confidence intervals being calculated using the bootstrap method. Moreover we validate the consistency of underlying fitting results, by calculating other physical properties of interest, such as the enthalpy, that can be derived directly from the heat capacity of the studied material.

An overview of available functionalities in the *R* software package for segmented regression [6, 7] will be discussed in solving the nonlinear equations arising from complex model that has to be applied.

References

[1] O. Kubaschewski, C.B. Alcock, P.J. Spencer (1993). Materials thermochemistry.
[2] G. Grimvall (1986). Thermophysical properties of materials.
[3] B. Sundman, H. Lukas, S. Fries (2007). Computational Thermodynamics: The Calphad Method.
[4] G. A. F. Seber, C. J. Wild (1989). Nonlinear Regression.
[5] J. D. Toms, M. L. Lesperance (2003). Piecewise regression: A tool for identifying ecological thresholds. *Ecology 84(8)*, pp.2034-2041
[6] Vito M.R. Muggeo (2010). Segmented relationships in regression models. R-Package: **segmented** (Version 0.2-7.3)
[7] Derek Sonderegger (2010). SiZer: Significant Zero Crossings. R-Package: **SiZer (**Version 0.1-3)

[a] CALPHAD = CALculation of Phase Diagramm

Sparse Bayesian kernel projections for classification of near-infrared spectroscopy data

Katarina Domijan[1⋆]

1. Department of Mathematics and Statistics, NUI Maynooth, Maynooth, Ireland

⋆Contact author: Katarina.Domijan@maths.nuim.ie

Keywords: Bayesian inference, reproducing kernel Hilbert spaces, Bayesian decision theory, NIR spectroscopy

A Bayesian classification method is applied to near infrared spectroscopic data obtained from from several food authenticity studies. The datasets record spectra over the visible and near infra-red wavelength range for relatively small numbers of training samples. Typically the dimension of the variables exceeds ten to twenty - fold the number of samples, however, the reflectance measurements are sequentially highly correlated. The classifier is based on the reproducing kernel Hilbert spaces (RKHS) theory which allows for nonlinear generalization of linear classifiers by implicitly mapping the classification problem into a high dimensional feature space where the data are thought to be linearly separable. The proposed algorithm performs the classification of the projections of the data to the principal axes of the feature space. The projections are uncorrelated and sparse, leading to large computational savings and improved classification performance. The likelihood is modeled through the multinomial logistic regression model and the relatively standard hierarchical prior structure for Bayesian generalized linear models is assumed. The degree of sparsity is regulated in a novel framework based on Bayesian decision theory. The Gibbs sampler is implemented to find the posterior distributions of the parameters, thus probability distributions of prediction can be obtained for new data points, which gives a more complete picture of classification. The classifier obtains good classification rates for the datasets considered and does not require pre-processing steps to reduce their dimensionality. The results from several classification algorithms available in *R* programming language are included for comparison. A collection of *R* functions for classification and data visualization is available.

Recovering Signals and Information From Radio Frequencies Using R (A high school student's experience)

Jinhie Skarda[1*]

1. Montgomery Blair High School, Silver Spring. MD 20901
*Contact author: rideonRF@gmail.com

Keywords: RF signal analysis and demodulation, Time-series analysis, Software oscilloscope and radio

Radio frequency (RF) is part of the electromagnetic spectrum (10 kHz-300 GHz) and is essential to communications and information transfer. For this study, RF measurements were carried out with antennas connected to a PC software oscilloscope, and a digital receiver combined with a software-defined radio tool kit (GNU radio). The *R* language and environment for statistical computing was used to analyze RF signal strengths as well as the information carried on amplitude modulated (AM) radio signals. The oscilloscope was programmed to take ten RF time histories (with 40,000 points over 2 milliseconds) approximately 2 seconds apart. The RF power spectra (up to 10 MHz in frequency) corresponding to the time histories were computed using the `fft` function, and averaged. Variations in RF signal strengths were then studied by taking the ten time history sets every 6 hours for several weeks, and every half hour over a couple of days. The signal strengths of the four AM stations chosen for detailed analysis were found to vary by as much as two orders of magnitude, with strongest signals typically occurring at noon. The information content of multiple AM signals was successfully analyzed using mixing, filtering, decimation, and interpolation routines written in *R* to demodulate a digital signal. The multiple carrier waves were contained in a single RF signal, which was received and stored using the GNU radio and a digital receiver. Where feasible, the results obtained using these routines were compared with the results from standard or specialized *R* packages such as **signal** and **tuneR**. The reduction of undesirable RF signals inside RF-free environments of homemade Faraday cages was also investigated, with data analysis performed in *R*. Faraday cages were effective in reducing RF signal strengths to 1/1000 of the unshielded signals outside the cages. The plotting capabilities of *R* were used for graphical analysis of the RF data in the form of conventional frequency and time history plots, box plots, and contour plots (sometimes referred to as waterfall plots in RF analysis).

References

Adler, Joseph (2010). *R in a Nutshell, A Desktop Quick Reference,* California, O'Reilly.

Greensted, Andrew (2010). FIR Filters by Windowing The Lab Book Pages, http://www.labbookpages.co.uk/audio/firWindowing.html#code/.

Hamming, R. W. (1998). *Digital Filters,* New York, Dover Publications Inc.

Lyons, Richard G. (2004). *Understanding Digital Signal Processing,* New Jersey, Prentice Hall.

Rouphael, Tony J. (2009). *RF and Digital Signal Processing for Software-Defined Radio, A Multi-Standard Multi-Mode Approach,* New York, Elsevier.

Schmitt, Ron (2002). *Electromagnetics Explained, A Handbook for Wireless/RF, EMC, and High-Speed Electronics,* New York, Elsevier.

Spector, Phil (2008). *Data Manipulation with R,* New York, Springer.

Straw, Dean R., Ed. (2007). *The ARRL Antenna Book,* Connecticut, The AARL Inc.

animatoR: dynamic graphics in *R*

Andrej Blejec[1,2,⋆]

1. National Institute of Biology, Ljubljana, Slovenia
2. University of Ljubljana, Department of Biology, Ljubljana, Slovenia
⋆Contact author: andrej.blejec@nib.si

Keywords: Graphics, Dynamic graphics, Teaching

Graphics, especially dynamic graphics, is an impressive tool in various demonstrations. In statistics teaching, there are many situations where graphics with animation of certain elements of the picture communicate the concepts in obvious way.

Since *R* graphic devices are in a sense static, several approaches towards dynamic graphics are used. On many occasions, one would like to move certain graphical element, for example one point, on otherwise static picture. One way is to hide the point by re-plotting it in exclusive OR (XOR) mode and plotting the point in a new position. This method can be fast since one is plotting only the elements that are changing on the otherwise static background which can be very complex. *R* graphic devices are not suitable for such technique. Another way, which is close to this technique, is hiding the dynamic elements by re-plotting them in the background color. This works only for pictures with solid single color background and proves to be unsatisfactory. Another technique is to simply plot a series of complete pictures, each one with relocated picture elements. If the pictures are not very complex, *R* is fast enough (if not too fast) for producing a flicker free dynamic impression. This is the most popular technique, which can provide satisfactory results.

To get an impression of smooth movement, the changes in successive pictures should be small and one needs to get many intermediate point or line positions. Here we provide technique and a set of functions that complement base graphics function for production of dynamic graphics. The basic idea is to define the starting and finishing coordinates of moving picture elements (points, lines, segments, etc.). Then we plot a series of pictures for successive intermediate positions, which are calculated by homotopy between starting and finishing values. If start position is x_0 and end position is x_1 than positions between them can be determined as

$$x_t = x_0(1-t) + x_1 t, \qquad t \in [0,1]$$

for different values of homotopy parameter t. Selection of suitable sequence for homotopy parameter t provides an impression of smooth movement along trajectories from starting to finishing positions.

Package **animatoR** is a developing package utilizing homotopy for production of smooth dynamic graphics, with the motive of presentations and use in teaching. In the presentation, several examples that demonstrate the use of dynamic graphics in statistics teaching will be shown.

Graphical Syntax for Structables and their Mosaic Plots

Erich Neuwirth[1], **Richard M. Heiberger**[2,⋆]

1. University of Vienna
2. Temple University
⋆Contact author: rmh@temple.edu

Keywords: RExcel, Rcmdr, aperm, graphical syntax, mosaic

A `"structable"` object is a representation in the **vcd** package in *R* of a k-dimensional contingency table. The structable object has an attribute `"split_vertical"` that carries two pieces of information: assignment of the factors to row or columns, and sequencing of the factors. The default plot of a structable is as a mosaic plot with recursive splits of the factors in the specified sequence. Each split is along the vertical or horizontal direction associated with the column or row assignment of its factor.

The printed display of a structable as a *flat* table in two dimensions shows the row and column assignment but is unable to illustrate the sequencing of the splits. As a consequence, multiple structables—and their associated mosaic plots—can yield the same printed flat table.

We have developed a graphical notation (with **RExcel** and **Rcmdr** implementations), and corresponding *R* functions with a command-line notation, that simplify the specification of the alternate sequencing of splits—hence alternate mosaic plots—associated with a printed flat table. The notation also permits reassignment of factors from rows to columns. The primary *R* function is an `aperm` (array permutation) method designed for structables. It extends the permutation argument of the default `aperm` method for arrays and tables to include the `"split_vertical"` information that distinguishes the multiple structables associated with the same flat printed table.

RMB: Visualizing categorical data with Relative Multiple Barcharts

Alexander Pilhöfer

1. Alexander Pilhöfer
Department of Computer Oriented Statistics and Data Analysis
Institute of Mathematics
University of Augsburg
86135 Augsburg, Germany
E-mail: alexander.pilhoefer@math.uni-augsburg.de
URL: http://www.rosuda.org

Keywords: categorical data, visualization, multiple barcharts, logistic regression

This talk will present a new graphic for displaying categorical data called "Relative Multiple Barcharts" (`rmb`-plot) and its implementation in the *R* package **extracat**. It is a new attempt to enrich the family of mosaicplots by combining the most important advantages of multiple barcharts (see Hofmann, 2000) and classical mosaicplots (see Friendly, 1994) in one display. The intention of `rmb`-plots is to precisely display relative frequencies of a target variable for each combination of explanatory variables divided over a grid-like graphical display and, simultaneously, their corresponding weights. The breakup of absolute frequencies into conditional distributions and weights is a common procedure in many methodologies for categorical data analysis, such as generalized linear models or correspondence analysis, but even so there seems to be a lack of graphical solutions for exploratory as well as illustrative purposes. After a brief introduction to the concepts of the plot the talk will focus on the implementation in *R*. Using the well-known Copenhagen housing dataset for the examples the talk will first present the main variants of the plot such as a generalized version of spineplots as well as the most important options like color schemes, ceiling censored zooming and residual-based shadings according to Poisson models or logistic regression models. Moreover an interactive version of the graphic based on the **iWidgets** package will be presented which provides interactive controls for all important options as well as a connection to classical mosaicplots in the **vcd** (Hornik et al., 2006) package.

References

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association 89*, 190–200.

Hofmann, H. (2000). Exploring categorical data: Interactive mosaic plots. *Metrika 51*(1), 11–26.

Hornik, K., A. Zeileis, and D. Meyer (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software 17*.

Pilhoefer, A. and A. Unwin (2010). Multiple barcharts for relative frequencies and parallel coordinates plots for categorical data - package extracat. *Journal of Statistical Software*. submitted.

ObsSensitivity: An R package for power analysis for sensitivity analyses of Observational Studies

Neil Diamond[1,*], **Ewa Sztendur**[2]

1. Monash University, Melbourne, Australia
2. Victoria University, Melbourne, Australia
*Contact author: neil.diamond@monash.edu

Keywords: Observational Studies, Power, Sensitivity Analysis

Researchers in the social sciences have been encouraged to use randomised controlled experiments (see, for example, Mosteller and Boruch(2002)). This has been aided by the development of easily used software such as that of Spybook et. al. (2009) which allows researchers to design both person randomised and cluster randomised trials. Despite this, observational studies are still often used. The design of observational studies is just as important as in randomised controlled experiments and "a well designed observational study resembles, as closely as possible, a simple randomized experiment ... [except that in an observational study] ... randomization is not used to assign treatments". (Rosenbaum, 2010, p.4). Many observational studies now use propensity scores to make the treatment and control groups comparable.

Although propensity scores remove the effect of measured covariates, they do not remove bias due to unmeasured variables. An essential aspect of the analysis and reporting of an observational study is to carry out a sensitivity analysis which determines the magnitude of the bias that would be needed to alter the conclusions of the study. The bias is measured in terms of a parameter Γ (Gamma): the odds of receiving the treatment rather than the control given the observed covariates. A value of $\Gamma = 1$ indicates that there is no bias due to unmeasured covariates while a value of $\Gamma = 2$ indicates that an unobserved covariate has the effect of making one of two apparently equal subjects twice as likely to be in the treatment group than the control group.

This talk summarises an R package, **ObsSensitivity**, that assists researchers to determine the appropriate sample size for an observational study. The software tool is provided in R/Excel using the R Commander interface to R, giving the advantages of R but with a convenient and easily learnt environment. Demonstration of the use of the software with examples of actual observational studies will be given.

Mosteller, F. and and Boruch, R., (Eds.), (2002). *Evidence Matters: Randomized Trials in Education Research*, Brookings Institution Press: Washington, DC.

Rosenbaum, P.R (2010). *Design of Observational Studies* Springer: New York.

Spybook, J., Raudenbush, S.W., Congdon, R., and Martínez, A. (2009). Optimal Design for Longitudinal and Multilevel Research: Documentation of the "Optimal Design" Software. http://www.wtgrantfoundation.org/resources/overview/research_tools Downloaded 31 March, 2011.

Visualizing Multilevel Propensity Score Analysis

Jason M. Bryer[1,*]

1. University at Albany
*Contact author: Jason@bryer.org

Keywords: propensity score analysis, multilevel analysis, causal inference, graphics

The use of propensity score analysis (Rosenbaum & Rubin, 1983) has gained increasing popularity for the estimation of causal effects within observational studies. However, its use in situations where data is multilevel, or clustered, is limited (Arpino & Mealli, 2008). This talk will introduce the `multilevelPSA` (Bryer, 2011) package for R that provides functions for estimating propensity scores for large datasets using logistic regression and conditional inference trees. Furthermore, a set of graphical functions that extends the framework of visualizing propensity score analysis introduced by Helmreich and Pruzek (2009) to multilevel analysis will be discussed. An application for estimating the effects of private schools on reading, mathematics, and science outcomes from the Programme for International Student Assessment (PISA; Organization for Economic Co-operation and Development, 2009) is provided.

References

Arpino, B. & Mealli, F. (2008). The specification of the propensity score in multilevel observational studies. *Dondena Working Paper 6*.

Bryer, J. (2011). `multilevelPSA`: Package for estimating and visualizing multilevel propensity score analysis. http://multilevelpsa.r-forge.r-project.org

Helmreich, J & Pruzek, R.M. (2009). `PSAgraphics`: An R package to Support Propensity Score Analysis. *Journal of Statistical Software, 29*, 06.

Organization for Economic Co-operation and Development (2009). *Programme for International Student Assessment*. http://www.pisa.oecd.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrica, 70*, 1, pp. 41-55.

ClustOfVar: an *R* package for the clustering of variables

Marie Chavent[1,2,⋆], **Vanessa Kuentz**[3], **Benoît Liquet**[4], **Jérôme Saracco**[1,2]

1. IMB, University of Bordeaux, France
2. CQFD team, INRIA Bordeaux Sud-Ouest, France
3. CEMAGREF, UR ADBX, France
4. ISPED, University of Bordeaux, France

⋆Contact author: marie.chavent@u-bordeaux2.fr

Keywords: Mixture of quantitative and qualitative variables, hierarchical clustering of variables, k-means clustering of variables, dimension reduction.

Clustering of variables is as a way to arrange variables into homogeneous clusters i.e. groups of variables which are strongly related to each other and thus bring the same information. Clustering of variables can then be useful for dimension reduction and variable selection. Several specific methods have been developed for the clustering of numerical variables. However concerning qualitative variables or mixtures of quantitative and qualitative variables, much less methods have been proposed. The **ClustOfVar** package has then been developped specifically for that purpose. The homogeneity criterion of a cluster is the sum of correlation ratios (for qualitative variables) and squared correlations (for quantitative variables) to a synthetic variable, summarizing "as good as possible" the variables in the cluster. This synthetic variable is the first principal component obtained with the PCAMIX method. Two algorithms for the clustering of variables are proposed: iterative relocation algorithm, ascendant hierarchical clustering. We also propose a bootstrap approach in order to determine suitable numbers of clusters. The proposed methodologies are illustrated on real datasets.

References

Chavent M, Kuentz V., Saracco J. (2009). A Partitioning Method for the clustering of Categorical variables. In *Classification as a Tool for Research*, Hermann Locarek-Junge, Claus Weihs (Eds), Springer, Proceedings of the IFCS'2009, Dresden.

Dhillon, I.S., Marcotte, E.M. and Roshan, U. (3003). Diametrical clustering for identifying anti-correlated gene clusters, *Bioinformatics*, **19**(13), 1612-1619.

Kiers, H.A.L., (1991). Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, **56**, 197–212.

Pagès, J. (2004). Analyse factorielle de données mixtes, *Revue de Statistique Appliquée*, **52**(4), 93-111.

Vigneau, E. and Qannari, E.M., (2003). Clustering of Variables Around Latent Components, *Communications in Statistics - Simulation and Computation*, **32**(4), 1131–1150.

Variable Screening and Parameter Estimation for High-Dimensional Generalized Linear Mixed Models Using ℓ_1-Penalization

Jürg Schelldorfer[1,⋆], **Peter Bühlmann**[1]

[1] Seminar für Statistik, Department of Mathematics, ETH Zurich, CH-8092 Zurich, Switzerland
[⋆] Contact author: schelldorfer@stat.math.ethz.ch

Keywords: generalized linear mixed models, Lasso, high-dimensional data, coordinatewise optimization, variable selection

We propose a two step procedure for dealing with high-dimensional generalized linear mixed models. Generalized linear mixed models (Breslow and Clayton, 1993; Bates, 2009a,b) are straighforward extensions of generalized linear models for clustered observations. In a first step, we perform a Lasso-type (Tibshirani, 1996) variable screening procedure in order to select a relatively small set of covariates. In the second step, we perform ordinary maximum likelihood using only the variables selected in the first step. The latter step is necessary for overcoming bias problems steming from the variable screening step.

In this talk, we present the key ingredients for fitting high-dimensional generalized linear mixed models and demonstrate the performance of the prodecure by presenting the new *R* package **glmmlasso**.

This work is an extension of Schelldorfer et al. (2011) for gaussian linear mixed models and the *R* package **lmmlasso**, which is available from R-Forge (http://lmmlasso.R-forge.R-project.org) and the first author's website (http://stat.ethz.ch/people/schell).

References

Bates, D. M. (2009a). Computational methods for mixed models. Vignette for lme4 on http://www.r-project.org/.

Bates, D. M. (2009b). Linear mixed model implementation in lme4. Vignette for lme4 on http://www.r-project.org/.

Breslow, N. and D. G. Clayton (1993). Approximate inference in generealized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

Schelldorfer, J., P. Bühlmann, and S. van de Geer (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. To appear in the Scandinavian Journal of Statistics, available at http://arxiv.org/abs/1002.3784.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B 58*, 267–288.

gamboostLSS: boosting generalized additive models for location, scale and shape

Benjamin Hofner[1,*]

1. Institut für Medizininformatik, Biometrie und Epidemiologie; Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
*Contact author: benjamin.hofner@imbe.med.uni-erlangen.de

Keywords: GAMLSS, high-dimensional data, prediction inference, spatial models, variable selection

Generalized additive models for location, scale and shape (GAMLSS) model not only the mean but every parameter of the conditional distribution of the outcome (e.g. location, scale and shape) using distinct sets of covariates (Rigby and Stasinopoulos, 2005). We present a boosting algorithm for GAMLSS (*gamboostLSS*; Mayr et al., 2010), which was developed to allow model fitting for potentially high-dimensional data and that overcomes limitations of the original fitting algorithm related to variable selection based on (potentially problematic) information criteria. Furthermore, our approach allows to include a wide variety of possible covariate effects such as linear, smooth, random, or even spatial effects. As the algorithm relies on boosting, estimation of the effects with intrinsic variable selection is possible.

We apply the *gamboostLSS* approach to data of the Munich Rental Guide, which is used by landlords and tenants as a reference for the rent of a flat depending on its characteristics and spatial features. The net-rent predictions resulting from the high-dimensional GAMLSS are highly competitive to classical GAMs while covariate-specific prediction intervals show a major improvement.

A software implementation of the algorithm is available in the *R* package **gamboostLSS** (Hofner et al., 2010).

References

Hofner, B., A. Mayr, N. Fenske, and M. Schmid (2010). *gamboostLSS: Boosting Methods for GAMLSS Models*. R package version 0.5-0.

Mayr, A., N. Fenske, B. Hofner, T. Kneib, and M. Schmid (2010). GAMLSS for high-dimensional data - a flexible approach based on boosting. Technical Report 98, Department of Statistics, Ludwig-Maximilans-Universität München.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics 54*, 507–554.

SCperf: An inventory management package for R

Marlene S. Marchena[1⋆]

1. Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro
⋆Contact author: marchenamarlene@gmail.com

Keywords: Supply chain management, inventory control, EOQ model, safety stock, bullwhip effect.

Supply Chain Management, i.e., the control of the material flow from suppliers of raw material to final customers, is a crucial problem for companies. If appropriately designed and executed, it may offer efficient business solutions, thereby minimizing costs and improving readiness or competitiveness. In this context, the use of mathematical inventory models can give a significant competitive advantage.

We have developed the **SCperf** package as the first package which implements different inventory models that can be used when developing inventory control systems. There are several basic considerations that must be reflected in the inventory model. For instance, models can be divided into deterministic models and stochastic models according to the predictability of demand involved. Our package presents functions to estimate the order quantity and the reorder point regarding different models. Also, other important variables like the safety stock level and the bullwhip effect are calculated. During the presentation, examples will be used to illustrate different inventory situations.

References

[1] Sven Axsäter (2006) Inventory control. Springer, New York. 2nd

[2] Frederick Hillier and Gerald Lieberman (2001). Introduction to operational research. McGraw-Hill. New York, 7th.

[3] Marlene Marchena (2010). The bullwhip effect under a generalized demand process: an R implementation. In Book of Contributed Abstract *useR! 2010, The R User Conference, (Gaithersburg, USA)*, pp. 101.

[4] Marlene Marchena (2011). Measuring and implementing the bullwhip effect under a generalized demand process. PhD Thesis Pontifical Catholic University of Rio de Janeiro, 2011.

[5] R Development Core Team (2010). 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.

Using R to test transaction cost measurement for supply chain relationship: A structural equation model

Pairach Piboonrungroj[1,2,⋆], **Stephen M. Disney**[1]

1. Logistics Systems Dynamics Group, Cardiff Business School, Cardiff University, United Kingdom
2. Chiang Mai School of Economics, Chiang Mai University, Thailand
⋆Contact author: pairach@piboonrungroj.com

Keywords: Transaction cost economics, Supply chain relationship, Structural equation model

Transaction cost economics (TCE) has been widely used to explain the existence and boundary of the firm (Williamson, 2005). Recently, TCE has been extended to inter-firm relationship in supply chains (Hobbs, 1996). However, a measurement of the transaction cost has rarely been tested empirically. Grover and Malhotra (2005) attempted to do so but that measurement has limited to an industrial context and did not cover transaction costs related to the governance problem and the opportunity cost. Thus, we revisited the measurement of transaction cost using both industrial and service perspectives. Moreover, we also considered the transaction cost metric that associated with the governance and the opportunity cost. Then a revised transaction cost metric was tested with empirical data from the tourism supply chains in Thailand using a structural equation model (SEM).

In this study, we used packages in *R*, specifically **sem** (Fox, 2006) and **OpenMx** (Boker et al., 2011) to test the measurement model. Results from *R* were then compared to those from other popular SEM software i.e., AMOS (Arbuckle, 1995) and LISREL (Jöreskog and Sörbom, 1997). An evaluation of SEM using packages in *R* and other softwares is also discussed.

Acknowledgement
The authors are grateful to the Royal Thai Government through the Commission on Higher Education for financial support of Mr. Piboonrungrojs study in Cardiff University (under the program Strategic Scholarships for Frontier Research).

References

Arbuckle, J. L. (1995). *Amos 16.0 User's Guide*. Chicago: SPSS.

Boker, S., M. Neale, H. Maes, M. Wilde, M. Spiegel, T. Brick, J. Spies, R. Estabrook, S. Kenny, T. Bates, P. Mehta, and J. Fox (2011). Openmx: An open source extended structural equation modeling framework. *Psychometrika* (In Press), 1–12.

Fox, J. (2006). Structural equation modeling with the sem package in r. *Structural Equation Modeling 13*(3), 465–486.

Grover, V. and M. K. Malhotra (2005). Transaction cost framework in operations and supply chain management research: theory and measurement. *Journal of Operations Management 21*(4), 457–473.

Hobbs, J. E. (1996). A transaction cost approach to supply chain management. *Supply Chain Management 1*(2), 15–27.

Jöreskog, K. G. and D. Sörbom (1997). *LISREL 8 user's reference guide*. Lincolnwood, IL: Scientific Software International.

Williamson, O. E. (2005). Transaction cost economics and business administration. *Scandivavian Journal of Management 21*(1), 19–40.

Integrating R and Excel for automatic business forecasting

Giovanni Millo[1,⋆], **Fabrizio Ortolani**[1]

1. Assicurazioni Generali S.p.A., R&D Department

⋆Contact author: Giovanni_Millo@generali.com

Keywords: RExcel, Forecasting.

We present a simple exercise in bridging the gap between statistics and everyday business practice, based on two powerful tools already available in the *R* system: the **forecast** package Hyndman (2011) for automatic time series forecasting and the *RExcel* add-in for MS Excel Baier and Neuwirth (2007), allowing to embed *R* functionality into spreadsheets and to interact with their built-in macro language. The application we developed makes forecasting practice accessible to those who are not familiar with statistical programs and, possibly, do not even have a sound statistical background.

Many processes inside the firm involve forecasting. Some build on models and relationships between balance sheet items, but sometimes an a-theoretical extrapolation of past tendencies is needed. As few firms can afford to have trained statisticians dedicated to supply-chain forecasting and the like, budgeting and other activities are often based on simple, heuristic extrapolation of past data. It is commonplace, especially in small enterprises, to ”pick last year/month’s value”, either in terms of stocks or of increments, as the best estimate for the coming period.

Fully automatic forecasting of time series, based on model fitting and model comparing algorhythms selecting the ’best’ model for the data at hand, provides a statistically well founded solution to the forecasting problem and can be of great use to the firm in obtaining accurate predictions for variables like sales, commodities’ input needs and the like, where forecast errors cost money.

Such fully automatic procedures are implemented in a variety of commercial software. We show how an open-source solution is also very easy to set up.

The ideal way is thus to have the forecaster dealing only with Excel for data input, commmand issuing and results’ retrieval, while a ’real’ statistical engine transparently does the computing in the background.

Now the forecaster just has to select the data vector, press the trigger keys for showing up the userform, select the data frequency and press OK. He will get the forecasts at the end of the original series.

References

Baier, T. and E. Neuwirth (2007). Excel :: Com :: R. *Computational Statistics 22*(1), 91–101.

Hyndman, R. J. (2011). *forecast: Forecasting functions for time series*. R package version 2.13.

Using R to quantify the buildup in extent of free exploration in mice

Tal Galili[1*], Yoav Benjamini[1]

1. Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel
*Contact author: Tal.Galili@gmail.com

Keywords: Quantile loess, sequences of repeated motion, dynamics of behavior, open field test, phenotyping mouse behavior

Background: (ref 1) To obtain a perspective on an animal's own functional world, we study its behavior in situations that allow the animal to regulate the growth rate of its behavior and provide us with the opportunity to quantify its moment-by-moment developmental dynamics. Thus, we are able to show that a mouse's exploratory behavior consists of sequences of repeated motion: iterative processes that increase in extent and complexity, whose presumed function is a systematic active management of input acquired during the exploration of a novel environment. We use this study to demonstrate our approach to quantifying behavior: targeting aspects of behavior that are shown to be actively managed by the animal, and using measures that are discriminative across strains and treatments and replicable across laboratories.

The *R* perspective: In our research, *R* was our central tool of choice (ref 1, citation 39). We employed various existing *R* facilities for preparing, analyzing, and visualizing the data. We implemented the `Quantile.loess` algorithm in *R* (the code was published online ref 2) in order to quantify various measurements of the buildup in the mouse's exploration of the ring.

In this talk I will provide the context of this study and present how *R* was used for devising and implementing known and new methods in order to support our investigation.

References

1. Yoav Benjamini, Ehud Fonio, Tal Galili, Gregor Z. Havkin, and Ilan Golani (2011). Quantifying the buildup in extent and complexity of free exploration in mice. *PNAS (Published online before print March 7, 2011)*
2. Tal Galili (2010). R-statistcs blog, http://www.r-statistics.com/2010/04/quantile-loess-combining-a-moving-quantile-window-with-loess-r-function/

Changepoint analysis with the changepoint package in *R*

R. Killick[1,⋆], **I.A. Eckley**[1]

1. Department of Mathematics & Statistics, Lancaster University, Lancaster, U.K.
⋆Contact author: r.killick@lancs.ac.uk

There exists a rich literature exploring changepoint problems which dates back to the 1950s. To date there have been few implementations of changepoint methods within existing *R* packages. At the time of writing, those packages which are currently available implement novel methodology only. As a consequence it is more difficult (i) to compare the performance of new methods with those proposed in the established literature and (ii) for practitioners to test changepoint methods on their data.

The changepoint package attempts to resolve the above by providing both well-established and new methods within a single package. In doing so, we hope to make it easier for practitioners to implement existing methods and for researchers to compare the performance of new approaches against the established literature.. The package has therefore been designed to give users access to many techniques for changepoint analysis within a few easy to use functions. It includes functions that detect changes in mean, variance and regression under various distributional and distribution-free assumptions.

Given the capacity to sample time series at ever higher frequencies the need for computationally efficient search methods is increasing. This is currently an active area of research and so we include four search options in this initial version of the package. The search options which are available are At Most One Change (AMOC), Binary Segmentation (Scott and Knott, 1974), Segment Neighbourhoods (Auger and Lawrence, 1989) and PELT (Killick et al., 2011).

The presentation will be structured as an introduction to changepoint analysis followed by a demonstration of the methods within the **changepoint** package. The methods will be illustrated using datasets from a variety of application areas including genetics, oceanography, finance and linguistics.

References

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54.

Killick, R., Fearnhead, P., and Eckley, I. A. (2011). Optimal detection of changepoints with a linear computational cost. *In Submission*.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512.

Clustering patterns in streamflow to produce regionally or anthropogenically similar groups

Karen R. Ryberg[1,2,⋆]

1. U.S. Geological Survey
2. North Dakota State University
⋆Contact author: kryberg@usgs.gov

Keywords: streamflow, environmetrics, clustering, trends, hydrology

Trends in streamflow characteristics (flood peaks, critical low flows, annual-mean flows) are often discussed in various climate change scenarios, whether that climate change is human induced or the result of natural climatic cycles. Increasing or decreasing trends are important for emergency management, agricultural, and municipal supply, as well as other industrial uses. Trends are often summarized graphically with a few sites highlighted in a paper, for example, or summarized in tables of trend values. However, when looking at a large number of sites with varying drainage basins sizes over a large geographic region, the number of sites involved makes these graphical and tabular methods difficult to comprehend and summarize.

R is used here to calculate a number of streamflow-related variables and extract features from the daily streamflow time series of approximately 500 long-term, mostly unregulated, streamgage sites operated by the U.S. Geological Survey. Cluster analysis is then performed to identify regionally similar areas and to identify sites that may be geographically distant but experiencing similar trends, such as a downward trend in streamflow because of large-scale irrigation which depletes groundwater and baseflow to the streams. A number of cluster variables and techniques (including techniques in **cluster** and **pvclust**) are explored and the results compared to existing knowledge of hydrologic trend and regional differences. The clustering is then visualized geospatially (using **maptools** and **rgdal**) with other hydrologic and geographic data. The clustering and visualization then allows researchers to focus on key sites in the regions (sites with the longest periods of record, largest basins, critical sites where flooding has economic impacts, and so on), as well as identify outliers. Outliers are of interest because they could be sites with unique conditions such as extensive land use change, urbanization, or regulation not indicated in the source database.

Panel on Starting & Building a Local R User Group

Derek McCrae Norton, Atlanta RUG, groups.google.com/group/atlanta-r-users

Szilard Pafka, Los Angeles RUG, www.meetup.com/LAarea-R-usergroup/

Richard Pugh, Mango Solutions, www.mango-solutions.com

David Smith, Revolution Analytics, www.revolutionanalytics.com

In the past two years 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 Derek Norton.

RTextTools

Loren Collingwood,[1,2,*] Tim Jurka[3], Amber Boydstun[3], Emiliano Grossman[4], and Wouter Van Atteveldt[5]

1. Political Science Department, University of Washington
2. Center for Statistics and the Social Sciences, University of Washington
3. Political Science Department, University of California, Davis
4. Centre of European Studies, Sciences Po
5. Communication Science Department, Vrije Universiteit

*Contact author: lorenc2@uw.edu

Keywords: Supervised Learning, Text Analysis, Machine Learning, Social Science

Machine learning has only recently entered the world of social and political science. For years, scholars have used undergraduate research assistants for various classification tasks—such as labeling congressional bill titles, classifying parliamentary speeches, and coding party platforms. Many social scientists are in search of new tools to automate tedious coding tasks, and social scientists have now begun using supervised learning to automate the labeling of documents. *RTextTools* is a recently developed *R* package that provides a uniform interface to several existing *R* algorithms to label text.

From the *tm* package, we include functions to generate document term matrices, stopword removal, and stemming. Currently, the package includes standardized training and classification access to *svm, NaiveBayes, glmnet, randomForest, tree, AdaBoost, Bagging*, and *nnet* algorithms. In addition, we include a C++ maximum entropy train and classification function. We also provide a function for cross validating each algorithm. Finally, several accuracy and ensemble agreement functions are provided to examine how well each algorithm does in terms of predictive accuracy and ensemble agreement. Researchers can quickly identify which text documents are coded with high degrees of accuracy and which documents need to be coded by humans for active learning.

The Role of R in Lab Automation

Jason Waddell[1], **Tobias Verbeke**[1,⋆]

1. OpenAnalytics BVBA
⋆Contact author: tobias.verbeke@openanalytics.eu

Keywords: lab automation, LIMS, R, software integration, R Service Bus

Wet labs of life science companies increasingly make use of advanced measurement technologies that are organized in automated workflows. Many of the steps in these workflows require statistical analyses prior to feeding the results of one step into the next step. Also, the final experiment data require automated reporting to support scientists in the experiment assessment and decision making.

The advantages of using R in this context are well known as it provides companies with cutting edge statistical algorithms that can be directly transferred from Academia to the workbench in industry. In this presentation we will share our experiences in bridging R and lab equipment computers, LIMS systems and end user applications from a variety of settings. We will distill best practices from examples in the analysis of RTqPCR data, EEG signal processing, behavioral experiment data or compound activity screens. The secret weapon to support the diverse workflows and computational requirements is the R Service Bus, an open source integration tool specifically designed to cope with R integration exercises.

OpenAnalytics (2010–2011). The R Service Bus. http://www.openanalytics.eu/r-service-bus.

Using *R* data functions with TIBCO Spotfire

Peter Shaw[1*], Louis Bajuk-Yorgan[1]

1. TIBCO Spotfire Software
*Contact author: pshaw@tibco.com

Keywords: Spotfire, *R*

R scripts and functions can be called from TIBCO Spotfire, and the results returned to the Spotfire environment. Spotfire is a highly dynamic, interactive graphical environment for visualizing data, with filtering and drill-down capabilities. A user can thus manipulate data, drill down and highlight data of interest, and seamlessly send this subset to *R* for analysis, returning the results to Spotfire.

The combination of Spotfire with *R* scripts and functions enables powerful, dynamic analyses. Spotfire provides a convenient means of manipulating data and selecting the data for analysis; *R* extends the capabilities of Spotfire. The results returned by *R* to Spotfire can be any combination of numeric (e.g., model coefficients, forecasts etc), text (e.g., summary diagnostics) or *R* graphical objects.

We will show how to install an *R* script and a CRAN package into Spotfire and demonstrate the resulting R-Spotfire interactive environment, including an example of an *R* graphic returned into Spotfire.

spTimer: Spatio-Temporal Bayesian Modelling using R

Khandoker Shuvo Bakar & Sujit K. Sahu

April 1, 2011

Abstract

Hierarchical Bayesian modelling of large point referenced space-time data are increasingly becoming feasible in many environmental applications due to the recent advances in both statistical methodology and computation power. Bayesian model based analysis methods using the Markov chain Monte Carlo (MCMC) techniques are, however, a very formidable task for large data sets rich in both space and time. Most such analyses are best performed using user written computer code in a low level language that takes hours and sometimes days to run in most personal computers. Currently there does not exist an `R` package which can fit and predict space-time data effectively, although the package `spBayes` (Finley, *et al.*, 2007)can analyse some moderately sized data sets.

This paper develops the `R` package, `spTimer` specifically for modelling spatio-temporal data. This allows us to fit and forecast temporal data over large spatial domains even in mid-range personal computers available today. The package implements a recently developed Bayesian hierarchical autoregressive model (Sahu, *et al.*, 2007) suitable for moderately sized problems. For modelling even larger space-time data, a predictive process approximation method proposed by Banerjee, *et al.*, (2008) is extended, implemented and illustrated with a large data set on ozone monitoring data observed in the eastern United States. mean square error results over other competitive methods currently available in the literature.

Applying geospatial techniques to temporal data

Jason Lessels*, Thomas Bishop, Michael Nelson

The Faculty of Agriculture, Food and Natural Resources. The University of Sydney. Australia.
*Contact author: jason.lessels@sydney.edu.au

Keywords: Water quality, Geospatial, Generalised linear mixed models

Studies involving water quality often focus on load estimation to quantify the characteristics of a catchment, whilst government guidelines often provide concentration threshold targets. Water quality studies often involve the use of unequally spaced temporal data, which prohibits the use of traditional time series methods. Currently simple linear regression models are the most commonly used to estimate concentration levels through time. The use of such models fail to account for the temporal auto-correlation in the observations or provide meaningful results in regards to threshold guidelines. This paper will demonstrate the use of spatial generalised liner (mixed) models to account for the temporal auto-correlation. There are currently two packages within *R* to provide statistically sound methods to improve water quality estimation through time. The package **geoR** provides the ability to use a general linear model with a temporal auto-correlation structure to estimate concentration through time. The second package **geoRglm** provides the ability to use a generalised linear mixed model (GLMM) to estimate the probability of a threshold being exceeded. Both models account for the temporal auto-correlation, where simple linear regression models have failed. In addition the GLMM provides a new method for the estimation of the probability of threshold exceedence through time.

Structured Additive Regression Models: An *R* Interface to BayesX

Nikolaus Umlauf[1,⋆], **Thomas Kneib**[2], **Stefan Lang**[1], **Achim Zeileis**[1]

1. Department of Statistics, Universität Innsbruck, Austria
2. Department of Mathematics, Carl von Ossietzky Universität Oldenburg, Germany
⋆Contact author: Nikolaus.Umlauf@uibk.ac.at

Keywords: MCMC, geoadditive models, mixed models, space-time regression, structured additive regression.

Structured additive regression (STAR) models provide a flexible framework for modeling possible nonlinear effects of covariates: They contain the well established frameworks of generalized linear models (GLM) and generalized additive models (GAM) as special cases but also allow a wider class of effects, e.g., for geographical or spatio-temporal data. This allows for the specification of complex and realistic models that can typically be conveniently estimated using Bayesian inference based on modern Markov chain Monte Carlo (MCMC) simulation techniques or a mixed model representation.

Although there is already a quite extensive existing toolset in *R* supporting GLMs and GAMs, many of the more complex models from the STAR class, especially those utilizing Bayesian inference, are currently not easily available. They are, however, provided in the standalone software package BayesX: a very comprehensive Bayesian semiparametric regression toolbox based on open-source *C++* code. BayesX not only covers models for responses from univariate exponential families, but also models from non-standard regression situations such as models for categorical responses with either ordered or unordered categories, continuous time survival data, or continuous time multi-state models.

Since there has been increasing interest in an *R* interface to BayesX, the already existing CRAN package **BayesX**, which previously provided only functions for exploring estimation results, is now extended to a full interactive interface. With the new version of the package, STAR models can be conveniently specified using *R*'s formula language (with some extended terms), fitted using the BayesX binary, represented in *R* with objects of suitable classes, and finally printed/summarized/plotted.

The talk outlines the usage of the *R* interface to BayesX and its application in complex regression problems, emphasizing its strength in estimating and visualizing models with geographical effects.

References

Brezger, A., T. Kneib, and S. Lang (2005). BayesX: Analyzing Bayesian structural additive regression models. *Journal of Statistical Software 14*(11), 1–22. http://www.jstatsoft.org/v14/i11/

Brezger, A. and S. Lang (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis 50*, 967–991.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica 14*, 731–761.

Kneib, T. (2005). *Mixed model based inference in structured additive regression*. Dr. Hut-Verlag, München, available at http://edoc.ub.uni-muenchen.de/5011/.

The *R* package isocir for Isotonic Inference for Circular Data. Applications to Problems Encountered in Cell Biology.

Sandra Barragán[1,⋆], **Cristina Rueda**[1], **Miguel A. Fernández**[1], **Shyamal D. Peddada**[2]

1. Universidad de Valladolid, Spain.
2. National Institute of Environmental Health Sciences, USA.
⋆Contact author: sandraba@eio.uva.es

Keywords: Circular Data, Isotropic Order, *CIRE*, Conditional Test, R package **isocir**.

Estimation of angular parameters when they are intrinsically ordered around a unit circle is a question of great interest for several researchers. Standard statistical methods, developed for Euclidean parameter space, are not directly applicable for circular data. Particularly, in the presence of restrictions among the parameters, estimators and hypotheses tests of hypotheses have to be properly defined in order to cope with the peculiarities of circular data. Motivated by applications to problems encountered in cell biology, Rueda et al. (2009) introduced the notion of isotropic order and developed a methodology for estimating parameters under this constraint. Given the recent interest among cell biologists in identifying cell cycle genes that are conserved among multiple species, Fernandez et al. (2011) developed a methodology for dealing with isotropic testing problems. The *R* package **isocir** provides a user fiendly software for running all these methods in any context where circular data may appear.

References

Fernandez, M., C. Rueda, and S. Peddada (2011). Isotropic order among core set of orthologs conserved between budding and fission yeast. *Preprint*.

Rueda, C., M. Fernandez, and S. Peddada (2009). Estimation of parameters subject to order restrictions on a circle with application to estimation of phase angles of cell-cycle genes. *Journal of the American Statistical Association 104*(485), 338–347.

CircNNTSR: An *R* Package for the Statistical Analysis of Circular Data Based on Nonnegative Trigonometric Sums

Juan José Fernández-Durán [1,2], María Mercedes Gregorio-Domínguez [2,3]

1. Instituto Tecnológico Autónomo de México, ITAM
2. Department of Statistics and School of Business
3. Department of Actuarial Science

Contact author: jfdez@itam.mx

Keywords: Circular Data, Fourier Series, *R* Software.

In Fernández-Durán (2004), a new family of circular distributions based on nonnegative trigonometric sums (NNTS models) is developed. Contrary to the great majority of families of circular distributions, this family allows to model datasets that present multimodality and/or skewness. Initially, the maximum likelihood estimates of the parameters of the NNTS family were obtained by algorithms based on Sequential Quadratic Programming (SQP) that were difficult to implement in *R* and take a long time to converge. Because the parameter space of the NNTS family is the surface of the hypersphere, an efficient Newton-like algorithm on manifolds is generated in order to obtain the maximum likelihood estimates of the parameters. This algorithm is implemented in the *R* package **CircNNTSR**. Examples of the application of **CircNNTSR** for testing seasonality and homogeneity in problems in biology, actuarial science and environmetrics are presented.

References

[1] Fernández-Durán, J.J. (2004) Circular Distributions Based on Nonnegative Trigonometric Sums. *Biometrics*, 60, pp. 499-503.

[2] Fernández-Durán, J.J. and Gregorio-Domínguez, M.M. (2010) Maximum Likelihood Estimation of Nonnegative Trigonometric Sum Models Using a Newton-like Algorithm on Manifolds. *Electronic Journal of Statistics*, 4: 1402-1410.

[3] Fernández-Durán, J.J. and Gregorio-Domínguez, M.M. (2010b) **CircNNTSR**: An *R* package for the statistical analysis of circular data using nonnegative trigonometric sums (NNTS) models. *R* package version 1.0-1. http://CRAN.R-project.org/package=CircNNTSR

[4] *R* Development Core Team. *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/, 2010.

Summary statistics selection for ABC inference in *R*

Matthew A. Nunes[1,⋆], **David J. Balding**[2]

1. Department of Mathematics & Statistics, Lancaster University, Lancaster, U.K.
2. UCL Genetics Institute, University College London, London, U.K.
⋆Contact author: m.nunes@lancs.ac.uk

Keywords: Approximate Bayesian Computation, entropy minimization, summary statistics

For the purposes of statistical inference, high-dimensional datasets are often summarized using a number of statistics. Due to the intractable likelihoods involved in models for these complex datasets, this is often performed using Approximate Bayesian Computation (ABC), in which parameter inference is achieved by comparing the summaries of an observed dataset to those from simulated datasets under a chosen model (Beaumont et al., 2002).

The question of how best to summarize high-dimensional datasets to maximize potential for inference has been recently addressed in Nunes and Balding (2010). The authors propose two techniques to select a good set of summaries for ABC inference from a collection of possible statistics.

Since high-dimensional datasets arise in many areas of science, these methods could provide practical guidance to scientists on how to represent datasets for optimal inference. The minimum entropy (ME) and two-stage error prediction methods introduced in Nunes and Balding (2010) are implemented for the "rejection-ABC" algorithm (Tavaré et al., 1997) in the **ABCME** *R* package. The package also includes code to perform optional "regression adjustments" for the mean and variance of parameter values which can improve overall quality of inference of ABC algorithms (Fan and Yao, 1998; Yu and Jones, 2004; Beaumont et al., 2002). Some of the computational cost of the ABC algorithm and summary selection procedures is reduced through the use of C routines.

We provide an example of the techniques in the **ABCME** package, demonstrating posterior inference of the parameters from a coalescent model of population DNA sequences (Nordborg, 2007).

References

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate Bayesian Computation in population genetics. *Genetics 162*(4), 2025–2035.

Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85*(3), 645–660.

Nordborg, M. (2007). Coalescent theory. In D. J. Balding, M. J. Bishop, and C. C (Eds.), *Handbook of Statistical Genetics* (3rd ed.)., pp. 179–208. Wiley: Chichester.

Nunes, M. A. and D. J. Balding (2010). On optimal selection of summary statistics for Approximate Bayesian Computation. *Stat. Appl. Genet. Mol. Biol. 9*(1).

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly (1997). Inferring coalescence times from DNA sequence data. *Genetics 145*(2), 505–518.

Yu, K. and M. C. Jones (2004). Likelihood-based local linear estimation of the conditional variance function. *J. Am. Stat. Assoc. 99*(465), 139–144.

Power and minimal sample size for multivariate analysis of microarrays

Maarten van Iterson[1,⋆], José A. Ferreira[2], Judith M. Boer[1,3,4] and Renée X. Menezes[5]

1. Center for Human and Clinical Genetics, LUMC, Leiden
2. RIVM, Bilthoven
3. Department of Pediatrics Oncology and Hematology, Erasmus MC Sophia Childrens Hospital, Rotterdam
4. Netherlands Bioinformatics Center
5. Department of Epidemiology and Biostatistics, VUmc, Amsterdam.

⋆Contact author: M.van_iterson.HG@lumc.nl

Keywords: average power, sample size determination, high-throughput data, kernel deconvolution estimator, moderated t test statistics.

Choosing the appropriate sample size for high-throughput experiments, such as those involving microarrays and next-generation sequencing is complicated. Traditional univariate sample size determinations relate power and significance level to sample size, effect size and sample variability. However, for high-dimensional data these quantities need to be redefined: average power instead of power, significance level needs to take multiple testing into account, and both effect sizes and variances have many values.
Some authors (see Ferreira and Zwinderman, 2006 and Dobbin and Simon, 2005) have proposed such methods for two-group comparisons of high-dimensional data. The most general of those, by Ferreira and Zwinderman, uses the entire set of test statistics from pilot data to estimate the effect size distribution, power and minimal sample size, as opposed to most other published methods that either fix an effect size of interest, or assume (partial) homoscedasticity. Ferreira and Zwinderman assume that the test statistics follow a normal distribution, which is unlikely to hold in practice as many comparisons involve a Student-t test statistic, and the performance of the method in such cases was not evaluated.
We aimed at a generalization of power and sample size estimation more applicable to high-throughput genomics data. First, we extended Ferreira and Zwindermans method to the case of a Student-t test. Second, we considered t-test statistics generated by testing if a coefficient of a general linear model is equal to zero. Furthermore, we considered Student-t tests that use a shrunken variance estimator, such as those produced by empirical Bayes linear models as implemented in the BioConductor package **limma** (Smyth, 2005). These extensions represent a considerable improvement on the power and sample size estimation compared to when the normal assumption is used, which we illustrate via a simulation study. Finally, we will extend the method to generalized linear models aimed at power and sample size estimation for RNA-seq data. The extensions will be implemented as part of our BioConductor package **SSPA** (van Iterson *et al.*, 2009), forming a valuable tool for experimental design of microarray experiments.

References

Dobbin K, Simon R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*. 6(1):27-38.

Ferreira JA, Zwinderman A. (2006). Approximate sample size calculations with microarray data: an illustration. *Statistical Applications in Genetics and Molecular Biology*. 5(25).

Smyth GK (2005). Limma: linear models for microarray data, Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman and V. Carey and S. Dudoit and R. Irizarry and W. Huber, Springer, New York, 397–420.

van Iterson *et al.* (2009). van Iterson M, 't Hoen PA, Pedotti P, Hooiveld GJ, den Dunnen JT, van Ommen GJ, Boer JM, Menezes RX. Relative power and sample size analysis on gene expression profiling data. *BMC Genomics*. 10:439.

Kenward-Roger modification of the F-statistic for some linear mixed models fitted with `lmer`

Ulrich Halekoh[1*], **Søren Højsgaard**[1]

1. Department of Molecular Biology and Genetics, Aarhus University, Denmark
*Contact author: Ulrich Halekoh

Keywords: Kenward-Roger, degrees of freedom, linear mixed models, bootstrap tests

For linear mixed models a frequent task is to test for the reduction of a model by the removal of some fixed effect. For models fitted with the `lmer` function of the *R* package **lme4** (Bates et al., 2011) such tests can be performed by via a maximum likelihood ratio test and the approximate χ^2 distribution of the test statistic.

This approximation may yield rather anti-conservative tests (Pinheiro and Bates, 2000, p. 88) with erroneously small p-values. An alternative is to use F-tests which are exact in some balanced situations. Kenward and Roger (1997) proposed a modification of the F-test statistic possibly shrinking the statistic and adjusting the residual degrees of freedom in order to achieve a better approximation to a F-distribution. Our function implements such an approximation for linear mixed models fitted with `lmer`. The implementation is restricted to covariance structures which can be expressed as a linear combination of known matrices. Such models comprise variance component and random coefficient models.

It has been shown in simulation studies (Kenward and Roger, 1997; Spilke et al., 2005) that sometimes the Kenward-Roger modification represents a satisfactory approximation to the F-distribution. But others have argued (e.g. in a discussion on the **R** help list) that it is generally not clear that such an approximation holds under general circumstances.

We offer therefore additional functions to calculate a p-value via parametric bootstrap and possible improvements by the Bartlett correction of the test statistic.

In the talk we will shortly depict the theory and show some applications also pointing to some numerical problems we encounter.

The functions will be made available via the CRAN package **doBy** (Højsgaard and Halekoh, 2011).

References

Bates, D., M. Maechler, and B. Bolker (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-39.

Højsgaard, S. and U. Halekoh (2011). *doBy: Groupwise summary statistics, general linear contrasts, LSMEANS (least-squares-means), and other utilities*. R package version 4.3.0.

Kenward, M. G. and J. H. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics 53*(3), 983–997.

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

Spilke, J., H.-P. Piepho, and X. Hu (2005). A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data. *Journal of Agricultural, Biological, and Environmental Statistics 10*(3), 374–389.

lqmm: Estimating Quantile Regression Models for Independent and Hierarchical Data with R

Marco Geraci[1,⋆]

1. MRC Centre of Epidemiology for Child Health, UCL Institute of Child Health, London, UK
⋆Contact author: m.geraci@ucl.ac.uk

Keywords: Quantile regression, random effects, longitudinal data, hierarchical models

Conditional quantile regression (QR) pertains to the estimation of unknown quantiles of an outcome as a function of a set of covariates and a vector of fixed regression coefficients. In the last few years, the need for extending the capabilities of QR for independent data to deal with clustered sampling designs (e.g., repeated measures) has led to several and quite distinct approaches. Here, I consider the likelihood–based approach that hinges on the strict relationship between the weighted L_1 norm problem associated with a conditional QR model and the asymmetric Laplace distribution (Geraci and Bottai, 2007).

In this presentation, I will illustrate the use of the *R* package **lqmm** to perform QR with mixed (fixed and random) effects for a two–level nested model. The estimation of the fixed regression coefficients and of the random effects' covariance matrix is based on a combination of Gaussian quadrature approximations and optimization algorithms. The former include Gauss-Hermite and Gauss-Laguerre quadratures for, respectively, normal and double–exponential (i.e., symmetric Laplace) random effects; the latter include a modified compass search algorithm and general purpose optimizers (`optim` and `optimize`). Modelling and inferential issues are detailed in Geraci and Bottai (2011) (a preliminary draft is available upon request). The package also provides commands for the case of independent data.

References

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics 8(1)*, 140–154.

Geraci, M. and Bottai, M. (2011). Linear quantile mixed models, in preparation.

lcmm: an *R* package for estimation of latent class mixed models and joint latent class models

Cécile Proust-Lima[1,2,⋆], **Benoit Liquet**[1,2]

1. INSERM U897, Bordeaux, France

2. ISPED, Bordeaux Segalen University, France

⋆Contact author: cecile.proust@isped.u-bordeaux2.fr

Keywords: classification, joint models, mixed models, growth mixture model, longitudinal data, non-Gaussian data, time-to-event data.

The linear mixed model is routinely used to describe change over time of a quantitative outcome in longitudinal studies. However, it is limited to the analysis of a quantitative Gaussian outcome, in a homogeneous population, and does not handle association with a time-to-event although it is frequent in practice. The *R* package **lcmm** extends the linear mixed model to (1) the study of heterogeneous populations through the estimation of latent class mixed model, and (2) the joint analysis of longitudinal and time-to-event data through the estimation of joint latent class models. In each case, both Gaussian or non Gaussian quantitative and ordinal outcomes can be analysed.

Latent class mixed models consist in exploring the latent profiles of trajectories in heterogeneous population. They combine the mixed models theory to account for the individual correlation in repeated measures, and the latent class models theory to discriminate homogeneous latent groups when modelling trajectories. Despite a large interest in this approach also known as growth mixture models, implementation in free softwares is very limited. Within the **lcmm** package, the function `hlme` estimates latent class mixed models assuming a quantitative Gaussian outcome and `lcmm` extends this approach to handle non Gaussian quantitative and ordinal outcomes that are very frequent especially in psychological and quality of life studies.

Joint models to analyse jointly longitudinal and time-to-event data have also become increasingly popular in statistics. There exist two kinds of joint models, shared random-effect models in which functions of the random-effects from mixed model are included in the survival model, and joint latent class models which assume the population is constituted of latent classes with a specific longitudinal outcome trajectory and a specific risk of event. While an *R* package **JM** was recently developed to estimate shared random-effect models, no free software exists for joint latent class models. We propose in **lcmm** the function `jointlcmm` to estimate such joint models, both for Gaussian or non Gaussian quantitative and ordinal outcomes.

Whatever the models, estimation in **lcmm** is performed using a maximum likelihood method using a modified Marquardt algorithm with strict convergence criteria. Only for ordinal outcomes, a numerical integration is required. `hlme`, `lcmm` and `jointlcmm` allow any shape of trajectory, and covariates as predictors of the latent class structure as well as of class-specific trajectories. The joint model includes different baseline functions and common or class-specific covariate effects. Finally, posterior classification and goodness-of-fit measures are provided within post estimation functions.

References

Proust, Jacqmin-Gadda (2005). Estimation of linear mixed models with a mixture of distribution for the random-effects. *Computer Methods and Programs in Biomedicine*,**78**,165-73

Proust-Lima, Joly et al. (2009). Joint modelling of multivariate longitudinal outcomes & a time-to-event: a nonlinear latent class approach, *CSDA*, **53**, 1142-54

Proust-Lima & Taylor (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach, *Biostatistics*, **10**, 535-49

Mixed-effects Maximum Likelihood Difference Scaling

Kenneth Knoblauch[1⋆], **Laurence T. Maloney**[2]

1. INSERM, U846, Stem Cell and Brain Research Institute, Dept. Integrative Neurosciences, Bron, France
2. Psychology Department & Center for Neurosciences, New York University, New York, NY, USA
⋆Contact author: ken.knoblauch@inserm.fr

Keywords: psychophysics, difference-scaling, glm, mixed-effects

Difference scaling is a psychophysical procedure used to estimate interval, perceptual scales along a stimulus continuum. On each of a set of trials, an observer is presented with a quadruple, (I_a, I_b, I_c, I_d), drawn from an ordered set of stimuli, $\{I_1 < I_2 < \ldots < I_p\}$. The observer judges which pair shows the greater perceptual difference. Alternatively, a trial can consist of a triple, (I_a, I_b, I_c) and the observer judges whether the difference between (I_a, I_b) or (I_b, I_c) appears greatest. The fitting procedure estimates perceptual scale values, $\{\psi_1, \psi_2, \ldots, \psi_p\}$, that best capture the observer's judgments of the perceptual differences between the stimuli. The perceived difference between a pair of stimuli, (I_a, I_b), is modeled as the difference of perceptual scale values, $L_{ab} = \psi_b - \psi_a$. In comparing, pairs of pairs, we assume that the observer forms the noise-contaminated decision variable, $\Delta = L_{ab} - L_{cd} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$ and judges the difference between the pair (I_a, I_b) to be greater than that between (I_c, I_d) precisely when $\Delta > 0$. The parameter σ represents the observer's precision in judgment. We estimate σ and the remaining free parameters, $\psi_2, \ldots, \psi_p$, either by directly maximizing the likelihood, using the `optim` function or, because the decision rule is linear, via a generalized linear model, using `glm`. The package **MLDS** on CRAN implements these two approaches.

We extend these analyses to mixed-effects models, using `glmer` and `lmer` in the **lme4** package. This allows us to incorporate variability of observers and other sources as random effects, The design matrix constructed for `glm`, however, distributes the stimulus levels over different columns, somewhat like a factor variable, making it difficult to treat them as a single entity. We consider two strategies to deal with this situation using `glmer` and one using `lmer`.

1. redefining the decision variable in terms of a parametric function that describes well the perceptual scale and using this decision variable as a regressor in the model formula,
2. redefining the decision variable in terms of an observer's own estimated perceptual scale and using this decision variable as a regressor in the model formula,
3. fitting the estimated perceptual scale values as a linear mixed-effects model, using `lmer`.

We demonstrate the three approaches with example data sets and discuss their pros and cons.

References

Bates, D., M. Maechler, and B. Bolker (2011). **lme4***: Linear mixed-effects models using S4 classes*. R package version 0.999375-39/r1298.

Knoblauch, K. and L. T. Maloney (2008). **MLDS**: Maximum likelihood difference scaling in *R*. *Journal of Statistical Software 25*, 1–26.

Maloney, L. T. and J. N. Yang (2003). Maximum likelihood difference scaling. *Journal of Vision 3*(8), 573–585.

Schneider, B., S. Parker, and D. Stein (1974). The measurement of loudness using direct comparisons of sensory intervals. *Journal of Mathematical Psychology 11*, 259–273.

Tricks and Traps for Young Players

Ray Brownrigg[1*]

1. School of Engineering and Computer Science, Victoria University of Wellington, NZ
*Contact author: Ray.Brownrigg@ecs.vuw.ac.nz

Keywords: programming, performance, vectorisation

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

Issues to be discussed include:

- Vectorisation
  - user-defined functions (using `curve`, `optimise`)
  - pseudo vectorisation
  - multi-dimensional vectorisation
- `sort`, `order` and `rank`
- When to use Fortran
- Local versions of standard functions
- Using a matrix to index an array
- Resolution of pdf graphs
- `get`
- `file.choose`

Note: This paper was presented at UseR'08 in Dortmund. It seemed to be very popular, so I feel it is worth presenting the same material again.

Software design patterns in R

Friedrich Schuster[1,*]

1. HMS Analytical Software GmbH, Heidelberg, Germany
*Contact author: friedrich.schuster@analytical-software.de

Keywords: Software engineering, Design patterns, Software quality, Development process

The long-term development of R-programs and packages not only leads to a growing code base, but also to increased code complexity and stronger interdependencies between packages and functions. As a result the required training times as well as the development and maintenance costs tend to rise substantially, especially for large applications. In a corporate environment, software quality and maintenance are a major cost factor. One answer to this challenge of software engineering are "Software Design Patterns". Design patterns are reusable general solutions to commonly occurring problems in software design. Design patterns assist the software development process in different ways: (1) by providing tested and proven development paradigms, patterns speed up development and reduce errors. (2) Patterns facilitate the familiarization with unknown code. (3) Patterns improve communication between developers. This presentation will focus on well-known development patterns and their applicability to R and on the identification of specific patterns for R.

References

Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides (1994). Design Patterns. Elements of Reusable Object-Oriented Software. Addison-Wesley.

Random input testing with *R*

Patrick Burns[1,*]

1. Burns Statistics
*Contact author: patrick@burns-stat.com

Keywords: Software testing, Random generation

Traditional software testing uses specific inputs and then sees if the results are correct. This is necessary but not always sufficient to have faith that the software operates properly. When the number of inputs is large, the combinatorial explosion means that full coverage is impossible.

An alternative form of testing is to create random inputs and then infer the suitability of the result. This means many more combinations can be tested, and in particular avoids bias that may be in the traditional test suite. Another advantage of this type of testing is that it exercises the code throwing errors and warnings, which is seldom the case for traditional tests.

R is an excellent environment both for generating random inputs, and for examining the resulting output. We'll highlight a specific example of portfolio optimization.

An Open Source Visual R Debugger in StatET

Stephan Wahlbrink[1], **Tobias Verbeke**[2,⋆]

1. WalWare / wahlbrink.eu
2. OpenAnalytics BVBA
⋆Contact author: tobias.verbeke@openanalytics.eu

Keywords: IDE, debugger, developer tools, R

The StatET plug-ins turn Eclipse into a first class integrated development environment for *R* programmers and data analysts. It offers a set of mature tools for *R* coding and package building and provides a fully integrated R Console, Object Browser and R Help System. In this paper we will present a visual debugger for R that has been seamlessly implemented using the Eclipse debugger framework and allows developers to visually step through the code and inspect variables and expressions in dedicated Eclipse views. We will provide a detailed overview of all debugging functionality that has been implemented (over the last year) and (convincingly) demonstrate the added value a visual debugger can bring to the R development process.

WalWare et al (2004 – 2011). Eclipse Plug-In for R: StatET. http://www.walware.de/goto/statet/.

Predicting the offender's age

Stephan Stahlschmidt[1,2,⋆]

1. Humboldt-Universität zu Berlin, School of Business and Economics, Chair of Statistics, Spandauer Straße 1, 10178 Berlin, Germany
2. Ruprecht-Karls-Universität Heidelberg, Centre for Social Investment, Adenauerplatz 1, 69115 Heidelberg, Germany
⋆Contact author: stahlschmidt@wiwi.hu-berlin.de

Keywords: Offender Profiling, Prediction Power, Statistical Learning

Predicting the age of an unknown offender is notoriously difficult for police profilers. These experts rely solely on traces found on the crime scene and combine this information to deduce the exact events which happened on the crime scene. Only afterwards they try to characterise the offender based upon the unassured knowledge drawn from the crime scene. These statements on the offender are often based on asumptions and lack certitude. However, an assured prediction on, for example, the offender's age would prove very useful to narrow the group of potential suspects.

We therefore compare the performance of several prediction techniques on a data set of sex-related homicides. The applied techniques include linear regression, k-Nearest-Neighbour, regression trees, Random Forest and Support Vector Machine. We evaluate each approach by its prediction power concerning the offender's age and account for the specific requirements of forensic data, namely the restricted access to information via the crime scene and uncertainty of available information. Our results show, if and how the police's investigation could be enhanced by implementing one of these well-known prediction techniques.

References

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33, 1-22.

Hornik, K. , Buchta, C. and Zeileis, A. (2009). Open-Source Machine Learning: R Meets Weka. *Computational Statistics*, 24, 225-232.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods. *Journal of Statistical Software*, 11, 1-20.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2, 18–22.

Leveraging Online Social Network Data and External Data Sources to Predict Personality

Daniel Chapsky*

*Contact author: danchapsky@gmail.com

Keywords: Bayesian Networks, Personality, Facebook

People express their personalities through online social networks in a variety of ways, such as their relationships with their friends and their listed interests. In this work I present a method for automatically predicting an individual's personality by combining his Facebook profile information with external data sources using a machine learning method known as a Bayesian Network. The developed models use representations of people's connections to other people, places, cultures, and ideas, as expressed through Facebook. Due to the nature of Bayesian Networks, the semantics underlying the models are clear enough to not only predict personality, but also use knowledge of one's personality to predict his behavioral attributes and actions. I will present some of the more interesting models of personality that my systems have produced thus far. These models demonstrate the potential of my methodology in two ways: First, they are able to explain up to 70% of all variation in a personality trait from a sample of 615 individuals. Second, they are able to clearly describe underlying relationships in the model through findings such as how to predict a man's agreeableness based on his age, hometown, number of Facebook wall posts, and his willingness to disclose his preference for music made by Lady Gaga.

I will also present the necessary background in Bayesian Networks and Personality theory to understand the above results, and present how all data collection and modeling was automated using an academic edition of REvolution R Enterprise and various R packages.

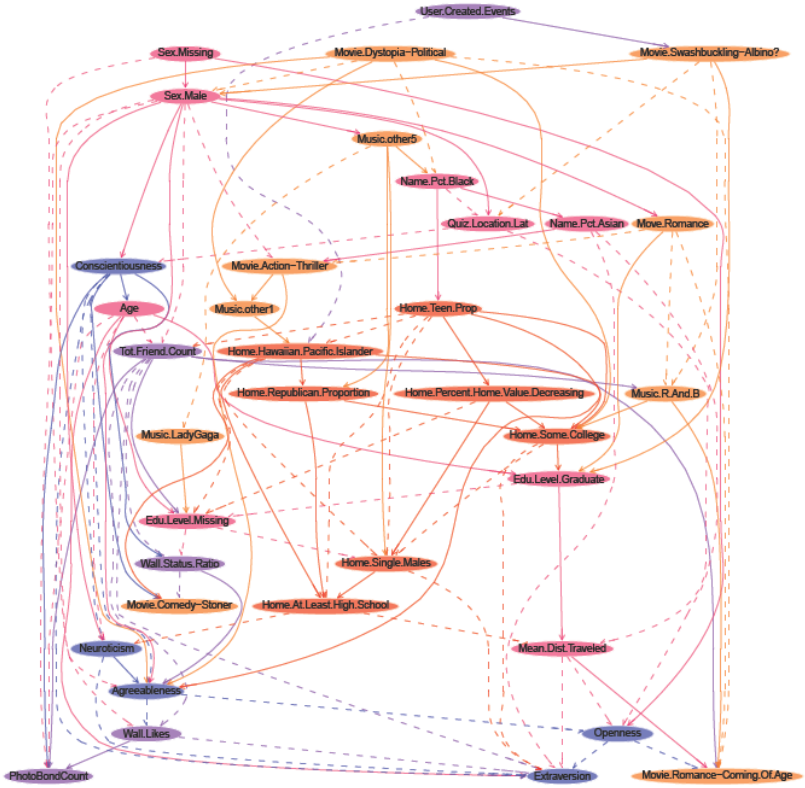


Figure 1: An Example Bayesian Network modeling personality. Blue nodes are personality traits. Dotted lines denote a positive relationship while straight lines denote a negative one.

Using R to Model Click-Stream Data to Understand Users' Path To Conversion

Douglas Galagate[1*,2] , Professor Wolfgang Jank[1,3]

1. University of Maryland, College Park
2. Department of Mathematics
3. Robert H. Smith School of Business
*Contact author: galagate@math.umd.edu

Keywords: Click-stream data, data mining, advertising, internet

Advertisers spend a lot of time and effort directing consumers' attention to their client's website, enticing them to purchase a product or a service. While most models attribute a consumer's conversion only to the last site they visited ("last click model"), it is generally accepted that all of the information in a consumer's search path (i.e. in the sequence of all websites visited) play some role in the ultimate purchase decision. In this project, we use data mining techniques to characterize and model a consumer's "path to conversion." That is, we characterize the network of websites and sequence of clicks in order to gauge the impact of different types of online content, its order and its relationship to one another on the probability of a consumer's conversion. We use many *R* packages in the analysis.

References

Agresti, Alan (2002). Categorical Data Analysis
Wickham, Hadley (2009). ggplot2: Elegant Graphics for Data Analysis (Use R)
Liu, Bing (2010). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)

Packaging *R* for Ubuntu: Recent Changes and Future Opportunities

Michael A. Rutter[1,⋆]

1. Department of Mathematics, Penn State Erie, The Behrend College
⋆Contact author: marutter@gmail.com

Keywords: *R*, CRAN, Packages, Ubuntu

Ubuntu is a GNU/Linux distribution based on Debian Linux that has become popular on desktops, servers, and in cloud-based environments. In addition to being free and open source, one of the important features of Ubuntu is the packaging system. The packaging system allows users and administrators to easily install and update software packages while automatically handling the software dependencies that may be needed in order for the software to run.

CRAN currently has binary packages of the latest versions of *R* as well as the recommended *R* packages in both 32-bit (i386) and 64-bit (amd64) versions for a number of Ubuntu releases, including the latest two long term service (LTS) releases. Due to changes in hardware availability, the building process for these packages has changed. In this presentation, I will detail how the build process has moved from personal servers to Launchpad, a suite of tools provided by Ubuntu. Launchpad includes the ability to build binary packages for a variety of different architectures and Ubuntu releases. These builds are then synced to CRAN, allowing them to be mirrored on various servers across the world.

However, even with the strong packaging system of Ubuntu, only a small number of the 2,000 plus *R* packages are available via CRAN or the main Ubuntu repositories. Installing additional packages can be done within *R*, but this sometimes requires tracking down the required libraries in Ubuntu. In addition, once these packages are installed they are not automatically updated during Ubuntu system updates, which may cause users to miss bug fixes or feature updates. In order to make more packages available to *R* users via Ubuntu, I will also discuss the possibility of using Launchpad as the back-end of an automated package building system for Ubuntu. Using components of cran2deb, developed by Charles Blundell and Dirk Eddelbuettel, it should be possible to create the source packages on a single server and allow Launchpad do to the heavy lifting of building the packages for multiple architectures. Initially, the number of packages available will be limited to some of the more popular packages (e.g. **ggplot2**) and those packages used in popular *R* references. While this type of resource will be helpful for users new to *R* and Ubuntu, an additional goal is to provide the packages needed to easily deploy *R* in cloud-based Ubuntu environments.

References

cran2deb website: *http://debian.cran.r-project.org/*

Launchpad website: *https://launchpad.net/*

Interpreter Internals: Unearthing Buried Treasure with CXXR

Andrew Runnalls[1]

1. School of Computing, University of Kent, UK, A.R.Runnalls@kent.ac.uk

Keywords: *R*, CXXR, *C++*, packages, bignum

CXXR (www.cs.kent.ac.uk/projects/cxxr) is a project to reengineer the interpreter of the *R* language, currently written for the most part in *C*, into *C++*, whilst as far as possible retaining the full functionality of the standard *R* environment, including compatibility with existing *R* packages (cf. Runnalls, 2010). It is intended 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 open up new avenues of development for the *R* project. 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 2.12.1.

From the beginning, different kinds of *R* object have been represented within the CXXR interpreter using a C++ class hierarchy, with the intention that this hierarchy could be extended by *R* developers. This reflects two strategic objectives:

- Given a *C++* class (provided for example by a third-party library) it should be straightforward to introduce an *R* (CXXR) class wrapper around it, so that the functionality of that class can be exploited within the *R* language framework.
- Conversely, given an existing *R* class (provided typically by an *R* package), it should be straightforward to migrate its functionality—to whatever extent desired—into a new underlying *C++* class.

Unfortunately, much C code within the standard *R* interpreter that could greatly support these objectives is not offered to packages as an API (Application Program Interface), and is not documented to API standards. Moreover, in large measure this code is hard-wired around the built-in *R* data types (`REALSXP` etc.), and so cannot easily be applied to other *C++* data types.

This paper will explain how CXXR is endeavouring to change this situation by rewriting key pieces of interpreter code at a higher level of abstraction (often using *C++* templates), and making it available to package writers via carefully documented interfaces.

Current progress will be illustrated by considering how to implement a package allowing the use of 'BigInts' (integers of arbitrarily large magnitude) within R. This is a task that has already been ably accomplished, building on the standard *R* interpreter, by Lucas *et al.* as part of their *gmp* package, which utilises GMP, the GNU Multiple Precision Arithmetic Library. But the paper will show how the CXXR framework makes it very much easier to establish key elements of the necessary functionality.

References

Andrew Runnalls (2010). CXXR and Add-on Packages. http://user2010.org/slides/Runnalls.pdf. Presented at useR! 2010, Gaithersburg, MD.

Antoine Lucas, Immanuel Scholz, Rainer Boehme and Sylvain Jasson (2010). Package 'gmp'. http://cran.r-project.org/web/packages/gmp/index.html.

Torbjörn Granlund *et al.* (2010). The GNU Multiple Precision Arithmetic Library. http://gmplib.org/

R's Participation in the Google Summer of Code 2011

Claudia Beleites[1,2,⋆], **. . . , and John C. Nash**[3]

1. CENMAT and Dept. of Industrial and Information Engineering, University of Trieste, Trieste/Italy
2. Institute of Photonic Technology, Jena/Germany
3.Telfer School of Mangement, University of Ottawa, Ottawa/Canada
⋆Contact author: cbeleites@units.it

Keywords: R packages, Google Summer of Code

We plan to present this year's Google Summer of Code projects that take place under the umbrella of The R Project for Statistical Computing. The work will start end of May and finish in August (suggested "pencils down" at August 16$^{\text{th}}$). Thus the UseR! 2011 conference will be a good occasion to present the brand new tools and packages and also to make the google summer of code more widely known to the user community of R.

Clearly the deadline for abstracts for the UseR! 2011 conference is prior to the final assignment of the students (April 25$^{\text{th}}$), so this "dummy" abstract will be expanded with a more detailed one on the GSoC-R projects page once we have active projects.

Converting a spatial network to a graph in R

Binbin Lu, Martin Charlton
National Centre for Geocomputation, National University of Ireland, Maynooth
*Contact author: binbin.lu.2009@nuim.ie

Keywords: spatial network, graph, igraph, shp2graph

Spatial networks have been an important subject of many studies in quantitative geography and sociology (Barthelemy, 2010). In practical applications, they are traditionally managed in a geospatial vector format, in which spatial entities are recorded as polylines that are spaghetti collections of 2D/3D geospatial coordinates (George & Shekhar, 2008). On the other hand, a network can be regarded as a set of nodes, occupying particular positions and joined in pairs by physical or ephemeral constructs (Gastner & Newman, 2006). In this sense, networks are visualized in a link-node mode, termed as graph data model. Due to its great advantages in simplifying the representation of a network and facilitating related computations powered by graph theory, this model has been widely used as one of the central tools in spatial analysis (Hostis, 2007). Based on the **R** platform, a considerable number of packages have emerged for processing both kinds of objects. On the R-spatial website (Bivand, 2007), numerous packages are available for processing spatial data, especially like **sp** and **maptools**. As for graph, there are two popular packages to create and manipulate undirected and directed graphs, i.e. **graph** and **igraph**. However, there is no such thing to convert a *spatial* object to a *graph-class* object. This work aims to fill this gap.

Overall, all the functions are encapsulated in an **R** package named **shp2graph**. Its main utility is to convert a *SpatialLines* or *SpatialLinesDataFrame* object to a *graphNEL*, *graphAM* or *igraph* object. The principle is to abstract geospatial details, select two endpoints of each *spatialline* as spatially enabled nodes and then link them with an edge. Furthermore, the following issues have been provided for this conversion:

- Attribute heritage. Properties are guaranteed to inherit properly.
- Structure optimization. As for the potential redundant info, like self-loops, multiple-edges and pseudo-nodes, three optimization functions have been designed to remove them.
- Topology. Identified from the converting principle, topology errors could cause serious connectivity problems, and correct topology is the key to a successful conversion. Consequently, a self-test function of topology has been included in this package.
- Integration of data points and network. When we are dealing with a set of data points together with a spatial network, they may not be closely related. In this case, how to integrate these points into the network is concerned according to different accuracy requirements.

References

Barthelemy M (2010). Spatial Networks, http://arxiv.org/abs/1010.0302v2
George B, Shekhar S (2008). Road Maps, Digital. *In: Shashi S and Hui X (Ed) Encyclopedia of GIS*, Springer-Verlag. New York.
Gastner MT, Newman MEJ (2006). The spatial structure of networks, *The European Physical Journal B* 29, 247-252
Hostis AL (2007). Graph theory and representation of distances: chronomaps, and other representations. *In: Mathis P(Ed.), Graphs and Networks: Multilevel Modeling*, Wiley-ISTE, London.
Bivand R(2007), Spatial data in R, http://r-spatial.sourceforge.net/

Spatial modelling with the *R*–GRASS Interface

Rainer M Krug[1,⋆]

1. Centre of Excellence for Invasion Biology, Stellenbosch University, South Africa
⋆Contact author: Rainer@krugs.de

Keywords: Ecological modelling, spatial, monte carlo simulation, GRASS, GIS

R is a widely used tool for analysis and simulation of non-spatial data, even considered by many as the "lingua franca" for statistical modelling. But for working with spatial data, the toolbox provided by *R* does not consist of all the functionality to do complex analysis in *R* directly, unless one is prepared to spend a substantial amount of time with programming routines. This problem extends to the storage of spatial data, where, especially when dealing with large areas, the data sets can become extremely large and one is quickly confronted with memory problems. As a solution to these problems, interfaces with different GIS programs are provided for in R. These interfaces (for linking to GIS **spgrass6**, **rsaga**, importing and exporting **rgdal**, **sqlitemap** and others, but also **spgrass6**, and **rsaga**) provide functions to import the data into R and export it again into a format readable by the GIS, but also to utilise the functions of the GIS applications from within *R*, i. e. to use *R* as a scripting language for those GIS applications. This scripting becomes an important tool in the analysis and simulation of spatial data, as large maps can be manipulated without reading them into *R*, instead using functions of GIS programs which were written to accommodate large data sets (there is a package in *R* (**raster**) which uses the same approach, i. e. only loading smaller sections of the whole map into memory and processing those, but it obviously does not have all the functions included in dedicated GIS applications).

As elegant the solution is of using *R* as a scripting language to combine the power of *R* with the power of GIS applications like GRASS and SAGA, it has still problems. These, however, show only when these interfaces are used intensively.

To illustrate my points and to highlight a few areas which the *R*–GRASS interface could be improved, I will present a study conducted to investigate the impact of different budgets and management scenarios on the effectiveness of alien plant clearing. The study was conducted using a spatio-temporal explicit simulation model, covering areas of up to 215.000 hectares with a cell size of one hectare. The simulation used several vector and raster input layers, and generated hundreds of raster output layers. The interface between *R* and GRASS (**spgrass6**) was used intensively, as certain processes were not available in *R* or the implementation in GRASS was considerably faster due to implementation in *C*.

Main aspects identified during the project as "could be improved" range from direct reading of GRASS data without the need of additional software and simple implementation of processing with a MASK to the general question of backends for spatial data to parallisation especially of spatial routines.

sos4R - Accessing Sensor Web Data from R

Daniel Nüst[1,2,⋆]

1. 52°North Initiative for Geospatial Open Source Software GmbH
2. Institute for Geoinformatics, University of Muenster, Germany
⋆Contact author: d.nuest@52north.org

Keywords: Sensor Observation Service, Open Geospatial Consortium, Web-service Client, XML, Reproducible Research

The Sensor Web is a collection of standards and software solutions that allows the management of and data retrieval from sensors through the internet. The Open Geospatial Consortium's (OGC) Sensor Observation Service (SOS) is the core web service to provide sensor data in an interoperable, standardized way. Because its goals are so generic, this service is relatively complex, e.g. it needs other standards for request and query markup, and standards for data markup and encoding. This complexity makes it hard for non-Sensor Web specialists to benefit from publishing their data using SOSs, or analyzing data retrieved from other's or even someone's own SOS instances. The recently published **sos4R** CRAN package tries to overcome this hurdle by providing a relatively simple *R* interface to the SOS. The main contribution of the package is a set of *R* functions for the core SOS operations for data retrieval. These suffice for many common use cases by encapsulating the complexity of the SOS interface.

This work presents the technical background of accessing a XML-based web service from R using the packages **RCurl** and **XML**. It shows how the generic work flow of request building, encoding, data transfer and data decoding can be modelled in *R* classes and methods. Flexible mechanisms allow users to easily add features and adapt entire processing steps, or just the required parts, to their needs.

We also present example analyses based on publicly available SOS that illustrate the potential and advantages of building analyses and visualizations in *R* directly on SOS. These programmes can be based on near real-time data, but can also be the base for reproducibility of analyses. Reproducible research is already supported well by a variety of *R* packages, and these are now further complemented with open, online data sources such as those found in the Sensor Web.

Finally, we describe future work, especially on the integration of the SOS respectively OGC data models with existing and upcoming endeavours for spatio-temporal data in *R*, like the **spacetime** CRAN package.

References

Botts, M., G. Percivall, C. Reed, and J. Davidson (2008). OGC Sensor Web Enablement: Overview and High Level Architecture. In S. Nittel, A. Labrinidis, and A. Stefanidis (Eds.), *GeoSensor Networks*, Volume 4540 of *Lecture Notes in Computer Science*, pp. 175–190. Springer Berlin / Heidelberg.

Lang, D. T. (2007). R as a Web Client the RCurl package. *Journal of Statistical Software*.

Lang, D. T. (2010). XML: Tools for parsing and generating XML within R and S-Plus. online. version 3.2-0.

Nüst, D. (2010). sos4R - The OGC Sensor Observation Service Client for the R Project.

Open Geospatial Consortium, Inc. (2010). Sensor Observation Service. http://www.opengeospatial.org/standards/sos.

Pebesma, E. (2010). *spacetime: classes and methods for spatio-temporal data*. R package version 0.1-6.

MALDIquant: Quantitative Analysis of MALDI-TOF Proteomics Data

Sebastian Gibb[1,⋆] **and Korbinian Strimmer**[1]

1. Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Härtelstr. 16–18, D-04107 Leipzig, Germany
⋆Corresponding author: sebastian.gibb@studserv.uni-leipzig.de

Keywords: Proteomics, mass spectrometry, MALDI-TOF, Bruker *flex.

MALDI-TOF is a well established technology for mass spectrometric profiling of proteomics data. There are several ongoing efforts to provide open-source analysis software for proteomics studies, such as OpenMS (Kohlbacher et al., 2007) or PROcess (Li, 2005). However, a complete analysis pipeline for MALDI-TOF data for the *R* platform is lacking.

Common tasks and challenges in mass spectrometric analysis are data input, normalization, calibration, peak peaking and other preprocessing task. Due to technological limitations the intensity values of identified peaks cannot be directly compared across multiple spectra. This renders MALDI-TOF data difficult to use for quantitative analysis.

Here we introduce the `MALDIquant` *R* package for analysis of MALDI-TOF proteomics data. Our package provide routines for importing native Bruker *flex binary format, baseline removal, peak picking and, most importantly, procedures for coherent assignment of intensity values across multiple spectra. This approach generalizes the widely used standard procedures that rely on total ion count calibration or 0/1 truncation of peak intensities. The analysis pipeline is illustrated as well as compared with its competitors by classification of cancer proteomics data from the University Hospital Leipzig.

`MALDIquant` and associated *R* packages are available from CRAN.

References

Kohlbacher, O., K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, and M. Sturm (2007). TOPP–the OpenMS proteomics pipeline. *Bioinformatics 23*(2), 191–197.

Li, X. (2005). PROcess: Ciphergen SELDI-TOF Processing. Bioconductor R package archive. R package version 1.26.0.

QuACN: Analysis of Complex Biological Networks using R

Laurin AJ Mueller[1,*,†], **Karl G Kugler**[1,†], **Matthias Dehmer**[1]

1. Institute for Bioinformatics and Translational Research, UMIT, Hall in Tirol, Austria

*Contact author: matthias.dehmer@umit.at † contributed equally

Keywords: Network analysis, bioinformatics, network topology, information-theory

Classical analysis of biological data is mainly based on investigating single, isolated features. As it is now understood that most diseases are driven by sets of interacting genes or proteins this has shifted towards a more complex and holistic perception of this problem Strohman (2002). The first step in network-based analysis of complex biological data is inferring valid and robust network representations of the data. A plethora of packages for this task are available in *R*, e.g. Langfelder and Horvath (2008); Meyer et al. (2008); Altay and Emmert-Streib (2010). A topological analysis of the network provides new insights, as the structure of a network represents the biological function. A quantitative approach is to use so called topological network descriptors, which represent a network structure by a numeric value. Note, that different descriptors capture different structural patterns of the network topology. A small selection of topological network descriptors is available in the **igraph** package Csardi and Nepusz (2006). Recently we published the package **QuACN** on CRAN Mueller et al. (2011). **QuACN** provides a multitude of different topological network descriptors. We want to draw ones attention to the group of parametric graph entropy measures Dehmer and Emmert-streib (2008); Dehmer and Mowshowitz (2011), which are exclusively available in **QuACN**. These measures assign a probability value to each vertex of a graph by using information functionals to calculate the graph's information-content. Topological network descriptors can be used for a multitude of applications in the structural analysis of (biological) networks, e.g. supervised and unsupervised machine learning, or the integrative analysis of networks. Here, we present selected examples of descriptor-based approaches.

References

Altay, G. and F. Emmert-Streib (2010). Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology 4*(1), 132.

Csardi, G. and T. Nepusz (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695.

Dehmer, M. and F. Emmert-streib (2008). Structural information content of networks : Graph entropy based on local vertex functionals. *Computational Biology and Chemistry 32*, 131–138.

Dehmer, M. and A. Mowshowitz (2011, January). A history of graph entropy measures. *Information Sciences 181*(1), 57–78.

Langfelder, P. and S. Horvath (2008, January). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics 9*, 559.

Meyer, P. E., F. Lafitte, and G. Bontempi (2008). minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics 9*, 461.

Mueller, L. A., K. G. Kugler, A. Dander, A. Graber, and M. Dehmer (2011, November). QuACN - An R Package for Analyzing Complex Biological Networks Quantitatively. *Bioinformatics (Oxford, England) 27*(1), 140–141.

Strohman, R. (2002). Maneuvering in the Complex Path from Genotype to Phenotype. *Science 296*, 701–703.

Investigate clusters of co-expressed and co-located genes at a genomic scale using CoCoMap

Marion Ouedraogo [1], **Frédéric Lecerf** [2], **Sébastien Lê** [3]

1. INRA, Agrocampus Ouest UMR 0598 Génétique Animale, Rennes, France
2. Agrocampus Ouest, INRA UMR 0598 Génétique Animale, Rennes, France
3. Agrocampus Ouest, CNRS UMR 6625 Mathématiques Appliquées, Rennes, France

*Contact author: marion.ouedraogo@rennes.inra.fr

Keywords: Multivariate analysis, Biostatistics, Genomics, Genes co-expression, Genes co-location

The study of the genome structure and its role in the gene function regulation had revealed new insights in the regulation of genes expressions by their chromosomal locations. The regulation of the genes expression is in some part due to the genome architecture. The organization of the chromosomes within the nucleus allows interactions between distant regions. Those interacting regions might present some co-expressed genes. The biological interest is to identify the co-located genes within a region which present a co-expression. In our data, genes are considered as statistical variables and individuals as statistical units; we will say that genes are co-expressed if they induce the same structure on the individuals.
The aim of the **CoCoMap** package is to identify co-expression between genomic regions. It uses transcriptomic data and the corresponding chromosomal locations of the genes and is mainly based on multivariate exploratory approaches:
i) A sequence of Principal Component Analyses among the genome is used to identify the co-located genes which present a co-expression.
We introduce the autovariogram representation to identify and discriminate regions of co-expressed genes.
ii) The Multiple Factor Analyses and its features are used to describes the co-expression between the chromosomal regions.
The final output is the groups of co-located genes which present a co-expression. The method has proven its efficiency on simulated data. Preliminary analyses on two different experimental sets identified some common clusters in interaction.
The **CoCoMap** package proposes an interface to the method and its features to investigate transcriptomic data.

References

Baker M. (2011). Genome in three dimensions. *Nature* 470: 289–294

De S. and Babu M. (2010). Genomic neighbourhood and the regulation of gene expression. *Current Opinions in Cell Biology*, 22:326–333.

Sexton T. et al. (2009). Genomic interactions: Chromatin loops and gene meeting points in transcriptional regulation. *Seminars in Cell & Developmental Biology* 20: 849–855

Beta Regression: Shaken, Stirred, Mixed, and Partitioned

Achim Zeileis[1], **Bettina Grün**[2,3⋆], **Francisco Cribari-Neto**[4]

1. Department of Statistics, Universität Innsbruck
2. Institut für Statistik, Ludwig-Maximilians-Universität München
3. Institut für Angewandte Statistik, Johannes Kepler Universität Linz
4. Departamento de Estatística, Universidade Federal de Pernambuco

⋆Contact author: Bettina.Gruen@stat.uni-muenchen.de

Keywords: beta regression, finite mixture, proportion, rate, model-based recursive partitioning

The class of beta regression models is commonly used by practitioners to model variables that assume values in the open standard unit interval $(0, 1)$. It is based on the assumption that the dependent variable is beta-distributed with its mean and precision parameters depending on some regressors. We explore various flavors of parametrizations for the dependence of the response on the regressors. To shake and stirr, either a single or double index model is employed for mean and/or precision, i.e., the parameters are related to potentially different sets of regressors through a linear predictor plus link function(s). This approach naturally incorporates features such as heteroskedasticity or skewness which are commonly observed in data taking values in the standard unit interval, such as rates or proportions. Furthermore, additional heterogeneity in the data can be captured by mixing or partitioning: If covariates are available that explain the heterogeneity, a model-based recursive partitioning approach can be employed. If not, latent class regression is an alternative. All flavors of models are implemented in the *R* package **betareg**, leveraging the **flexmix** and **party** packages.

References

Cribari-Neto, F. and A. Zeileis (2010). Beta regression in *R*. *Journal of Statistical Software 34*(2), 1–24.

Grün, B. and F. Leisch (2008). Flexmix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software 28*(4), 1–35.

Zeileis, A., T. Hothorn, and K. Hornik (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics 17*(2), 492–514.

Regression Models for Ordinal Data: Introducing *R*-package ordinal

Rune Haubo B Christensen[1,⋆]

1. Technical University of Denmark, Department of Informatics and Mathematical Modelling, Section for Mathematical Statistics, Richard Petersens Plads, Building 305, Room 122, DK-2800 Kgs. Lyngby, Denmark
⋆Contact author: rhbc@imm.dtu.dk

Keywords: Cumulative Link Models, Mixed Effects, Location-Scale, Partial Proportional Odds

Ordered categorical data, or simply *ordinal* data, are commonplace in scientific disciplines where humans are used as measurement instruments. Examples include school gradings, ratings of preference in consumer studies, degree of tumor involvement in MR images and animal fitness in field ecology. Cumulative link models (Agresti, 2002) are a powerful model class for such data since observations are treated rightfully as categorical, the ordered nature is exploited and the flexible regression framework allows in-depth analyses. A pertinent latent variable interpretation of cumulative link models is an important aspect in many applications in sensometrics, psychometrics and other social sciences. Cumulative link (mixed) models are implemented in functions `clm` and `clmm` in package **ordinal** (Christensen, 2011).

The **MASS** function `polr` is a popular implementation of basic cumulative link models taking its name from the proportional odds model—a cumulative link model with a logit link. `clm` and `clmm` extends `polr` in a number of ways by allowing for random effects, scale effects, nominal effects, flexible link functions and structured thresholds, further, several estimation methods are available including a fast Newton scheme. Mixed effects models are estimated with standard Gauss-Hermite quadrature, the highly accurate adaptive Gauss-Hermite quadrature or the flexible Laplace approximation accommodating nested as well as crossed random effect structures.

Collectively these options facilitate a fuller analysis of ordinal data. The model framework embrace location-scale models (McCullagh, 1980; Cox, 1995), allows for so-called partial proportional odds (Peterson and Harrell Jr., 1990), facilitates inference for the link function (Genter and Farewell, 1985) and allows assessment of linearity of the response scale. `profile` likelihood methods help visualize the likelihood function and provide accurate confidence intervals via a `confint` method. `dropterm`, `addterm` and `anova` methods facilitate model comparison. The implementation is primarily in *R*-code with computer intensive parts in *C*.

References

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

Christensen, R. H. B. (2011). ordinal—regression models for ordinal data. R package version 2010.12-15 http://www.cran.r-project.org/package=ordinal/.

Cox, C. (1995). Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in medicine 14*, pp. 1191–1203.

Genter, F. C. and V. T. Farewell (1985). Goodness-of-link testing in ordinal regression models. *The Canadian Journal of Statistics 13*(1), pp. 37–44.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B 42*, pp. 109–142.

Peterson, B. and F. E. Harrell Jr. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics 39*, pp. 205–217.

Multiple choice models: why not the same answer? A comparison among LIMDEP, R, SAS and Stata

Giuseppe Bruno [1]

1. Bank of Italy Economic Research and International relations. giuseppe.bruno@bancaditalia.it

Keywords: Multinomial models, Probit, Gibbs sampler, MCMC.

The recent past has seen a great deal of research into categorical response models especially in the econometrics and behavioral literature. The most popular categorical response models are the Multinomial Logit **(MNL)** and the Multinomial Probit **(MNP)**. The choice space for these models can be dichotomous (yes or no) or polytomous (e.g. points on a Likert scale). When dealing with more than two choices, a trade-off comes out between numerical tractability and model flexibility. In particular the more simple mathematical features of the **MNL** model entail the validity of the Independence of Irrelevant Alternatives **(IIA)** assumption. On the other hand, the adoption of the **MNP** model provides the maximum modeling flexibility at a much higher computational price. In this paper we investigate the availability and the accuracy of canned estimation algorithms for the multinomial models in the most spread packages employed by the researchers at the Bank of Italy. Here we compare the classical Maximum Likelihood method along with the Markov Chain Monte Carlo **(MCMC)** method available in the R package named **MNP**. A thorough comparison of the algorithms available in the packages **LIMDEP**, **R**, **SAS** and **Stata** is provided along with accuracy and performances indications. Preliminary results are quite insightful. As it is foreseeable, the availability of a closed-form expression for the choice probabilities in the multinomial logit model, shown in the following equation: $Pr(y_i = j|X_i) = \frac{exp(X_i \cdot \beta_j + C_{ij}\lambda)}{\Sigma_{k=1}^{J} exp(X_i \cdot \beta_k + C_{ik}\lambda)}$, allows the employment of stable and replicable algorithms. On the other hand the estimation of a multinomial probit with Maximum Likelihood involves the integration of a multivariate normal distribution. This task is not numerically straightforward. Different packages provides different answers and sometimes they don't even provide an answer. R provides the **MNP** package for fitting a Bayesian multinomial probit model with Markov chain Monte Carlo methods (see Imai and van Dyk 2005) . **SAS**, **STATA** and **LIMDEP** fit the multinomial probit model via Maximum Likelihood. Some empirical applications taken from the health insurance and the travel mode choices are presented. It is shown how the employment of the estimation algorithms for these models would benefit from the comparison with the implementation of the same algorithm in different packages and how helpful the comparison with already available numerical benchmarks would be (see Econometric Benchmarks 2010).

References

Econometric Benchmarks (2010).
`http://www.stanford.edu/~clint/bench/#logit`.

G. Glasgow and R.M. Alvarez (2008). Discrete Choice Methods,
`http://www.polsci.ucsb.edu/faculty/glasgow/glasgow_alvarez_final.pdf`.

Odysseus vs. Ajax: How to build an *R* presence in a corporate *SAS* environment

Derek McCrae Norton[1,⋆]

1. Norton Analytics

⋆Contact author: Derek.Norton@nortonanalytics.com

Keywords: Business Analyitcs, Open Source, SAS

A lot has changed in the year since I gave my talk at the UseR!2010 [1]. There have been multiple prominent articles about *R* as detailed in [2]. There has been a noticable decline in *SAS* and a steady increase in *R* as detailed in [3]. As well as many other factors, but there is still resistance in the corporate environment. There is certainly less than there was a year ago, but it is still there.

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.

Last year, I likened *R* to David and *SAS* to Goliath, however I revise that now as shown in the title. Both *R* and *SAS* are powerful forces in a war against poor analytics (Troy), but *R* is the crafty intelligent fighter like Odysseus while *R* is the formitable, but brutish fighter like Ajax [4].

References

1. Revolution Analytics Blog - RMedia, http://blog.revolutionanalytics.com/rmedia/.
2. Norton, Derek M. (2010). David v. Goliath: How to build an *R* presence in a corporate *SAS* environment. In *useR! 2010, The R User Conference, (Gaithersburg, Maryland, USA)*.
3. Muenchen, Robert A. (2011). The Popularity of Data Analysis Software. http://sites.google.com/site/r4statistics/popularity.
4. Homer. The Odyssey.

A Validation/Qualification Solution for *R*

Michael O'Connell[1], Ian Cook[1,*]

1. Spotfire, TIBCO Software Inc.
*Contact author: icook@tibco.com

Keywords: validation, qualification, regulatory compliance, FDA, clinical trial

To date, *R* is not widely used in regulated environments, e.g., clinical trials for pharmaceuticals and medical devices. A common misperception exists that *R* cannot support the various regulatory requirements for validation/qualification. We present a straightforward framework for successfully complying with regulatory software validation requirements including FDA 21 CFR Part 11 and other GxP documents. Recognizing that validation/qualification applies to the software and to its installation and operation, we present a solution that facilitates qualification of an *R* installation to meet IQ/OQ/PQ standards. We cite previous *useR!* proceedings on the topic, and discuss the combination of factors enabling growth in the use of *R* in regulated environments, including guidance from the R Foundation and the availability of tools supporting validation/qualification.

References

FDA (2010). Code of Federal Regulations Title 21, Part 11, http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?cfrpart=11.

Harrell, Frank E Jr. (2007). R for Clinical Trial Reporting: Reproducible Research, Quality and Validation. *useR! 2007 (Iowa State University)*, http://biostat.mc.vanderbilt.edu/twiki/pub/Main/FHHandouts/dmcreport.pdf.

R Foundation for Statistical Computing (2008). R: Regulatory Compliance and Validation Issues, A Guidance Document for the Use of R in Regulated Clinical Trial Environments, http://www.r-project.org/doc/R-FDA.pdf.

Schwartz, Marc (2007). Use of R in Clinical Trials and Industry-Sponsored Medical Research. *useR! 2007 (Iowa State University)*, http://user2007.org/program/presentations/schwartz.pdf.

R as a statistical tool for human factor engineering

Enrico Branca[1,⋆]

1. R-Adamant
⋆Contact author: ebranca@r-adamant.org

Keywords: R-Adamant, human resources, jobs, applied statistics

In the last decades global market has evolved toward a new dimension by creating an environment where different countries , attitudes , cultures and work organization coexist together. Companies are so exposed to a new management challenge to keep high business performances.

In this scenario Human Resources Management has become more than ever a key factor to build company's success. Tasks as: find and leverage talents, promote a culture of excellence by incouraging outstanding achievements, assemble a collective knowledge by sharing experience for common good, reduce employee turnovers are, among others, crucial to ensure company's survival.

Considering that in some cases Universities have not been providing with enough graduates, offset the gap in those sectors with skills shortage is a competitive investment in corporate branding.

As reported by a recent research at *Yale School of Management* in 2009: *"Those who graduate in bad economies may suffer from underemployment and are more likely to experience job mismatching since they have fewer jobs from which to choose. The disadvantage might be eliminated if workers can easily shift into jobs and career paths."*

An analytical approach can be used to create a functional relation between the HR function and enterprise risk management process. This allows to quantify and mitigate the possible risks related with human capital that can affect company's business performances.
Time series models (ARIMA, moving average) and multivariate analysis techniques (ANOVA, Generalised linear models) are indeed used as a support for both HR managers and risk managers.

Statistical methods and tools widely used for Business Intelligence and financial risk management may promote a new corporate culture more focused on Sense of Accomplishment and networks building than compliance to static and outdated behavioural models.

The tutorial scope is to show how a valuable tool as **R-Adamant** can be a powerful ally for researchers and university students as well as for companies.

References

Lisa B. Kahn (First Draft: March, 2003 Current Draft: August 13, 2009). The Long-Term Labor Market Consequences of Graduating from College in a Bad Economy,*Yale School of Management*.

U.S. Department of Labor - U.S. Bureau of Labor Statistics (Summary 10-05 / June 2010), *Issues in Labor Statistics*.

Council for Adult and Experiential Learning (CAEL) and Southern Regional Education Board (SREB) (2007). *Increasing Degree Completion Among Adult Learners: Policies and Practices to Build Greater State Capacity*. Chicago, IL: CAEL.

GWSDAT (GroundWater Spatiotemporal Data Analysis Tool)

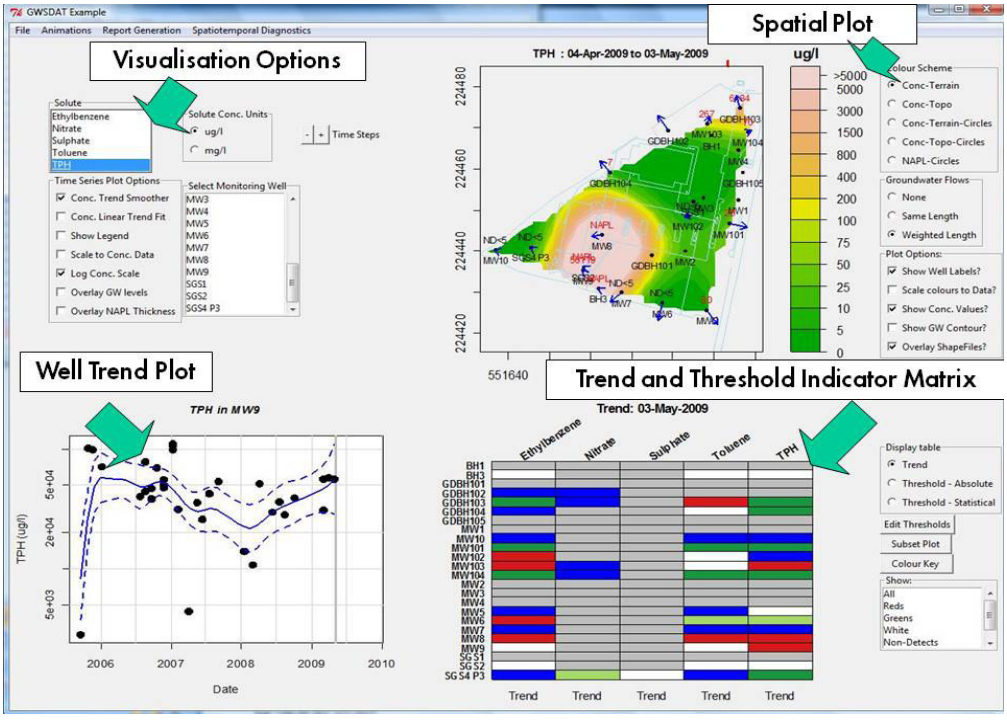
Wayne Jones[1,*], Michael Spence[2]

1. Shell Projects and Technology, Statistics and Chemometrics, Shell Technology Centre Thornton, Chester, United Kingdom
2. Shell Projects and Technology, Soil and Groundwater Group, Shell Technology Centre Thornton, Chester, United Kingdom
*Contact author: wayne.w.jones@shell.com

Keywords: Environmental monitoring, spatiotemporal modelling, application development, interface with other languages.

The GroundWater Spatiotemporal Data Analysis Tool (GWSDAT) is a user-friendly software application developed by Shell for the analysis and visualisation of visualise trends in environmental (groundwater) monitoring data. In this presentation we will discuss the underlying software architecture and demonstrate GWSDAT functionality which includes:

- Linking GWSDAT to Microsoft products for a user friendly application entry point (e.g. Excel, PowerPoint, Word) and automatic report generation (R packages: **RExcel** and **rcom**).
- Graphical User Interface (R packages: **tcltk** and **rpanel**).
- Time series trend detection (R packages: **sm**, **Kendall** and **zoo**).
- Methods for visualising and handling Spatial data (R packages **deldir**, **sp**, **splancs** and **maptools).**
- Smoothing and spatial plot animations for spatiotemporal trend detection.



GWSDAT Screenshot.

IntR – Interactive GUI for *R*

Fabio Veronesi[1,*], Ron Corstanje[1]

1. Cranfield University, Bld.37 - School of Applied Sciences, MK43 0AL, Cranfield, Bedfordshire
*Contact author: f.veronesi@cranfield.ac.uk

Keywords: GUI, interactive interface, *R*, geostatistics, python,

In recent years, *R* is increasingly being used by soil scientists. Its popularity is due to the fact that it is free, open-source, but also due to the large number of packages dedicated to spatial statistics and geostatistical interpolation. On the other hand, the necessity to learn a brand new programming language, with a relatively steep learning curve, has kept and still keeps most soil scientists away from *R*.

This software is designed to solve this problem and speed up the learning process. IntR is a graphical user interface, wrote in Python, that eases the procedure of the creation and execution of an *R* script for 2D and 3D geostatistical interpolation. This interface works by asking "questions" to the user: for instance to select the data file, the model of the variogram ect., and compiling the *R* script based on the commands inserted. The user is therefore in full control of every part of the process. The script is run in batch mode and the results are shown at the end of the process and saved. Normally the output of a geostatistical analysis is a table of the results, an image of the variogram, and images of the prediction and uncertainty maps and, where available, a plot of the observed versus the predicted values. The script is also saved for the user to control it and thus learn the language. The algorithms that can be used in IntR are: for the 2D version, inverse distance interpolation, ordinary kriging, universal kriging and regression kriging, random forest and CART, plus a module for obtaining a point grid from a polygon shape file and a module for the variogram and the anisotropy analysis. For the 3D version, the user can choose between inverse distance, ordinary kriging and universal kriging, plus a module to create a 3D prediction grid, a module for variogram and anisotropy analysis, a module for creating a series of slice images of the 3D map and a module for sticking the slices in an animation video. The packages used to perform all the analysis are: **gstat**, **sp**, **maptools**, **randomForest**, **tree**, **rgdal**, **lattice**, **akima**.

The software was developed with two different datasets, one in 2D and the other in 3D. The bidimensional dataset is composed by 30 textural data samples (sand, clay and silt) collected at three depth interval: 0-10 cm, 10-30 cm and 30-70 cm. The second dataset is composed by 57 cone-index samples taken on a regular grid, with a vertical resolution of 4 cm. In both cases we used geophysical covariates for the prediction, namely EM38, EM31 and Gamma-ray data. Regarding the 2D case study, the results show that universal kriging is the best performer among the major prediction interpolators incorporated into IntR. In the 3D case study, a comparison between a 3D ordinary kriging and 3D universal kriging was undertaken. As expected, the best predictor in 3D universal kriging. For this reason, universal kriging was used to create a 3D map of the Lany field.

References

Bivand R.S., Pebesma E.J., Gomez-Rubio V. (2008). Applied spatial data analysis with R. *Springer, NY.*

ESRI (1998). ESRI Shapefile Technical Description.

Liaw A. and Wiener M. (2002). Classification and Regression by randomForest. In *R News 2(3).*

Pebesma, E.J. (2004). Multivariable geostatistics in S: the gstat package. In *Computers & Geosciences, 30: 683-691.*

Visualisation and modelling of soil data using the aqp package

Pierre Roudier[1,⋆], **Dylan Beaudette**[2]

1. Landcare Research, New Zealand
2. Natural Resources Conservation Service, USDA, USA
⋆Contact author: roudierp@landcareresearch.co.nz

Keywords: Soil science, visualisation, aggregation, classification

Soils support nearly all terrestrial food webs, are composed of a dynamic mixture of organic and mineral constituents, and occur within the relatively fragile intersection between atmosphere, biosphere, and lithosphere. With the growth of the human population and a limited soil resource to feed those populations, the study of such a complex environment arises as a prime challenge (Sanchez et al., 2009). However, the importance of soils has been acknowledged for a long time. An impressive quantity of soil information has been collected to support soil survey operations, natural resource inventories, and research over at least the last 100 years. Soils are routinely sampled and studied according to characteristic layers (called *horizons*), resulting in a complex data structure, with the dimensions of location, depth, and property space.

The high dimensionality and grouped nature of large soil profile collections can complicate standard analysis, summarization, and visualization, especially within a research community that is a traditional spreadsheet user. Soil data manipulation and analysis is also complicated by difficulties associated with processing horizon data that vary widely in depth and thickness. Scalability and reproducibility are also becoming significant issues as the size of the databases grow. Finally, while the investigation of soil profile characteristics and horizon-level morphology is strongly based on visual and tactile cues, the challenge of communicating these data is traditionally addressed using written narrative or tabular form.

R is a suitable platform to address those challenges, and to develop tools that would provide soil scientists with aggregation, modelling and visualisations for soil data. An *R* package, **aqp** (Algorithms for quantitative pedology), has been developed to extend *R*'s methods to the specificities of soil data. Specialized S4 classes have recently been added to support the multivariate hierarchy of linked spatial data (e.g. coordinates), site data (e.g. landscape position), and horizon data (e.g. clay content at 10 cm). Various new aggregation and classification methods are also available, and make use of the parallelised environment provided by the **plyr** package (Wickham, 2011). Examples of the **aqp** functionalities are proposed on different soil databases, for data visualisation, analysis and classification. The future developments of the package, especially its interactions with other packages used for soil data analysis (e.g. **sp** Pebesma and Bivand, 2005), is also discussed.

References

Pebesma, E. J. and R. S. Bivand (2005, November). Classes and methods for spatial data in R. *R News 5*(2), 9–13.

Sanchez, P. A., S. Ahamed, F. Carre, A. E. Hartemink, J. Hempel, J. Huising, P. Lagacherie, A. B. McBratney, N. J. McKenzie, M. d. L. Mendonca-Santos, B. Minasny, L. Montanarella, P. Okoth, C. A. Palm, J. D. Sachs, K. D. Shepherd, T.-G. Vagen, B. Vanlauwe, M. G. Walsh, L. A. Winowiecki, and G.-L. Zhang (2009). Digital Soil Map of the World. *Science 325*(5941), 680–681.

Wickham, H. (2011). *plyr: Tools for splitting, applying and combining data*. R package version 1.4.

survAUC: Estimators of Prediction Accuracy for Time-to-Event Data

Matthias Schmid[1,⋆], **Sergej Potapov**[1] **and Werner Adler**[1]

1. Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander University Erlangen-Nuremberg
⋆Contact author: matthias.schmid@imbe.med.uni-erlangen.de

Keywords: Survival analysis; prediction accuracy; risk prediction

The evaluation of predictions for continuous time-to-event outcomes has become a key interest in biostatistical research. In contrast to the situation where predictions for uncensored outcomes have to be evaluated, deriving measures of prediction accuracy is not straightforward in the presence of censored observations. This is because traditional performance measures for continuous outcomes (such as the mean squared error or the R^2 fraction of explained variation) are biased if applied to censored data.

The `survAUC` package (Potapov et al. 2011) implements a variety of estimators to evaluate survival predictive accuracy. The focus is on estimators of discrimination indices that measure how well a prediction model separates observations having an event from those having no event (Pepe et al. 2008). In addition, `survAUC` provides `R` functions to estimate likelihood-based coefficients (O'Quigley et al. 2005) and measures based on scoring rules (Gerds and Schumacher 2006).

References

Potapov S, Adler W, Schmid M (2011). *survAUC: Estimators of Prediction Accuracy for Time-to-Event Data*. R package version 1.0-0. http://cran.at.r-project.org/web/packages/survAUC/index.html.

Pepe MS, Zheng Y, Jin Y, Huang Y, Parikh CR, Levy WC (2008). Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis 14*, 86–113.

Gerds TA, Schumacher M (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal 48*, 1029–1040.

O'Quigley J, Xu R, Stare J (2005). Explained randomness in proportional hazards models. *Statistics in Medicine 24*, 479–489.

Higher-order likelihood inference in meta-analysis using *R*

Annamaria Guolo[1*], **Cristiano Varin**[2]

1. Università degli Studi di Verona, Italy
2. Università Ca' Foscari, Venezia, Italy
*Contact author: annamaria.guolo@univr.it

Keywords: higher-order asymptotics, linear mixed-effects model, meta-analysis, Skovgaard's statistic, small sample inference

Standard likelihood inference is known to improve common non-likelihood techniques in meta-analysis (Hardy and Thompson, 1996). Nevertheless, relying on typical first-order approximations, as for example the χ^2 distribution for Wald and likelihood ratio statistics, can give rise to inaccurate results. This drawback is a consequence of small sample sizes, which are typical in meta-analysis. Resorting to the theory of higher-order asymptotics provides remarkably more precise results than first-order counterpart. See, for example, Severini (2000) and Brazzale et al. (2007) for a general overview of higher-order asymptotics. We present an *R* package which implements first-order likelihood inference and the second-order adjustment to the log-likelihood ratio statistic of Skovgaard (2006), either for meta-analysis and meta-regression problems, following the results in Guolo (2011). The package allows inference on fixed- and random-effect components of linear mixed models used in meta-analysis. The functionality of the package will be illustrated on a real example from the medical literature.

References

Brazzale, A R, Davison, A C, and Reid N (2007). *Applied Asympotics: Case Studies in Small Sample Statistics*. Cambridge University Press: Cambridge.

Guolo A (2011). Higher-order likelihood inference in meta-analysis and meta-regression. *Submitted*.

Hardy R J, Thompson S G (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine 15*, 619–629.

Severini, T A (2000). *Likelihood Methods in Statistics*. Oxford University Press: Oxford.

Skovgaard I M (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli 2*, 145–165.

Gaussian copula regression using *R*

Guido Masarotto[1], **Cristiano Varin**[2*]

1. Università degli Studi di Padova, Italy
2. Università Ca' Foscari, Venezia, Italy
*Contact author: sammy@unive.it

Keywords: discrete time series; Gaussian copula; likelihood inference; longitudinal data; spatial data.

Marginal regression models for non-normal correlated responses are typically fitted by the popular generalized estimating equations approach of Liang and Zeger (1986). Despite several theoretical and practical advantages, likelihood analysis of non-normal marginal regression models is much less diffuse, see *e.g.* Diggle et al. (2002). The main reason is the difficult identification of general classes of multivariate distributions for categorical and discrete responses. Gaussian copulas provide a possible solution with a general framework for modelling dependent responses of any type (Song, 2000). Gaussian copulas combine the simplicity of interpretation in marginal modelling with the flexibility in the specification of the dependence structure. Despite this, Gaussian copula regression had still a limited use since for noncontinuous dependent responses the likelihood function requires the approximation of high-dimensional integrals. Masarotto and Varin (2010) propose an adaptation of the Geweke-Hajivassiliou-Keane importance sampling algorithm (Keane, 1994; Train 2003) to overcome the numerical difficulties of the likelihood inference.

The *R* package **mr** implements the methodology discusses in Masarotto and Varin (2010). The package allows a flexible specification of the marginals and the dependence structure. At the time of writing, the package contains methods for inference in regression models for longitudinal and clustered responses, time series, spatially correlated observations and cross-correlated studies. The functionality of the package will be illustrated on several real data examples arising in Biostatistics.

References

Diggle, P J , Heagerty, P, Liang, K-Y, Zeger, S L (2002). *Analysis of Longitudinal Data*. Second edition. Oxford University Press: Oxford.

Keane, M P (1994). A computationally practical simulation estimator for panel data. *Econometrica 62*, 95–116.

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

Masarotto, G, Varin, C (2010). Gaussian dependence models for non-Gaussian marginal regression. *Submitted*.

Song, P X-K (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics 27*, 305–320.

Train, K E (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press: Cambridge.

Multinomial Processing Tree Models in R

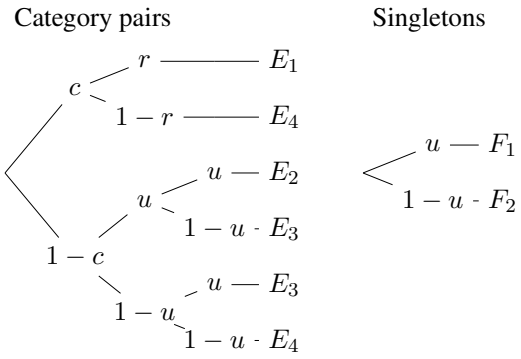
Florian Wickelmaier[1,⋆]

1. Department of Psychology, University of Tübingen
⋆Contact author: florian.wickelmaier@uni-tuebingen.de

Keywords: Cognitive Psychometrics, Multinomial Models, Latent Variables, Categorical Data

Multinomial processing tree models are a class of statistical models for categorical data with latent parameters. These parameters are the link probabilities of a tree-like graph and represent the cognitive processing steps executed to arrive at observable response categories (Batchelder & Riefer, 1999; Erdfelder et al., 2009; Riefer & Batchelder, 1988).

In this presentation, the **mpt** package (Wickelmaier, 2011) in *R* is introduced which provides functions for fitting and testing such models. The model structure is represented symbolically using a simple formula interface. Parameter estimation is carried out by the expectation-maximization algorithm described in Hu and Batchelder (1994). The statistical procedures are illustrated using examples from cognitive psychology and memory research.



References

Batchelder, W. H. & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6, 57–86.

Erdfelder, E., Auer, T., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie*, 217, 108–124.

Hu, X. & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, 59, 21–47.

Riefer, D. & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, 95, 318–339.

Wickelmaier, F. (2011). **mpt:** Multinomial processing tree (MPT) models. *R* package version 0.3-0. http://CRAN.R-project.org/package=mpt

Detecting Invariance in Psychometric Models with the psychotree Package

Carolin Strobl[1,*], Florian Wickelmaier[2], Julia Kopf[1] and Achim Zeileis[3]

1. Deparment of Statistics, LMU Munich
2. Psychological Institute, Universität Tübingen
3. Department of Statistics, Universität Innsbruck

*Contact author: carolin.strobl@stat.uni-muenchen.de

Keywords: Bradley-Terry Model, Rasch Model, Differential Item Functioning (DIF), Parameter Instability, Model-Based Recursive Partitioning.

The **psychotree** package offers a statistical toolbox for detecting parameter differences in psychometric models, including different worth parameters in Bradley-Terry models (Strobl, Wickelmaier and Zeileis, 2011) and differential item functioning (DIF) in the Rasch model (Strobl, Kopf and Zeileis, 2010a,b). The method for detecting different worth parameters in Bradley-Terry models is implemented in the `bttree` function, the DIF detection method for the the Rasch model is implemented in the `raschtree` function. Both methods are based on a general model-based recursive partitioning framework employing generalized M-fuctuation tests for detecting differences in the model parameters between different groups of subjects (Zeileis and Hornik, 2007; Zeileis, Hothorn and Hornik, 2008). The main advantage of this approach is that it allows to detect groups of subjects exhibiting different model parameters, that are not pre-specifed, but are detected automatically from combinations of covariates. The talk outlines the statistical methodology behind **psychotree** as well as its practical application by means of illustrative examples.

References

Strobl, C., J. Kopf, and A. Zeileis. A New Method for Detecting Differential Item Functioning in the Rasch Model. Technical Report 92, Department of Statistics, Ludwig-Maximilians-Universität München, Germany, 2010. URL: `http://epub.ub.uni-muenchen.de/11915/`.

Strobl, C., J. Kopf, and A. Zeileis. "Wissen Frauen weniger oder nur das Falsche? – Ein statistisches Modell für unterschiedliche Aufgaben-Schwierigkeiten in Teilstichproben." *Allgemeinbildung in Deutschland – Erkenntnisse aus dem SPIEGEL Studentenpisa-Test*. Ed. S. Trepte and M. Verbeet Wiesbaden: VS Verlag, 2010, 255–272.

Strobl, C., F. Wickelmaier, and A. Zeileis. "Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning." (To appear.). *Journal of Educational and Behavioral Statistics* (2011).

Zeileis, A., T. Hothorn, and K. Hornik. "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics* 17 (2008): 492–514.

Zeileis, Achim and Kurt Hornik. "Generalized M-Fluctuation Tests for Parameter Instability." *Statistica Neerlandica* 61 (2007): 488–508.

Investigating multidimensional unfolding models using R2WinBUGS

Shiu-Lien Wu[1] , Wen-Chung Wang[2]

1. National Chung-Cheng University, Taiwan
2. The Hong Kong Institute of Education
*Contact author: subert4@gmail.com

Abstract

It has been argued that unfolding models could be more appropriate for describing responses to attitude items or Likert items than dominance models. However, most existing unfolding models are either unidimensional or fitting to binary data under the framework of item response theory (IRT). When there are multiple tests of Likert items, or when a Likert item measures more than one latent trait simultaneously, unidimensional unfolding models become inefficient or inappropriate. Meanwhile, if a model can be fit only to binary data, the applications are limited. To resolve these problems, we developed the confirmatory multidimensional generalized graded unfolding model, which is a multidimensional extension of the generalized graded unfolding model (Roberts, Donoghue, & Laughlin, 2000), and conducted a series of simulations to evaluate its parameter recovery by using *R* and the *R* package **R2WinBUGS**. The simulation study demonstrated that the parameters of the new model can be recovered fairly well. In addition, we analyzed a real data set about tattoo attitude to depict the implication and applications of the new model and to demonstrate its advantages over the unidimensional model. The results showed that the multidimensional model had a better fit than the unidimensional one (log PsBF = 27.2); the multidimensional model yielded higher reliability estimates (.92, .89, .83) for the 3 latent traits than the unidimensional one (.84, .85, .83); and the multidimensional model yielded higher correlation estimates among the 3 latent traits (.20 ~ .84) than the unidimensional model (.04 ~ .30).

Keywords: multidimensional item response theory, unfolding models, Markov chain Monte Carlo, R2WinBUGS

References

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software, 12*, 1-16.

Tests for Multivariate Linear Models with the car Package

John Fox[1,⋆]

1. McMaster University, Canada
⋆Contact author: jfox@mcmaster.ca

Keywords: Multivariate Analysis of Variance, Repeated Measures, Linear Hypothesis Tests

It is straightforward to fit multivariate linear models in *R* with the `lm` function: Simply specify the left-hand side of the linear-model formula as a matrix of responses. *R* also has facilities for testing multivariate linear models via the `anova` function (as described in Dalgaard, 2007). Although the `anova` function is very flexible, applied to a multivariate linear model it calculates sequential (often termed "type I") tests, which rarely are of interest, and performing other common tests, especially for repeated-measures designs, is relatively inconvenient. In contrast, the `Anova` function in the **car** package (associated with Fox and Weisberg, 2011) can perform partial tests for the terms in a multivariate linear model, either obeying the principle of marginality ("type II" tests) or violating it ("type III" tests), including simply specified multivariate and univariate tests for repeated-measures models. In addition, the `linearHypothesis` function in the **car** package can test arbitrary linear hypothesis for multivariate linear models, including models for repeated measures. Both the `Anova` and `linearHypothesis` functions and their associated `summary` methods return a variety of information useful in further computation on multivariate linear models, such as the graphical display of hypothesis tests (see, e.g., Fox et al., 2009).

References

Dalgaard, P. (2007). New functions for multivariate analysis. *R News 7*(2), 2–7.

Fox, J., M. Friendly, and G. Monette (2009). Visualizing hypothesis tests in multivariate linear models: The heplots package for r. *Computational Statistics 24*, 233–246.

Fox, J. and S. Weisberg (2011). *An R Companion to Applied Regression*. Sage Publication.

missMDA: a package to handle missing values in and with multivariate exploratory data analysis methods

Julie Josse[1], **Francois Husson**[1]

1. Applied Mathematics Department, Agrocampus, Rennes, France
*Contact author: julie.josse@agrocampus-ouest.fr

Keywords: Missing values, Principal component methods, Multiple imputation, Confidence ellipses

In this presentation, we describe the **missMDA** package which is dedicated to handle missing values in exploratory data analysis methods such as principal component analysis (PCA) and multiple correspondence analysis. This package provides the classical outputs (scores, loadings, graphical representations, etc.) of principal component methods despite the missing values. It also gives confidence areas around the position of the points (individuals and variables) representing the uncertainty due to missing values. The package can also be used to perform single or multiple imputation for continuous and categorical variables in a general framework. In this presentation, we describe the underlying method through PCA.

A common approach to handle missing values in PCA consists in minimizing the loss function (the reconstruction error) over all nonmissing elements. This can be achieved by the iterative PCA algorithm (also named expectation maximization PCA, EM-PCA) described in Kiers (1997). It consists in setting the missing elements at initial values, performing the analysis (the PCA) on the completed data set, filling-in the missing values with the reconstruction formula (the PCA model, Caussinus (1986)) and iterate these two steps until convergence. The parameters (axes and components) and the missing values are then simultaneously estimated. Consequently, this algorithm can be seen as a single imputation method. To avoid overfitting problems, regularized version of the EM-PCA algorithm have been proposed (Josse and Husson, 2011; Ilin and Raiko, 2010).

After the point estimate, it is natural to focus on the variability of the parameters. However, the variance of the axes and components estimated from the completed data set (obtained with the EM-PCA algorithm) is underestimated. Indeed, the imputed values are considered as observed values and consequently the uncertainty of the prediction is not taken into account in the subsequent analysis. It is possible to resort to multiple imputation (Rubin, 1987) to avoid this problem. Multiple imputation consists first in generating different plausible values for each missing values. Then it consists in performing the statistical analysis on each imputed data set and combining the results. Josse and Husson (2011) have proposed a new method to generate multiple imputed data sets from the PCA model. They have also proposed two ways to visualize the influence of the different predictions of the missing values onto the PCA results. It leads to confidence areas around the position of the individuals and of the variables on the PCA maps.

References

Caussinus, H. (1986). Models and uses of principal component analysis. In J. de Leeuw, W. Heiser, J. Meulman, and F. Critchley (Eds.), *Multidimensional Data Analysis*, pp. 149–178. DSWO Press.

Ilin, A. and T. Raiko (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research 11*, 1957–2000.

Josse, J. and F. Husson (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*.

Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrica 62*, 251–266.

Rubin, D. B. (1987). *Multiple imputation for non-response in survey*. Wiley.

MAINT.DATA: Modeling and Analysing Interval Data in R

A. Pedro Duarte Silva[1,⋆], **Paula Brito**[2]

1. Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa at Porto, Porto, Portugal
2. Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto, Porto, Portugal
⋆Contact author: psilva@porto.ucp.pt

Keywords: Symbolic data, Interval data, Parametric modelling of interval data, Statistical tests for interval data, Skew-Normal distribution

In the classical model of multivariate data analysis, data is represented in a $n \times p$ data-array where n "individuals" (usually in rows) take exactly one value for each variable (usually in columns). Symbolic Data Analysis (Diday and Noirhomme-Fraiture (2008), Noirhomme-Fraiture and Brito (2011)) provided a framework where new variable types allow to take directly into account variability and/or uncertainty associated to each single "individual", by allowing multiple, possibly weighted, values for each variable. New variable types - interval, categorical multi-valued and modal variables - have been introduced. We focus on the analysis of interval data, i.e., where elements are described by variables whose values are intervals of $\mathbb{R}$. Parametric inference methodologies based on probabilistic models for interval variables are developed in Brito and Duarte Silva (2011) where each interval is represented by its midpoint and log-range, for which Normal and Skew-Normal (Azzalini and Dalla Valle (1996)) distributions are assumed. The intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix, which are represented by five different possible configurations.

In this work, we introduce the package MAINT.DATA, which implements the proposed methodologies in R. It introduces a data class for representing interval data. MAINT.DATA includes functions for modeling and analysing interval data, in particular maximum likelihood estimation and statistical tests for the different considered configurations. Methods for (M)ANOVA and Linear and Quadratic Discriminant Analysis of this data class are also provided.

References

Azzalini, A. and Dalla Valle, A. (1996). The multivariate Skew-Normal distribution, *Biometrika 83 (4)*, pp. 715–726.

Brito, P. and Duarte Silva, A.P. (2011). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, (in press).

Noirhomme-Fraiture, M. and Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining* (in press).

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester.

Web 2.0 for R scripts & workflows: Tiki & PluginR

Xavier de Pedro Puente[1,2,⋆], **Alex Sánchez Pla**[1,2]

1. Department of Statistics. Faculty of biology. University of Barcelona.
2. Statistics and Bioinformatics Unit. Vall d'Hebron Research Institute. Catalonia. Spain
⋆Contact author: xdpedro@ir.vhebron.net

Keywords: GUI, Web 2.0, Free/libre Software, Tiki Wiki CMS Groupware, PluginR.

Abstract

The need to work with colleagues from other institutions is very common around Academia and Science. Teams often find tools to communicate and coordinate with other web platforms to improve collaboration across space and time. Although analysis and visualization of data with *R* is becoming very popular, development teams frequently look also for web-based graphical user interfaces for the end users of those *R* scripts. The list of prototypes and publicly announced free tools (R (2011)) includes programs of all kinds. However, a quick review of these tools led us to similar conclusion reached by other researchers such as Saunders (2009): most of these programs seem to present problems in the short to medium term. Those problems arise from the fact that either such programs no longer work with current stable versions of standard and free web technology, because its development seems to have been discontinued for years. Or because they are too difficult to install or use for most scientists or people who are not professionals in web technology. Therefore, we decided in our research groups to contribute to the development of a relatively new approach, different from the latest approaches presented in the latest years (Ooms (2009), Nakano and Nakama (2009) and others): a plugin for *Tiki Wiki CMS Groupware* (also known as "*Tiki*"), a mature collaborative web 2.0 framework released as free/libre open source software, somewhat similar to the R extension for Mediawiki, but with all the extra features from this "Tightly Integrated Knowledge Infrastructure" that Tiki represents), along with its decentralized but truly successful development model (Tiki (2011)). This new PluginR (De Pedro (2011)), has so far allowed the development of a Web application of use in research on the Teaching and Learning field (De Pedro et al. (2010)), as well as to develop web interfaces for Basic Pipelines in Bioinformatics for medical research (De Pedro and Sánchez (2010b)). The communication will highlight the main advantages (and disadvantages) found up to date with the use of Tiki + PluginR to solve many of the needs of our research groups, including the new progresses achieved after the presentation at the last Spanish R Users meeting (De Pedro and Sánchez (2010a))

References

De Pedro, X. (2011). Tiki documentation: Plugin r. https://doc.tiki.org/PluginR.

De Pedro, X., M. Calvo, A. Carnicer, J. Cuadros, and A. Miñarro (2010). Assessing student activity through log analysis from computer supported learning assignments. In *Proceedings of the International Congress of University Teaching and Innovation*. http://cochise.bib.ub.es.

De Pedro, X. and A. Sánchez (2010a). Usando de forma segura r vía web con tiki. In *II Jornadas de Usuarios de R en Castellano*. http://r-es.pangea.org/II+Jornadas.

De Pedro, X. and A. Sánchez (2010b). Using r in tiki for bioinformatics (*poster*). In *Xth Spanish Symposium on Bioinformatics (JBI2010)*. http://estbioinfo.stat.ub.es/?p=219.

Nakano, J. and E.-j. Nakama (2009). Web interface to r for high-performance computing. In *The R UseR Conference 2009, Rennes*.

Ooms, J. (2009). Building web applications with r. In *The R UseR Conference 2009, Rennes*.

R (2011). R faq. cran.r-project.org/doc/FAQ/R-FAQ.html#R-Web-Interfaces.

Saunders, N. (2009). A brief survey of r web interfaces. http://nsaunders.wordpress.com/2009/11/30/a-brief-survey-of-r-web-interfaces/.

Tiki (2011). Development model. https://tiki.org/Model.

Browser Based Applications Supported by R in Pipeline Pilot

DaveNicolaides[1,*], Noj Malcolm[1], Stephane Vellay[1], Dana Honeycutt[2], Tim Moran[2]

1. Accelrys Ltd., 334 Cambridge Science Park, Milton Road, Cambridge UK, CB4 0WN
2. Accelrys Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA
*Contact author: dave@accelrys.com

Examples of the use of R as the "statistical engine" behind a simplified interface designed to be used by scientists abound (Bajuk-Yorgan, Kaluzny, 2010; Weiss 2008; Neuwirth 2008; there are perhaps another dozen examples in the useR! meetings over the last two years alone).

We wish to share the success we've had in deploying R more widely to the research community in standard browsers, via our Pipeline Pilot informatics platform (Hassan et al. 2006). While there are similarities between our approach and those referred to previously, we think that a reliance on the scientific domain knowledge increasingly to be found in R packages distinguishes our approach. We will give specific examples of the use of R behind the scenes, in such diverse areas of research as pharmaceutical discovery, automated image classification, gene expression, heterogeneous catalysis, and process analytical technology.

The success of these activities have largely stemmed from their approach to integration not as an activity based on software technologies, but instead based on people and the ways they work and learn. Allowing scientists to leverage a sophisticated statistics engine in a familiar interface gives them greater insight into and confidence in their decisions. In seeking to extend this success we will have to address many complex issues, including:

- distinguishing transfer of statistical expertise (where the goal is learning) from mere automation (where the goal is research efficiency), and
- understanding how we might "enable ideas as software" in R, where those ideas already have a strong presence in our own and other commercial software products.

References

Lou Bajuk-Yorgan, Stephen Kaluzny (2010). Making R accessible to Business Analysts with TIBCO Spotfire. In *useR! 2010, The R User Conference (Gaithersburg, Maryland)*, Book of Abstracts p. 14.

ChristianWeiss (2008). Commercial meets Open Source - Tuning STATISTICA with R.In *useR! 2008, The R User Conference (Dortmund, Germany)*, Book of Abstracts p. 188.

Erich Neuwirth (2008). R meets the Workplace - Embedding R into Excel and making it more accessible.In *useR! 2008, The R User Conference (Dortmund, Germany)*, Book of Abstracts p. 135.

Hassan M, Brown RD, Varma-O'brien S and Rogers D. (2006). Cheminformatics analysis and learning in a data pipelining environment. *Moleculardiversity* 10(3), 283-99.

A new task-based GUI for R

Sheri Gilley[1,*]

1. Principal UI Designer, Revolution Analytics
*Contact author: sheri@revolutionanalytics.com

Keywords: GUI, Revolution R, user interface design

One of Revolution Analytics' main goals is to make data analysis done in *R* accessible to a wider community of people. To that end, we have designed a graphical user interface for *R*.

Last year at useR!, I presented[1] the design methodology used to develop a GUI for *R*. This year, I will follow up on that theme by discussing some of the issues discovered during usability testing at various stages in the development. A demo of the soon-to-be released product will show how these issues were addressed, including:

- A task-based interface with an easy way to browse through the dialogs to find the task you want to perform.
- Task dialogs to allow users to generate and run R code. This can be done without ever seeing the code, although a user may click a button to reveal the code at any time.
- Script editor to write code that does not have a corresponding dialog.
- Workspace explorer to help you to learn more about your data as well as make it easy to drop variables into a task or into code written in the script editor.
- Output and history also organized by the tasks.

References

1. Gilley (2010). Designing a flexible GUI for R. http://user2010.org/abstracts/Gilley.pdf.

Computational aspects of continuous-time-arma (CARMA) models: The `ctarma` package

Helgi Tomasson[1]

1. University of Iceland, Faculty of Ecnomics

*Contact author: helgito@hi.is

Keywords: Estimation, simulation, continuous-time, ARMA models.

Some computational aspects of the continuous-time ARMA, CARMA are reviewed. Methods of simulation and estimation have been implemented in an *R*-package, `ctarma`. The simulations can be either frequency-domain based or time-domain based. Several approaches of simulating CARMA processes with time- and frequency-domain methods are implemented in the package. The estimation is based on numerically maximizing the normal likelihood. The likelihood is computed via the Kalman-filter algorithm. The process is constrained to be stationary by transforming the parameter space. Two alternative transformations enforcing the stationary conditions on the parameter space are given. The package can model irregularly spaced time series. Starting values of the parameters can be generated by using the Whittle-estimator when a usable empirical spectrum is available. A scheme of increasing the size of the model, i.e., generating a CARMA(p+1,q+1) model from a CARMA(p,q) is built into the package. The functionality of the package is illustrated with simulations and some real-data series.

robKalman—An *R* package for robust Kalman filtering revisited

Bernhard Spangl[1,*], **Peter Ruckdeschel**[2]

1. University of Natural Resources and Life Sciences, Vienna
2. Fraunhofer ITWM, Department of Financial Mathematics, Kaiserslautern
*Contact author: bernhard.spangl@boku.ac.at

Keywords: EM algorithm, Kalman filter, recursive nonlinear filter, robustness

Building up on talks on this issue given at previous Use*R*! conferences, we report on progress made in the development of the package **robKalman**. The focus of this talk will be on

- (robust) Kalman filtering and smoothing
  - enhancing the functionality allowing for time-invariant and time-variant hyper parameters, even functions
  - but keeping the easy extensibility and modular approach of the general recursive filtering and smoothing infrastructure
- (robust) estimation of hyper parameters
  - via the EM Algorithm and its robustification
- (robust) recursive nonlinear filtering implementations, namely,
  - the extended Kalman filter and
  - the unscented Kalman filter

References

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

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org/.

Ruckdeschel, P. (2010, May). Optimally Robust Kalman Filtering. Technical Report 185, Fraunhofer ITWM Kaiserslautern, Fraunhofer Platz 1, 67663 Kaiserslautern, Germany. http://www.itwm.fraunhofer.de/fileadmin/ITWM-Media/Zentral/Pdf/Berichte_ITWM/2010/bericht_185.pdf.

Ruckdeschel, P. and B. Spangl (2010). *robKalman: Robust Kalman Filtering*. R package version 0.3, http://robkalman.r-forge.r-project.org.

Shumway, R. and D. Stoffer (2000). *Time Series Analysis and Its Applications*. New York: Springer.

Spangl, B. (2008). *On Robust Spectral Density Estimation*. Ph. D. thesis, Dept. of Statistics and Probability Theory, Vienna University of Technology, Vienna.

Wan, E. and R. van der Merwe (2001). The unscented kalman filter. In Haykin (Ed.), *Kalman Filtering and Neural Networks*, Chapter 7, pp. 221–280. New York: Wiley.

(Robust) Online Filtering in Regime Switching Models and Application to Investment Strategies for Asset Allocation

Christina Erlwein[1], **Peter Ruckdeschel**[1,⋆]

Fraunhofer ITWM, Dept. of Fin. Mathematics, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany

⋆Contact author: peter.ruckdeschel@itwm.fraunhofer.de

Keywords: robustness, Hidden Markov Models, filtering, substitutive outliers, asset allocation

We implement to *R*, an online filtering algorithm for Hidden Markov Models with conditionally Gaussian observations by Elliott (1994), and are currently about to package this functionality to a new *R* package **robHMM**.

This algorithm consists of several steps: It involves a change of measure to an equivalent measure under which we have independence as well as a filtering and and a (ML-) parameter estimation step where the last two steps form an EM-algorithm.

The algorithm is modularized correspondingly such that in each step the respective function realizing it may easily be replaced by a suitable alternative (robust) function.

We study the vulnerability of each of these steps against substitutive outliers and propose corresponding robust alternatives extending Ruckdeschel (2010).

In a similar setting as in Erlwein et al. (2009), we apply this robustified algorithm to investment strategies for asset allocation with the rationale to better handle possible peaks or missings in asset returns, limiting their impact on optimal parameter estimates. The parameter estimates obtained are in turn used to make investment decisions within a regime-switching framework.

References

Elliott, R. (1994). Exact adaptive filters for markov chains observed in gaussian noise. *Automatica 30*, 1399–1408.

Erlwein, C., R. Mamon, and M. Davison (2009). An examination of hmm-based investment strategies for asset allocation. *Applied stochastic models in business and industry*. DOI: 10.1002/asmb.820.

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.

Ruckdeschel, P. (2010, May). Optimally Robust Kalman Filtering. Technical Report 185, Fraunhofer ITWM Kaiserslautern, Fraunhofer Platz 1, 67663 Kaiserslautern, Germany. http://www.itwm.fraunhofer.de/fileadmin/ITWM-Media/Zentral/Pdf/Berichte_ITWM/2010/bericht_185.pdf.

Using R for the Analysis of Bird Demography on a Europe-wide Scale

Christian Kampichler[1,2], Henk van der Jeugd[1], Alison Johnston[3], Rob Robinson[3], Stephen Baillie[3]

[1] Vogeltrekstation – Dutch Centre for Avian Migration and Demography, Netherlands Institute of Ecology, Wageningen, Netherlands;
[2] División de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco, Villahermosa, Mexico;
[3] British Trust for Ornithology, Thetford, U.K.
Contact: c.kampichler@nioo.knaw.nl

Bird populations are not constant in time but experience variation due to internal dynamics or to the action of external forces such as climatic change or habitat destruction. They are often used as indicators for the trends of wider biodiversity because our knowledge on them is generally good [1]. Beginning in the early 1980s, a Europe-wide network of more than 400 constant effort sites operating in 12 countries has been established, where birds are captured and ringed with a unique mark, according to a standardised protocol [2]. This scheme allows the monitoring of demographic parameters (abundance, reproduction rate, survival rate) of a large number of songbird species. The European database currently holds records of more than five million captures and recaptures.

Annual and site-specific reproduction rates are calculated by GLM based on the ratio of occurrences of adult and juvenile birds per site and year. The estimation of survival rates is somewhat more sophisticated and requires GLM with multinomial response variables for the application of specific capture-recapture models [3]. Basically, input data consist of a capture history for each ringed specimen of a given species where years are coded as 1 (specimen captured) and 0 (specimen not captured). Each possible capture history—for example, 111, 110, 101 or 100 for a three-years mark-recapture study—has a corresponding probability which is composed of the probability that a bird has survived from the preceding year (Φ) and the probability it is captured conditional on having survived (p). For the four capture histories above, the probabilities are $\Phi_1 p_2 \Phi_2 p_3$, $\Phi_1 p_2 (1 - \Phi_2 p_3)$, $\Phi_1 (1 - p_2) \Phi_2 p_3$ and $1 - \Phi_1 p_2 - \Phi_1 (1 - p_2) \Phi_2 p_3$. The capture history probabilities are weighed by their frequency in the dataset and the parameters are estimated using maximum likelihood or Markov chain Monte Carlo methods.

The most widely used application currently available for survival analysis using data from marked individuals is Program MARK [4] (URL http://www.phidot.org/software/mark/). On the one hand, it is extremely comprehensive, offering a wide range of different models. On the other hand, its data and model manipulation facilities are rather restricted—they are based on a conventional GUI with numerous pull-down menus and check-box options—making the repeated analyses of many species (or for many locations) a cumbersome and extremely error-prone task. We thus use RMark [5] (URL http://www.phidot.org/software/mark/rmark/) as an interface, allowing for the use of the powerful formula and design matrix functions in R. A key advantage of R for monitoring is that one does not need to recreate the analyses each year with extra data. Having a set of automated scripts makes producing trends easy when resources are often tight. Furthermore, the possibility of using and exchanging scripts facilitates the communication among the various institutions running national constant effort site programmes and assures a unified analytical approach.

References

[1] Gregory RD, van Strien A, Vorisek P, Gmelig Meyling AW, Noble DG, Foppen RPB, Gibbons DW, 2005. Developing indicators for European birds. *Philosophical Transactions of the Royal Society B* **360**, 269–288

[2] Robinson RA, Julliard R, Saracco JF, 2009. Constant effort: studying avian population processes using standardised ringing. *Ringing & Migration* **24**, 199-204

[3] Amstrup S, MacDonald L, Manly B, 2006. *Handbook of Capture-Recapture Analysis*. Princeton University Press

[4] White GC, Burnham KP 1999. Program MARK: Survival estimation from populations of marked animals. *Bird Study* **46** Supplement, 120-138

[5] Laake J, 2010. RMark: R Code for MARK Analysis. R package version 1.9.6.

Using OpenBUGS and lmer to study variation in plant demographic rates over several spatial and temporal scales

Johan P. Dahlgren

Department of Botany, Stockholm University, Stockholm, Sweden.
Contact author: johan.dahlgren@botan.su.se

Keywords: Bayesian lasso, Ecology, Hierarchical OpenBUGS model, lmer, Population dynamics

I am using the **lme4** and **R2OpenBUGS** *R* packages to study patterns in temporal and spatial variation in vital rates of the forest herb *Lathyrus vernus*. The combined effect of vital rate variation on population dynamics is examined in an Integral Projection Model (IPM), also run in *R*. Hierarchical plant demographic studies only recently became practically possible and have rarely been performed previously (Buckley et al. 2003), but available modeling techniques now enable such approaches (McMahon & Diez 2007, Evans et al. 2010). Understanding where variation occurs can shed much light on what factors affect plant demography. In contrast to previous studies I also use *Openbugs* and `lmer` to examine at what spatial scales temporal variation occurs. In addition to this study of variance components, I attempt to determine which out of a large set of environmental factors that were measured at the plot level have the largest effect on population dynamics. This is done by including Bayesian lasso components into the *OpenBUGS* vital rate models (Tibshirani 1996, Yi & Xu 2008).

Preliminary analyses suggest that most of the variation of all vital rates occur at the scale o f individuals, as opposed to plot or sub-population scales. Temporal variation seems to be equally distributed over spatial scales. Environmental factors have differing effects different years and are generally weak. This is in accordance with the fact that little variation occurred at the plot scale, where environmental factors were measured. I discuss the potential of density dependence to cause observed patterns.

On the technical side, I conclude that apart from the obvious "limitations" of a specialized function of not easily allowing utilization of techniques such as the Bayesian lasso used here, at least two aspects of lmer should make it an often more desirable choice than OpenBUGS for ecologists planning similar analyses (see also Gelman & Hill 2007). First, as seen from comparisons with our *BUGS* models, it is very accurate - even when estimating variance components for factors with few levels. Second, it is much faster than OpenBUGS, not only in computation time of correctly specified models, but because of the time it can take to specify a working *BUGS* model.

References

Buckley, Y.M., D.T. Briese & M. Rees (2003). Demography and management of the invasive plant species *Hypericum perforatum*. I. Using multi-level mixed-effects models for characterizing growth, survival and fecundity in a long-term data set. *Journal of Applied Ecology 40*, 481–93.

Evans, M.E.K., K.E. Holsinger & E.S. Menges (2010). Fire, vital rates, and population viability: a hierarchical Bayesian analysis of the endangered Florida scrub mint. *Ecological Monographs 80*, 627–49.

Gelman, A. & J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. University Press, Cambridge.

McMahon, S.M. & J.M. Diez (2007). Scales of association: hierarchical linear models and the measurement of ecological systems. *Ecology Letters 10*, 437–52.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, 58, 267–88.

Yi, N. & S. Xu (2008). Bayesian LASSO for Quantitative Trait Loci Mapping. *Genetics 179*, 1045-55.

An effort to improve nonlinear modeling practice

John C. Nash[1] with the NCEAS Nonlinear Modeling Working Group

1. Telfer School of Management (retired), University of Ottawa, Ottawa, ON K1N 6N5 Canada
Contact author: nashjc@uottawa.ca

Keywords: nonlinear modeling, optimization, BUGS

Funded by the National Center for Ecological Analysis and Synthesis, the Nonlinear Modeling Working Group is an assembly of approximately 20 researchers from a number of countries who are meeting for two 1-week working seminars, one in January 2011 and the other scheduled for late July 2011. At the first meeting, we chose a suite of test problems of an ecological nature requiring nonlinear models to be estimated. Each problem was then approached using open source modeling tools, in particular, *R*, *AD Model Builder,* and an open version of BUGS (either *OpenBUGS* or *JAGS*). This work will be continued at the second meeting and individually outside the formal events.

The goal of the Working Group (http://www.nceas.ucsb.edu/projects/12602) is to document the effort to build solutions in each package, to provide a set of ecological nonlinear modelling problems that are well-described, to provide advice on how to carry out the process, and to assess where each package has advantages or disadvantages with a view to improving the tools.

This paper will aim to share our preliminary findings and look to enlarge the activity and understanding of benchmarking not only software but the human process of problem-solving.

brglm: Bias reduction in generalized linear models

Ioannis Kosmidis[1,⋆]

1. Department of Statistical Science, University College, Gower Street London, WC1E 6BT, London, United Kingdom
⋆Contact author: ioannis@stats.ucl.ac.uk

Keywords: `glm` fitting methods, asymptotic bias reduction, adjusted score functions, Fisher scoring.

The **brglm** *R* package provides an alternative fitting method for the `glm` function for reducing the bias of the maximum likelihood estimator in generalized linear models (GLMs). The fitting method is based on the generic iteration developed in Kosmidis and Firth (2010a) for solving the bias-reducing adjusted score equations (Firth, 1993). It relies on the implementation of the first-order term in the asymptotic expansion of the bias of the maximum likelihood estimator for GLMs which has been derived in Cordeiro and McCullagh (1991). The bias-corrected estimates derived in the latter study are by-products of the general fitting method.

The benefits of reducing the bias in estimation are discussed, especially in models for discrete responses. Specifically, in such models there is a positive probability that the maximum likelihood estimate has infinite components, which can potentially cause problems in the use of standard inferential procedures. In contrast, for many well-used GLMs, the reduced-bias estimates have been found to always have finite components, motivating their study and use in practice (see, for example, Firth 1992; Mehrabi and Matthews 1995; Heinze and Schemper 2002; Bull et al. 2002; Zorn 2005; Kosmidis 2009; Kosmidis and Firth 2010b).

The **brglm** package also provides methods for the construction of confidence intervals through the profiles of appropriately constructed inference functions (Kosmidis, 2008; Kosmidis and Firth, 2009).

References

Bull, S. B., C. Mak, and C. Greenwood (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis 39*, 57–74.

Cordeiro, G. M. and P. McCullagh (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B: Methodological 53*(3), 629–643.

Firth, D. (1992). Bias reduction, the Jeffreys prior and GLIM. In L. Fahrmeir, B. Francis, R. Gilchrist, and G. Tutz (Eds.), *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM 92 Conference, Munich*, New York, pp. 91–100. Springer.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*(1), 27–38.

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

Kosmidis, I. (2008). The profilemodel R package: Profiling objectives for models with linear predictors. *R News 8/2*, 12–18.

Kosmidis, I. (2009). On iterative adjustment of responses for the reduction of bias in binary regression models. Technical Report 09-36, CRiSM working paper series.

Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika 96*(4), 793–804.

Kosmidis, I. and D. Firth (2010a). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics 4*, 1097–1112.

Kosmidis, I. and D. Firth (2010b). Multinomial logit bias reduction via poisson log-linear model. Technical Report 10-18, CRiSM working paper series. Accepted for publication in Biometrika.

Mehrabi, Y. and J. N. S. Matthews (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics 51*, 1543–1549.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis 13*, 157–170.

Large Scale, Massively Parallel Logistic Regression in R with the Netezza Analytics Package

C.Dendek, M. Klopotek, P. Biecek
Netezza Corporation, an IBM Company

April 4, 2011

Abstract

One of the important limitations of the standard *glm* procedure for estimation of the parametres of logistic regression model in R is the size of the data that is kept in the memory, i.e. the original sample and algorithm–specific temporary data, effectively restricting both cardinality and dimensionality of the sample.

The *biglm* package makes possible to overcome the restriction on the number of observations present in the sample. But it still leaves the dimensionality limitation, due to the second–order algorithm being used to fit the models.

The possiblity of use of the first–order, stochastic gradient–descent optimization method creates a tradeoff between the rate of convergence and the maximal dimensionality of the sample. Interestingly, in case of smoothly regularized logistic regression (e.g. L2–based, rigde estimate), it is possible to parallelize the first–order method w.r.t. data sample, greatly improving the computation time of a single iteration and – in practice – reducing the adventage of second–order method.

The gradient-based approach outlined above has been implemented in the Netezza Analytics package using Netezza Performance Server as a database

The binomTools package: Performing model diagnostics on binomial regression models

Merete K Hansen[1,*], Rune Haubo B Christensen[1]

1. Technical University of Denmark, Department of Informatics and Mathematical Modelling, Section for Mathematical Statistics, Richard Petersens Plads, Building 305, Room 122, DK-2800 Kgs. Lyngby, Denmark
*Contact author: mkh@imm.dtu.dk

Keywords: Binomial regression models, diagnostics

Binomial regression models are widely used for modelling observations with dichotomous outcome. Since diagnostics are important for validation of model adequacy, the accessibility of suitable diagnostic methods for binomial regression models is crucial. The **binomTools** package (Hansen and Christensen, 2011) provides a range of diagnostic methods for this class of models extending the basic tool set in base *R* and enabling a thorough examination of the fitted model.

Appropriate use of the deviance or Pearson's X^2 reported by `glm` for goodness-of-fit assessment requires that observations are suitably grouped, e.g., binary observations should be grouped to binomial form. Our `group` method for glm objects performs this otherwise cumbersome task. Additional goodness-of-fit tests for binary data available in **binomTools** include the Hosmer-Lemeshow test.

Residual analysis in binary and binomial models is complicated by the non-unique definition of residuals and the sparseness of appropriate graphical methods. `rstudent` from **stats** provide approximate deletion residuals, the so-called likelihood residuals (Williams, 1987); **binomTools** enhance this by providing the exact counterparts. The half-normal plot with simulated envelopes implemented in our `halfnorm` function is described by Collett (2003) and one of the most effective and sensitive residual plots for the notoriously difficult binary models. Further, the *coefficient of discrimination* (Tjur, 2008) and related figures are implemented for summarizing and describing predictive performance.

Our aim with **binomTools** is to provide a comprehensive collection of methods for residual analysis, model diagnostics and presentation of results from binomial regression models.

References:

Collett, D. (2003). *Modelling binary data* (Second edition). Chapman & Hall/CRC.

Hansen, M. K. and Christensen, R. H. B. (2011). binomTools – diagnostic tools for binomial regression models. 1.0.1. https://r-forge.r-project.org/projects/binomtools/.

Tjur, T. (2008). Coefficients of determination in logistic regression models – a new proposal: The coefficient of discrimination. *The American statistician 63*(4), 366-372.

Williams, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics 36*, 181-191.

”uniPlot” – A package to uniform and customize R graphics

Sina Rüeger[1,⋆]

1. Institute of Data Analysis and Process Design (Zurich University of Applied Sciences)
⋆Contact author: sina.rueeger@gmail.com

Keywords: uniform layout, customizing graphics, visual perception

Three packages widely used for producing graphics in *R* are **graphics**, **ggplot2** and **lattice**. Each has its own strengths and limitations. However, they all differ in design. This is undesirable if graphics produced by different packages are used in the same report. Therefore the package **uniPlot**, which uniforms graphics produced by **graphics**, **ggplot2** or **lattice**, is presented.

Daily *R* users who use the program to graph data within a particular data analysis are familiar with differences in appearance; they can extract information easily and mask the layout differences. Readers unfamiliar with *R* will be distracted when interpreting three different graphics made by **graphics**, **ggplot2** and **lattice** graphics because variation in layout affects the decoding of information. Therefore, a way should be found to produce consistent and calm graphics. In *R* this can be solved by adjusting several parameters, options or themes. Everyone who has ever prepared plots by adjusting parameters, options and themes knows that this is cumbersome and time consuming.

The package **uniPlot** solves this problem. It (a) designs an uniform look of graphics and (b) adapts graphics from packages **graphics**, **ggplot2** and **lattice** to achieve this uniform look. The idea is that the user would load different layout-styles by a simple function. Changes of settings and parameters could be made within this function. Restoring of the settings would be possible.

A first version of **uniPlot** is currently under development and will be available at CRAN by the end of July 2011.

sparkTable: Generating Graphical Tables for Websites and Documents with R

Alexander Kowarik[1,*], **Bernhard Meindl**[1] **and Matthias Templ**[1,2]

1. Statistics Austria, Guglgasse 13, 1110 Vienna, Austria
2. Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria
*Contact author: alexander.kowarik@statistik.gv.at

Keywords: Visualisation, Graphical Tables, Sparklines

With the R package **sparkTable** (Kowarik et al. (2010)) tables presenting quantitative information can be enhanced by including sparklines and sparkbars (initially proposed by Tufte (2001)). Sparklines and sparkbars are simple, intense and illustrative graphs, small enough to fit in a single line. Therefore they can easily enrich tables and continuous texts with additional information in a comprehensive visual way.

Another feature of **sparkTable** is the optimal allocation of geographical units to cells in a 2-dimensional table and the presentation of graphical and quantitative information within these cells. For example, graphical tables including information about countries with each country representing a cell in a 2-dimensional grid can be generated. Hereby, the position of countries in a thematic map are projected to the given grid in an optimal manner by solving a linear program.

The usage of `S4`-classes provide an easy and fast way to create fancy output for websites, presentation and documents. The output is presented and explained with real-world applications.

References

Kowarik, A., B. Meindl, and S. Zechner (2010). sparktable: Sparklines and graphical tables for tex and html. `http://CRAN.R-project.org/package=sparkTable`.

Tufte, E. R. (2001). *Visual Display of Quantitative Information*. Graphics Press.

compareGroups package, updated and improved

Héctor Sanz[2,1⋆], **Isaac Subirana**[3,1,4], **Joan Vila**[1,3]

1. Cardiovascular Epidemiology & Genetics group, Inflammatory and Cardiovascular Disease Programme, IMIM, Hospital del Mar Research Institute, Spain
2. UCICEC CAIBER. IMIM-Hospital del Mar
3. CIBER Epidemiology and Public Health (CIBERESP), Spain
4. Statistics Department, University of Barcelona, Spain

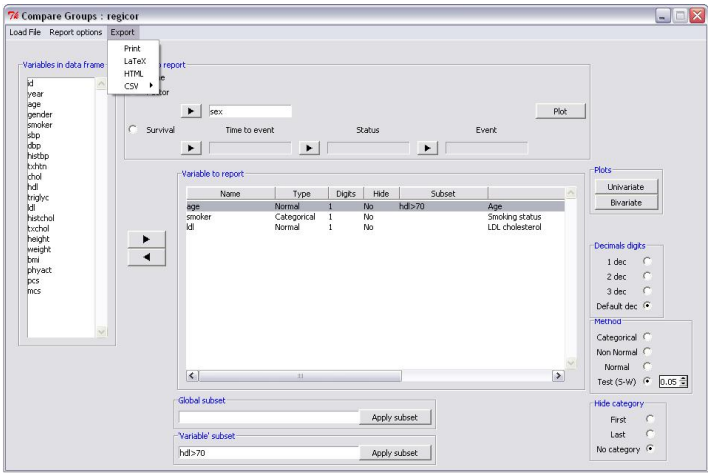
⋆Contact author: hsanz@imim.es

Keywords: Software Design, Bivariate Table, LaTeX, Descriptive Analysis

In many studies, such as epidemiological ones, it is needed to compare characteristics between groups of individuals or disease status. Usually these comparisons are presented in the form of tables (also called Bivariate Tables) of descriptive statistics where rows are characteristics, and each column is a group / status. Usually the number of characteristics is large, and thus construction of these tables is laborious.

To build them in an easy, quick and efficient way, we created the **compareGroups** package [1]. Here we present package improvements and extensions in the following issues:

- descriptives by status in a cohort study allowing to include right-censored time-to-response
- descriptives for groups or the entire sample.
- exporting tables to HTML format
- new GUI aspect with a single frame and a main menu
- importing data from workspace using the GUI
- subset specifically for each variable using the GUI
- incidence for right-censored time-to-event row-variables.
- extended and improved vignette
- exporting tables to LaTeX under the `longtable` environment
- proper plots for survival analysis



Var	Male N=1101	Female N=1193	p.overall
Age	54.8 (11.1)	54.7 (11.0)	0.840
Smoking status:			<0.001
Never smoker	301 (28.1%)	900 (77.5%)	
Current or former < 1y	410 (38.3%)	183 (15.7%)	
Never or former >= 1y	360 (33.6%)	79 (6.80%)	
Systolic blood pressure	134 (18.9)	129 (21.2)	<0.001
Diastolic blood pressure	81.7 (10.2)	77.8 (10.5)	<0.001
History of hypertension	341 (31.1%)	382 (32.1%)	0.644
HTN treatment	189 (17.5%)	239 (20.4%)	0.096
Total cholesterol	217 (42.7)	220 (47.4)	0.140
HDL cholesterol	47.5 (12.6)	57.5 (15.0)	<0.001
Triglycerides	131 (87.4)	101 (55.2)	<0.001
LDL cholesterol	145 (38.5)	142 (40.7)	0.092
Hystory of hypercol	353 (32.3%)	356 (30.2%)	0.308

References

1. Hector Sanz, Isaac Subirana, Joan Vila (2010) "Bivariate Analyses" UseR! 2010, The R User Conference 2010, (National Institute of Standards and Technology, Gaithersburg, Maryland, US), July 2010.

Six Sigma Quality Using R: Tools and Training

Emilio López[1,⋆], **Andrés Redchuk**[1], **Javier M.Moguerza**[1]

1. Department of Statistics and Operations Research. Rey Juan Carlos University (Madrid)
⋆Contact author: elopez@proyectum.es

Keywords: Six Sigma, Process Improvement, Engineering Statistics, Quality Control, Lean Six Sigma

Six Sigma is a known methodology for Process and Quality Improvement. It is also a philosophy, and a *set of tools*. It is based on the methodology **DMAIC** (Define, Measure, Analyze, Improve, Control). There are other business-process management methodologies related to Six Sigma, such as DFSS (Design For Six Sigma) or Lean Manufacturing (Lean Six Sigma).

Six Sigma is notable for using the Scientific Method, and Statistical Techniques. Some of the statistical tools that are used in Six Sigma projects are:

- Graphic Analysis: Pareto Charts, Histograms, Scatterplots, Box-Whisker Charts, Group Charts, Location Charts, Control Charts, Multivari Charts
- Design of Experiments
- Regression and Analysis of Variance (ANOVA)
- Confidence Intervals, Hypothesis Testing
- Gage R&R Studies
- Acceptance Sampling, Capability Analysis, Reliability Analysis

Commercial statistical software usually includes specific options for Quality Management, for example control charts. There are also a couple of contribution packages in *R* regarding control charts (**qcc**, **IQCC**) but there is no a complete set of tools for *Six Sigma*.

We are currently working in several initiatives aimed at **explain and facilitate Six Sigma practitioners to carry on their Six Sigma projects with *R***, such as a **SixSigma** package (already at `CRAN` repositories) with some functions deployed and many others in mind, as well as an on-line training course within the EC Lifelong Learning Programme VRTUOSI project, and several publications in preparation.

References

Allen, T. T. (2010). *Introduction to Engineering Statistics and Lean Six Sigma - Statistical Quality Control and Design of Experiments and Systems*. Springer.

Box, G. (1991). Teaching engineers experimental design with a paper helicopter. Report 76, Center for Quality and Productivity Improvement. University of Wisconsin.

Chambers, J. M. (2008). *Software for data analysis. Programming with R*. Statistics and Computing. Springer.

ISO (2009). *ISO/TS 16949: Quality management systems – Particular requirements for the application of ISO 9001:2008 for automotive production and relevant service part organizations*. International Organization for Standardization.

Montgomery, D. (2005). *Introduction to Statistical Quality Control* (5^{th} ed.). New York: Wiley.

Murrell, P. (2005). *R Graphics*. Chapman & HallCRC.

Process Performance and Capability Statistics for Non-Normal Distributions in *R*

Thomas Roth[1*]

The Department of Quality Science - Technical University of Berlin

*Contact author: Thomas Roth

Keywords: Distribution Identification, Process Capability, Process Performance

Estimating performance and capability statistics is a widespread practice in industry to assess and communicate the state of (manufacturing) processes. The calculation of these statistics is among others defined in the international standard ISO 21747:2006. Although the provided definitions are generic (i.e. distribution independent), calculation of these statistics is mainly based on the assumption of a normally distributed process. This practice causes difficulties with regularly occurring non-normally distributed process characteristics.

Methods are presented for conveniently fitting distributions and subsequently assessing process performance and capability regardless of the underlying distribution of the measured characteristic. In addition an insight into the terminology of the ISO 21747:2006 document is provided.

References

ISO (2005). Quality management systems - Fundamentals and vocabulary (ISO 9000:2005).

ISO (2007). Statistical methods - Process performance and capability statistics for measured quality characteristics (ISO 21747:2006).

Mittag, H.-J. and H. Rinne (1999). *Prozessfähigkeitsmessung für die industrielle Praxis*. München: Hanser.

Montgomery, D. C. (2005). *Introduction to statistical quality control* (5 ed.). Hoboken, N.J: John Wiley.

Roth, T. (2010). qualityTools: Statistical Methods for Quality Science.

R-Package JOP: Optimization of Multiple Responses

Nikolaus Rudak[1], **Sonja Kuhnt**[1]

1. Faculty of Statistics, TU Dortmund University, Germany
*Contact author: rudak@statistik.tu-dortmund.de

Keywords: Multiple responses, Robust parameter design, Simultaneous optimization, Pareto Optimality.

In technical applications, often a set of response variables with corresponding target values depends on a number of control variables. In these cases an off-line quality control prior to the actual manufacturing process frequently implies optimizing the mean as well as minimizing the variance of the responses.

Many existing methods for the optimization of multiple responses require some kind of weighting of these responses, for instance in terms of costs or desirabilities. Kuhnt and Erdbruegge (2004) present an alternative strategy using loss functions and a penalty matrix which can be decomposed into a standardizing and a weight matrix. The Joint Optimization Plot displays the effect of different weight matrices in terms of predicted response means and variances. Furthermore Erdbruegge et al. (2011) show that every point that minimizes the conditional mean of the loss function is Pareto optimal.

The new *R* package **JOP** (Kuhnt and Rudak, 2011) is an implementation of the Joint Optimization Plot and is available on CRAN in the version 2.0.1. **JOP** includes an automated procedure to fit double generalized linear models by means of the *R* package **Joint Modeling** (see (Ribatet and Iooss, 2010)). For the optimization actual three different optimization routines can be chosen, depending on the complexity of the problem, more precisely by `nlminb`, `gosolnp` (see (Ghalanos and Theussl, 2011)) or `genoud` (see (Walter et al., 2009)). The resulting optimal responses together with corresponding settings of the control variables are displayed by the Joint Optimization Plot. **JOP** returns an object containing the optimal response and control variable values as well as the fitted double generalized linear models. Furthermore the user can choose a compromise with the mouse directly on the plot and **JOP** returns the corresponding optimal control variable settings.

We demonstrate the use of the Joint Optimization Plot in various applications from mechanical engineering.

References

Erdbruegge, M., S. Kuhnt, and N. Rudak (2011). Joint optimization of several responses based on loss functions. *submitted*.

Ghalanos, A. and S. Theussl (2011). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.0-8.

Kuhnt, S. and M. Erdbruegge (2004). A strategy of robust paramater design for multiple responses. *Statistical Modelling 4*, 249–264.

Kuhnt, S. and N. Rudak (2011). JOP: Joint optimization plot. R Package version 2.0.1.

Ribatet, M. and B. Iooss (2010). *JointModeling: Joint Modelling of Mean and Dispersion*. R package version 1.0-2.

Walter, R. M., Jr., and J. S. Sekhon (2009). Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software*. Forthcoming.

Density Estimation Packages in R

Henry Deng (Rice University)
Mentor: Hadley Wickham (Rice University)

Contact author: hd4@rice.edu

Density estimation is an important statistical tool, and within *R*, there are over 10 packages for density estimation, including **np**, **locfit**, and **KernSmooth**. At different times, it is often difficult to know which to use. In this project, we will summarize the results of our study comparing these packages. We will present a brief outline of the theory behind each package, as well as a description of the functionality and comparison of performance. Some of the factors we touch on are dimensionality, flexibility, and control over bounds.

The benchden Package: Benchmark Densities for Nonparametric Density Estimation

Henrike Weinert[1,⋆], **Thoralf Mildenberger**[2]

1. Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany
2. Department of Mathematics, University of Bayreuth, 95440 Bayreuth, Germany

⋆Contact author: weinert@statistik.tu-dortmund.de

Keywords: nonparametic statistics, density estimation, benchmark densities

This talk gives an introduction to the *R*-package **benchden** which implements a set of 28 example densities for nonparametric density estimation (Mildenberger and Weinert, 2011). In the last decades, nonparametric curve estimation has become an important field of research. To assess the performance of nonparametric methods aside from theoretical analysis often simulation studies are used. Indeed, most published articles suggesting a new method contain a simulation study in which the proposed method is compared to at least a few competitors. The comparison of methods on real-life data sets has the disadvantage that the correct solution is unknown. By using artificial data sets generated under a completely known mechanism one can overcome this disadvantage. Examples of such data sets are the Donoho-Johnstone functions *Blocks, Bumps, Doppler* and *HeaviSine* originally introduced in Donoho and Johnstone (1994). These four functions have well-known features like discontinuities or certain textures that resemble the difficulties encountered in specific applications.

In the case of density estimation there seems to be no generally used set of test densities. Our package **benchden** (Mildenberger et al., 2011) aims at closing this gap. The set of 28 test bed densities first introduced by Berlinet and Devroye (1994) is sufficiently large to cover a wide variety of situations that are of interest for the comparison of different methods. These densities also differ widely in their mathematical properties such as smoothness or tail behaviour and include some densities with infinite peaks that are not square-integrable. They include both densities from standard families of distributions as well as some examples specifically constructed to pose special challenges to estimation procedures.

The package contains the usual *R*-functions for densities that evaluate the density, distribution and quantile functions or generate random variates. In addition, a function which gives some information on special properties of the densities is included, which should be useful in large simulation studies.

References

Berlinet, A. and L. Devroye (1994). A comparison of kernel density estimates. *Publications de l'Institute de Statistique de L'Universite de Paris 38*, 3–59.

Donoho, D. L. and I. Johnstone (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika 81*, 425–455.

Mildenberger, T. and H. Weinert (2011). The **benchden** package: Benchmark densities for nonparametric density estimation. Discussion Paper 14/2011, SFB 823, Technische Universitt Dortmund.

Mildenberger, T., H. Weinert, and S. Tiemeyer (2011). *benchden: 28 benchmark densities from Berlinet/Devroye (1994)*. R package version 1.0.4.

Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions

Taylor B. Arnold[1,⋆]

1. Yale University, Department of Statistics
⋆Contact author: taylor.arnold@yale.edu

Keywords: Kolmogorov-Smirnov, Numerical Instability, Power Analysis

Goodness-of-fit tests are used to assess whether data are consistent with a hypothesized null distribution. The χ^2 test is the best-known parametric goodness-of-fit test, while the most popular nonparametric tests are the classic test proposed by Kolmogorov and Smirnov followed closely by several variants on Cramér-von Mises tests.

In their most basic forms, these nonparametric goodness-of-fit tests are intended for continuous hypothesized distributions, but they have also been adapted for discrete distributions by Conover (1972), Choulakian et al. (1994), and Gleser (1985). Unfortunately, most modern statistical software packages and programming environments have failed to incorporate these discrete versions. As a result, researchers would typically rely upon the χ^2 test or a nonparametric test designed for a continuous null distribution. For smaller sample sizes, in particular, both of these choices can produce misleading inferences.

We will present a revision of R's `ks.test()` function and a new `cvm.test()` function to fill this void for researchers and practitioners in the *R* environment. This work was motivated by the need for such goodness-of-fit testing in a study of Olympic figure skating scoring (Emerson and Arnold, 2010). We will first present overviews of the theory and general implementation of the discrete Kolmogorov-Smirnov and Cramér-von Mises tests. We then will discuss the particular implementation of the tests in R and provide examples paying close attention to the many numerical issues that arise in the implementation.

References

Arnold, T. B. and J. W. Emerson (2010). *ks.test: A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions*. R package version 1.0.

Choulakian, V., R. A. Lockhart, and M. A. Stephens (1994). Cramér-von Mises statistics for discrete distributions. *The Canadian Journal of Statistics 22*(1), 125–137.

Conover, W. J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association 67*(339), 591–596.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Emerson, J. W. and T. B. Arnold (2010, August). The power of human nature in a study of olympic figure skating. Unpublished manuscript.

Gleser, L. (1985). Exact power of goodness-of-fit tests of Kolmogorov type for discontinuous distributions. *Journal of the American Statistical Association 80*(392), 954–958.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika 48*, 419–426.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Instuto Italiano degli Attuari 4*, 83–91.

An algorithm for the computation of the power of Monte Carlo tests with guaranteed precision

Patrick Rubin-Delanchy[1], **Axel Gandy**[1,⋆]

1. Deparment of Mathematics, Imperial College London, SW7 2AZ, London, UK
⋆Contact author: a.gandy@imperial.ac.uk

Keywords: Monte Carlo testing, power of a test, sequential testing.

The power of a hypothesis test is the probability of rejecting the null hypothesis if the data follows a given distribution. For Monte Carlo tests, for example bootstrap or permutation tests, this quantity must usually be evaluated by simulation. N datasets are generated from the specified data distribution, a Monte Carlo test is performed on each, and the frequency with which the null hypothesis is rejected serves as an estimate of the power of the test.

The existing research into computing the power of Monte Carlo tests focuses on obtaining an estimate that is as accurate as possible for a fixed computational effort. However, the methods that are proposed have no guarantee of precision, in the sense that they cannot report a (non-trivial) confidence interval for the power with a certain coverage probability.

In this presentation, we describe an algorithm that runs N Monte Carlo tests simultaneously and indefinitely until a user-specified confidence interval length and coverage probability is met. We demonstrate that under some minor regularity conditions the algorithm terminates in finite expected time, and discuss some optimisation issues.

We are currently augmenting the *R* **simctest** package to incorporate this work. The user must input a confidence interval length and coverage probability, and a function that generates N binary streams, where each value is one if the simulated statistic under the null is at least as extreme as the base test-statistic for this stream. The method returns an estimate of the power with a confidence interval that meets the user-specified requirements.

Simple haplotype analyses in *R*

Benjamin French[1,*], Nandita Mitra[1], Thomas P Cappola[2], Thomas Lumley[3]

1. Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA USA
2. Penn Cardiovascular Institute, Philadelphia, PA USA
3. Department of Statistics, University of Auckland, Auckland, New Zealand
*Contact author: bcfrench@upenn.edu

Keywords: Statistical genetics, regression models

Statistical methods of varying complexity have been proposed to efficiently estimate haplotype effects and haplotype-environment interactions in case-control and prospective studies. We have proposed an alternate approach that is based on a non-iterative, two-step estimation process: first, an expectation-maximization algorithm is used to compute posterior estimates of the probability of all potential haplotypes consistent with the observed genotype for each subject; second, the estimated probabilities are used as weights in a regression model for the disease outcome, possibly including environmental factors. Standard error estimates are based on a robust variance estimator. We have shown that the two-step process provides valid tests for genetic associations and reliable estimates of modest genetic effects of common haplotypes for case-control studies (French et al, 2006). The two-step process has also been applied to prospective studies with a survival outcome subject to censoring (Neuhausen et al, 2009). An advantage of the two-step process is its straightforward implementation in software, so that analyses combining genetic and environmental information can be conducted by researchers expert in that subject matter using standard software, rather than by statisticians using specialized software. We illustrate the use of the two-step process for case-control studies using our *R* package **haplo.ccs**, which implements weighted logistic regression, and for prospective studies with a survival outcome using our working *R* package **haplo.cph**, which implements weighted Cox regression. We illustrate our method and software using data from a study of chronic heart failure patients (Cappola et al, 2011) to estimate the effect of *CLCNKA* haplotypes on time to death or cardiac transplantation.

References

Cappola TP, Matkovich SJ, Wang W, et al. (2011). Loss-of-function DNA sequence variant in the *CLCNKA* chloride channel implicates the cardio-renal axis in interindividual heart failure risk variation. *Proceedings of the National Academy of Sciences*, doi:10.1073/pnas.1017494108

French B, Lumley T, Monks SA, et al. (2006). Simple estimate of haplotype relative risks for case-control data. *Genetic Epidemiology* 30, 485–494.

Neuhausen SL, Brummel S, Ding YC, et al. (2009). Genetic variation in insulin-like growth factor signaling genes and breast cancer risk among *BRAC1* and *BRAC2* carriers. *Breast Cancer Research* 11, R76.

Mixed models of large pedigrees in genetic association studies

Jing Hua Zhao[1⋆]

1. MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK
⋆Contact author: jinghua.zhao@mrc-epid.cam.ac.uk

Keywords: mixed models, pedigree data, genetic linkage and association analysis

Standard approaches in genetic linkage and association analysis of data from large pedigrees have limited ability to handle covariates, which are important for studying their main effects and interactions with genetic markers. We desribed mixed models appropriate for analyzing such data. In particular, our motivating example was a simulated data distributed through the Genetic Analysis Workshop (GAW17) for which three types of mixed models have been fitted: a linear model of quantitative trait, a logistic model of binary trait, and a Cox model of binary trait and age at onset. The relevant functions `lmekin` for linear model and `coxme` for Cox model are from the *R* package **kinship** and able to accommodate kinship and identity-by-descent (IBD) information, while function `pedigreemm` from the *R* package **pedigreemm** allows for Gaussian, binary and Poisson responses with kinship but not IBD information. We compared these with procedures in *SAS*. We found that avaibility of a good IBD information can be useful for positional cloning and fine mapping in genome-wide association studies involving single nucleotide polymorphisms. We believe that our work will be of practical use to researchers in their analysis of pedigree data.

References

Epstein, M. P., J. E. Hunter, E. G. Allen, S. L. Sherman, X. Lin, and M. Boehnke (2009). A variance-component framework for pedigree analysis of continuous and categorical outcomes. *Statistics in Bioscience 1*, 181–198.

Therneau, T. (2007). On mixed-effect Cox models, sparse matrices, and modeling data from large pedigrees. http://mayoresearch.mayo.edu/mayo/research/biostat/upload/kinship.pdf.

Vazquez, A. L., D. M. Bates, G. J. M. Rose, D. Gianola, and K. A. Weigel (2010). Technical note: An R package for fitting generalized linear mixed models in animal breeding. *Journal of Animal Science 88*, 497–504.

Zhao, J. H. (2005). Mixed-effects Cox models of alcohol dependence in extended families. *BMC Genetics 6 (Suppl I)*, S127.

Graphical tools for assessing Hardy-Weinberg equilibrium for bi-allelic genetic markers

Jan Graffelman[1,⋆]

1. Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain
⋆Contact author: jan.graffelman@upc.edu

Keywords: Exact test, chi-square test, ternary plot, acceptance region, log-ratio transformation.

If there are no disturbing forces (migration, mutation, selection, etc.), then the genotype frequencies AA, AB, and BB of a bi-allelic genetic marker are expected to occur with frequencies p^2, $2pq$ and q^2 respectively, where p is the allele frequency of A, and $q = 1 - p$ is the allele frequency of B. This basic principle in genetics is known as Hardy-Weinberg equilibrium (HWE). In genetics, markers are statistically tested prior to their subsequent use in for instance, association studies. A significant deviation from HWE can be ascribed to many factors. Gross genotyping error can cause deviation from HWE and forms one of the reasons to test markers. The HapMap project (2007) excludes genetic markers that have a p-value below 0.001 in an exact test for HWE.

Several statistical tests are available to check markers for HWE: the χ^2 test, the exact test, the likelihood ratio test and Bayesian tests. Weir (1996, Chapter 4) gives an overview of several tests for HWE. Several of these tests are implemented in The *R* package **HardyWeinberg**. The package is currently being extended with routines for power calculation and Bayesian tests. Most testing for HWE is done in an entirely numerical manner. Graffelman and Morales (2008) showed that testing can be done graphically inside a ternary plot representation. The HWE law defines a parabola inside a ternary plot of the three genotype frequencies. An acceptance region for the different tests can be drawn around the HWE parabola. The graphical testing facilities are implemented in the function `HWTernaryPlot` of the *R* package **HardyWeinberg**. The obtained graphics are very informative because they display genotype frequencies, allele frequencies and the (statistical) equilibrium condition in a single graph.

The genotype counts can also be treated as three-way compositions that sum up to 1. Tools from the field of compositional data analysis yield alternative graphical representations of the HWE law. Several different log-ratio transformations (additive, centred or isometric) of the genotype counts can be used. Graphically, the HWE parabola in the ternary plot is converted into a straight line in log-ratio coordinates. Functions `HWalrPlot`, `HWclrPlot` and `HWilrPlot` of the **HardyWeinberg** package can be used to create the log-ratio plots for HWE. Zero genotype counts are a problem for the log-ratio approach, and some adjustment for zero counts is necessary.

The proposed graphics can be used to assess HWE for multiple samples that are all typed for one marker (e.g. cases and controls), but can also be used to screen many markers simultaneously. This way, a whole genomic region can be screened for an excess or lack of heterozygotes. Several genetic data sets will be used in the talk to illustrate the proposed graphs.

References

Graffelman, J. and J. Morales-Camarena (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity 65*(2), 77–84. DOI: 10.1159/000108939.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature 449*, 851–861.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, Massachusetts: Sinauer Associates.

Detecting Drug Effects in the Brain

Heather Turner[1*], Foteini Strimenopoulou[1], Phil Brain[1]

1. Pfizer Global Research & Development, UK
*Contact author: ht@heatherturner.net

Keywords: Multivariate analysis, Statistical modelling, Pharmaceutical research

In the early phases of drug development, the effect of a candidate drug on the brain can be of key importance, either because the drug is specifically targeted at the brain or because of safety concerns regarding an effect on the brain. Quantitative electroencephalography (qEEG) uses multiple electrodes placed on the scalp of a subject (typically rat or human) to record the electrical activity produced by the firing of neurons within the brain. By monitoring these "brainwaves" under different treatment conditions, the effect of a candidate drug can be inferred.

qEEG produces a virtually continuous signal over time that is typically processed using a Fast Fourier Transform, producing a set of power spectra at successive time slices for each subject. The challenge is to relate these power spectra (a multivariate response representing the brain activity), to the concentration of the drug in the brain at the corresponding time. Of course, the actual concentration of the drug in the brain at a given time cannot be measured, but can be modelled using pharmacokinetic models typically with unknown parameters.

In this talk we will introduce an R package for Extended Semi-linear Canonical Correlation Analysis (ESLCCA, described in Brain et al, 2011). ESLCCA estimates the parameters of a given nonlinear pharmacometric model to maximize the correlation with a linear combination of multiple response variables (in this case, the power spectra). We shall illustrate how this method has been used to characterize brain activity under different treatment regimens in research and development projects at Pfizer.

References

Brain, P, Strimenopoulou, F & Ivarsson, M (2011). Analysing electroencephalogram (EEG) data using Extended Semi-Linear Canonical Correlation Analysis. *Submitted.*

Statistical Parametric Maps for Functional MRI Experiments in *R*: The Package fmri

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

1. Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS) Leibniz-Institut im Forschungsverbund Berlin e.V.
2. DFG Research Center MATHEON "Mathematics for key technologies"

⋆Contact author: karsten.tabelow@wias-berlin.de

Keywords: functional magnetic resonance imaging, structural adaptive smoothing, structural adaptive segmentation, random field theory, multiscale testing

The purpose of the package **fmri** is the analysis of single subject functional magnetic resonance imaging (fMRI) data. It provides fMRI analysis from time series modeling by a linear model to signal detection and publication quality images. Specifically, it implements structural adaptive smoothing methods with signal detection for adaptive noise reduction which avoids blurring of activation areas.

We describe the complete pipeline for fMRI analysis using **fmri**, i.e., reading from various medical imaging formats, the linear modeling used to create the statistical parametric maps, signal detection and visualization of results. We review the rationale behind the structural adaptive smoothing algorithms and explain their usage from the package **fmri**. We demonstrate how such an analysis is performed using experimental data.

References

K. Tabelow, J. Polzehl, H.U. Voss, V. Spokoiny (2006). Analyzing fMRI experiments with structural adaptive smoothing procedures. *NeuroImage 33*, 55–62.

K. Tabelow, V. Piëch, J. Polzehl, H.U. Voss, High-resolution fMRI: Overcoming the signal-to-noise problem, J. Neurosci. Meth., 178 (2009) pp. 357–365.

J. Polzehl, H.U. Voss, K. Tabelow (2010). Structural adaptive segmentation for statistical parametric mapping. *NeuroImage 52*, 515–523.

K. Tabelow, J.D. Clayden, P. Lafaye de Micheaux, J. Polzehl, V.J. Schmid and B. Whitcher (2011). Image analysis and statistical inference in neuroimaging with R, *NeuroImage 55*, pp. 1686–1693.

K. Tabelow (2010). Statistical parametric maps for functional MRI experiments in R: The package fmri. Preprint no. 1562, WIAS, Berlin, http://www.wias-berlin.de/preprint/1562/wias_preprints_1562.pdf

neuRosim: an *R* package for simulation of fMRI magnitude data with realistic noise

Marijke Welvaert[1,⋆], **Yves Rosseel**[1]

1. Department of Data Analysis, Ghent University, Belgium
⋆Contact author: Marijke.Welvaert@UGent.be

Keywords: fMRI, simulation, physiological noise

Statistical analysis techniques for highly complex structured data such as fMRI data should be thoroughly validated. In this process, knowing the ground truth is essential. Unfortunately, establishing the ground truth of fMRI data is only possible with highly invasive procedures (i.e. intracranial EEG). Therefore, generating the data artificially is often the only viable solution. However, there is currently no consensus among researchers on how to simulate fMRI data. Research groups develop their own methods and use only in-house software routines. A general validition of these methods is lacking, probably due to the nonexistance of well-documented and freely available software.

In a response to this gap, **neuRosim** is developed to offer a software package for the simulation of fMRI data. The ultimate goal of the package is to create a general standardized platform that contains fMRI simulation methods that are validated. To this end, the package will contain several functions to simulate BOLD activation and fMRI noise.

In the current version, the functionalities of the package are separated in two layers. First, low-level functions are intended for advanced useRs who want in-depth control over their simulated data. With these functions it is possible to build fMRI data consisting of activation and noise while keeping full knowledge of the structure of the data and having the possibility to manipulate several parameters. Second, high-level functions are created to be used in more standard simulation studies. The power of these functions lies in the fact that they allow the useR to generate a full 4D fMRI dataset using only 3 command lines.

During the presentation, we will discuss the current features of **neuRosim** and our plans for coming updates. Using a variety of examples, we will demonstrate the useR-friendliness for first time useRs and how these examples can be extended to more advanced simulations. Finally, we will show the difference between the implemented simulation methods and discuss briefly how the simulated data can be used in other related *R* packages.

It's a Boy!
An Analysis of Tens of Millions of Birth Records Using R

Susan I. Ranney, Ph.D.[1*]

1. VP Product Development, Revolution Analytics, Inc.
*Contact author: sue@revolutionanalytics.com

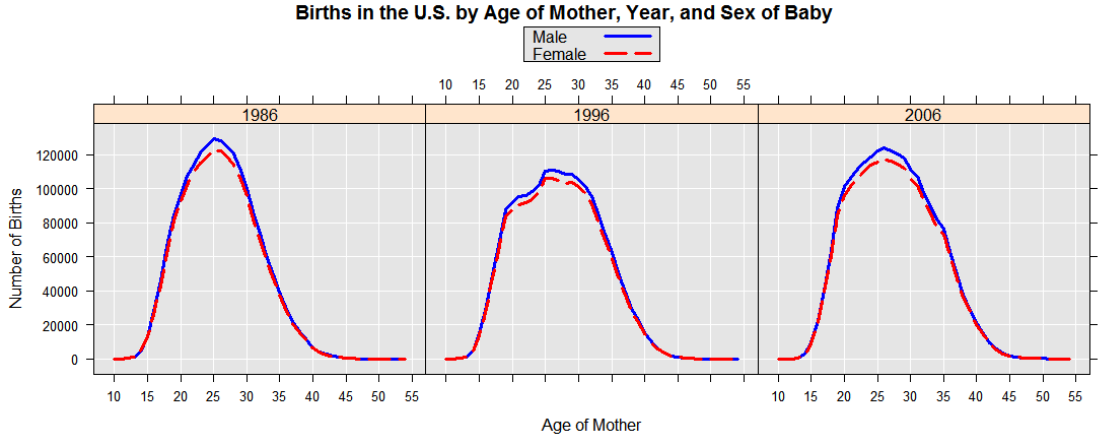
Keywords: Visualization, Data, Birth

The fact that more boys than girls are born each year is well established – across time and across cultures. But there are variations in the degree to which this is true. For example, there is evidence that the sex ratio at birth declines as the age of the mother increases, and babies of a higher birth order are more likely to be girls. Different sex ratios at birth are seen for different racial groups, and a downward trend in the sex ratio in the United States since the early 1970s has been observed. Although these effects are very small at the individual level, the impact can be large on the number of "excess" males born each year. To analyze the role of these factors in the sex ratio at birth, it is appropriate to use data on many individual births over multiple years.

Such data are in fact readily available. Public-use data sets containing information on all births in the United States are available on an annual basis from 1985 to 2008. But, as Joseph Adler points out in *R in a Nutshell*, "The natality files are gigantic; they're approximately 3.1 GM uncompressed. That's a little larger than *R* can easily process." An additional challenge to using these data files is that the format and contents of the data sets often change from year to year.

Using the **RevoScaleR** package, these hurdles are overcome and the power and flexibility of the *R* language can be applied to the analysis birth records. Relevant variables from each year are read from the fixed format data files into **RevoScaleR**'s .xdf file format using *R* functions. Variables are then recoded in *R* where necessary in order to create a set of variables common across years. The data are combined into a single, multi-year .xdf file containing tens of millions of birth records with information such as the sex of the baby, the birth order, the age and race of the mother, and the year of birth.

Detailed tabular data can be quickly extracted from the .xdf file and easily visualized using **lattice** graphics, as shown in the plot below. Trends in births, and more specifically the sex ratio at birth, are examined across time and demographic characteristics. Finally, logistic regressions are computed on the full .xdf file examining the conditioned effects of factors such age of mother, birth order, and race on the probability of having a boy baby.



References

Adler, Joseph (2010). *R in a Nutshell*.
CDC (1985-2008). Birth Data Files, http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm.
Matthews TJ, Hamilton BE (2005). Trend analysis of the sex ratio at birth in the United States, http://www.cdc.gov/nchs/data/nvsr/nvsr53/nvsr53_20.pdf.

Challenges of working with a large database of routinely collected health data: Combining *SQL* and *R*

Joanne Demmler[1], Caroline Brooks[1], Sarah Rodgers[1], Frank Dunstan[2], Ronan Lyons[1]

1. Swansea University, College of Medicine, Grove Building, Singleton Park, Swansea SA2 8PP
2. Cardiff University, Department of Primary Care & Public Health**,** Neuadd Meirionnydd, Heath Park, Cardiff CF14 4YS
*Contact author: j.demmler@swansea.ac.uk

Keywords: SAIL, anonymised data, RODBC, child health, data linkage

Introduction: Vast amounts of data are collected about patients and service users in the course of health and social care service delivery. Electronic data systems for patient records have the potential to revolutionise service delivery and research. But in order to achieve this, it is essential that the ability to link the data at the individual record level be retained whilst adhering to the principles of information governance. One such example is the Secure Anonymised Information Linkage (SAIL) databank, which contains health, social and education data for three million residents for a contiguous area in Wales, UK. There are currently 21 major datasets containing about 1.6 billion records, which can be linked anonymously at the individual level and household level.

Background: All work on SAIL data is executed through a secure remote desktop environment via a virtual private network (VPN), which has no internet connection and all output requires approval before it can be released for external viewing or publication. The processor speed equals 1 core of a Xeon X5550 @ 2.67 GHz processor, with an allocated memory of 2GB RAM per user.

Methods: We present here an example from the National Community Child Health Database, which contains height and weight measurements from school entry examinations for 849,238 children. Data are preselected using *SQL* and saved as a temporary table in SAIL to remove children with negative age at examination and to restrict examinations to the years 1990 to 2008. After removal of biologically unfeasible records with *SQL* 1,764,728 records for 594,720 children are imported into *R* using the `sqlfetch` command of the **RODBC** package. A simple algorithm is explored to remove remaining outliers in the data.

Results: Although *R* is very effective in some basic analysis and in the exploration of the data, it is not very efficient in dealing with such a vast dataset in certain situations. Memory limitations within the secure gateway platform mean that quite simple *R* scripts might run for a considerable time (days) and data might get lost in transfer operations (saving back to SAIL might fail depending on the table dimensions). At the present, this prevents the usage of more advanced statistical methods as well as modelling of the data.

Outlook: As a result of this analysis we have decided to migrate *R* onto a more powerful server. We are also investigating the possibility of multithreading *R* code, using the College of Medicine's BlueC replacement supercomputer (2 nodes, each with 30 Power7 cores @ 3.3GHz and 100GB of Ram).

Demographic: Classes and Methods for Data about Populations

John Bryant[1*]

1. Statistics New Zealand
*Contact author: john.bryant@stats.govt.nz

Keywords: S4 Classes, Demography, Official Statistics, Data Manipulation

Population data cross-classified by variables such as age, sex, and time is ubiquitous in official statistics and elsewhere. The presentation introduces **Demographic**, an *R* package under development that aims to ease some of the burden of manipulating this sort of data.

It is natural to represent cross-classified data as a multidimensional array, and to describe operations on cross-classified data as manipuations of arrays, such as splitting, collapsing, or expanding. Package **Demographic** facilitates this approach by providing an S4 class, `Demographic`, that builds on standard *R* arrays. Among other things, an object of class `Demographic`

- holds metadata such as the precise definitions of age intervals, or key words describing of the nature of each dimension (eg 'time', 'origin', or 'category'),
- has methods for most standard functions, many of which take advantage of the metadata by, for instance, checking that age intervals align correctly,
- encourages the use of names rather than numeric indices, in the interest of more transparent, less error-prone code,
- permits data.frame-like expressions such as `subset(x, age >= 65)`.

There are two main subclasses of `Demographic`: `Counts` and `Rates`. Objects from these two classes behave differently, reflecting differences in the way that analysts conventionally treat counts and rates data. For instance, when one `Counts` object is divided by another `Counts` object, a `Rates` object is produced, and if one of the `Counts` objects has a dimension that the other does not, the extra dimension is collapsed before the division is carried out.

The presentation will show how **Demographic** can be used for common tasks such as tidying messy data or doing simple projections. It will also provide an example of how a specialised package can take advanage of the general-purpose services provided by **Demographic**.

Correcting data violating linear restrictions using the deducorrect and editrules packages

Mark van der Loo[1,⋆], Edwin de Jonge[1] and Sander Scholtus[1]

1. Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands
⋆Contact author: m.vanderloo@cbs.nl

Keywords: Data editing, official statistics, linear restrictions, errror correction

editrules In many computational and statistical problems one needs to represent a set of m linear restrictions on a data record x of the form $Ax - b = 0$ or $Ax - b \geq 0$,where A is a matrix with real coefficients and b a vector in $\mathbb{R}^m$. Constructing and maintaining A manually is tedious and prone to errors. Moreover, in many cases the restrictions are stated verbosely, for example as "profit + cost must equal turnover". The **editrules** package can parse restrictions written in R language to matrix form. For example:

```
> editmatrix(c("x + 3*y == -z", "x>0"))
```

```
Edit matrix:
   x y z CONSTANT
e1 1 3 1        0
e2 1 0 0        0

Edit rules:
e1 : x + 3*y == -z
e2 : x > 0
```

The result is an `S3` object of class `editmatrix` which extends the standard `matrix` object. Here, the `editmatrix` function accepts linear restrictions in `character` or `data.frame` format. The latter offers the opportunity to name and comment the restrictions. The package also offers functionality to check data against the imposed restrictions and summarize errors in a useful way. In fact, the error checking functionality is independent of restrictions being of linear form, and can be used for any restriction including numerical and/or categorical data.

deducorrect Raw survey data is often plagued with errors which need to be solved before one can proceed with statistical analysis. The **deducorrect** package offers functionality to detect and correct typing errors (based on the Damerau-Levenshtein distance) and rounding errors in numerical data under linear restrictions. It also solves sign errors and value swaps, possibly masked by rounding errors. The methods used are (slight) generalizations of the methods described by Scholtus (2008) and Scholtus (2009). All datacorrection functions return corrected data where possible, a log of the applied corrections and the correction status. The package also offers functionality to determine if a matrix is totally unimodular, which is useful for solving errors in data involving balance accounts.

References

Scholtus, S. (2008). Algorithms for correcting obvious inconsistencies and rounding errors in business data. Technical Report 08015, Statistics Netherlands, Den Haag. *Accepted by J.Official Stat.*

Scholtus, S. (2009). Automatic correction of simple typing error in numerical data with balance edits. Technical Report 09046, Statistics Netherlands, Den Haag. *Accepted by J. Offical Stat.*

iWebPlots: Introducing a new *R* package for the creation of interactive web-based scatter plots

Eleni-Anthippi Chatzimichali[1,*] , Conrad Bessant[1]

1. Cranfield Health, Cranfield University, Bedfordshire, MK43 0AL. UK.
*Contact author: e.chatzimichali@cranfield.ac.uk

Keywords: Visualization, Image Maps, HTML, Web Graphics

Scatter plots constitute a most widely-used tool employed to investigate the correlations underlying a dataset. Nowadays, the demand for online data visualization in addition to the increasing need for availability of dynamic features and interactions, necessitate the construction of interactive web graphics.

The **iWebPlots** package simplifies the implementation of interactive web-based scatter plots, generated directly via the *R* statistical environment. The package's functions take as input a matrix of coordinates with associated metadata and generate a bitmap image with an HTML wrapper. Interactivity is implemented using the fundamental HTML image map technology, so no additional software such as applets or plug-ins is required. Developers can easily modify and expand the generated HTML pages, or incorporate them into web applications; thus, great extensibility is ensured.

Additional features include dynamic tooltips and text annotations as well as asynchronous alternation between two- and three-dimensional scatter plots. Furthermore, each plot can be interlinked with a fully interactive data table, which displays more information about the data in the plot.

The **iWebPlots** package has been used to develop the web front-end of a multivariate analysis pipeline, featuring techniques such as Principal Component Analysis (PCA) and k-means clustering, among many others. This work was carried out as part of the SYMBIOSIS-EU project, funded by European Commission Framework 7.

References

SYMBIOSIS-EU (2011), http://www.symbiosis-eu.net/

Rocessing: Interactive Visualizations in R

Ian Hansel[1]

1. Ernst & Young, Fraud Investigation and Dispute Services
*Contact author: ian.hansel@au.ey.com

Keywords: Visualization, Processing, Rjava, Interactive Graphs

Rocessing is a new package under development which combines *R* with *Processing* to produce highly interactive graphs and plots for Exploratory Data Analysis and presentation of data. *Processing* is an open source Java based language and environment which allows users to create images, animations and interact with data. There are two main objectives of this project:

1. Provide an in-depth visualization tool for analysing data in *R*.
2. Establish a platform that allows interactive applications written in *Processing* to be called from *R*.

Visualizing is often a key component of any analytical activity whether this is in the data cleansing stage, statistical modelling stage or the presentation of results. Being able to perform these tasks in an interactive manner can make these tasks simpler and quicker. In my presentation I will be going through a case study to show how **Rocessing** can be used at each step from data preperation and profiling right up to displaying the results of the analysis.

Rocessing utilises the **rJava** package to link *R* to *Processing*. The initial release scheduled for 20 April 2011 will provide interactive graphs such as scatterplots, histograms and mapping. In addition it will allow for user created *Processing* scripts to be called from *R*.

References

Fry, Ben (2008), *Visualizing Data*, O'Reilly Media, ISBN 0596514557.

Fry, Ben; Reas, Casey (2011). Processing, http://processing.org.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.Rproject.org .

Tufte, Edward R. (2001), *The Visual Display of Quantitative Information*, Graphics Press LLC.

Urbanek, Simon(2010). rJava: Low-level R to Java interface. R package version 0.8-6. http://CRAN.Rproject.org/package=rJava

Easy Interactive ggplots

Richard Cotton[1*]

1. Health and Safety Laboratory, Buxton, UK
*Contact author: richierocks@gmail.com

Keywords: ggplot2, gWidgets, plotting, data-viz, interactive

The **ggplot2** package allows the creation of high quality static graphics, but it does not currently support interaction. In general, making **ggplot2** graphics interactive is a difficult problem to solve, and is beyond the scope of this talk. Instead, a simpler problem is solved here: the ability to quickly and easily make small numbers of interactive features for specific usage, via control panels built using the **gWidgets*** packages. This talk demonstrates the creation of several of pieces of functionality that can be easily created and mixed together to provide a custom interactivity for your ggplots. Features demonstrated will include changing scale transformations, faceting, the title and axis labels, the dataset and geom layers.

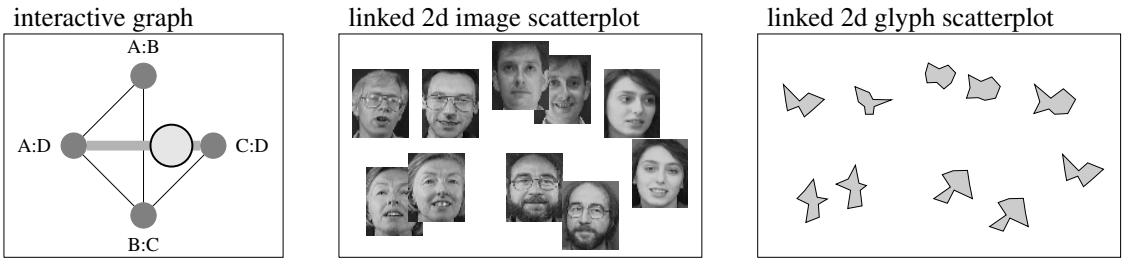
RnavGraph and the tk canvas widget

Adrian Waddell[1*], **Wayne Oldford**[1]

1. University of Waterloo
*Contact author: adrian@waddell.ch

Keywords: Visualization, Clustering, Multivariate statistics, Graphs, Images

We will demonstrate **RnavGraph**, an *R* package to explore and/or cluster interactively high dimensional data such as images, microarray or text data. The navigational infrastructure is provided by graphs (see Hurley and Oldford, 2011, Comp. Stat.). That is, the user moves on a graph a "You are here" circle, or "bullet", from one node to another along defined edges, causing some data visualization to be smoothly morphed from one plot (first node) into another plot (second node). Nodes of such graphs could for example represent 2d scatterplots and the edges rigid 3d rotations- or 4d transitions- from one scatterplot into another. We implemented our own 2d scatterplot display, called `tk2d` for this working example. `tk2d` can display either dots, images, star glyphs or text, and the data can be linked between displays.



The **RnavGraph** package heavily uses *tcl* and *tk* through the **tcltk** *R* package. Both the graph display and the `tk2d` display build upon the *tk* canvas widget. The `tk2d` display also makes use of the *C* API to achieve smooth morphing and image resizing results. Users can also easily extend the package and its predefined visualization instructions; this will be demonstrated during the talk.

For the second part of our presentation, we will present some of our insights gained while working with *tcl*, *tk* and *R*, especially with regard to the performance of the *tk* canvas widget. This will include some experimentation results, demonstration of some self-containing simple examples and some remarks about the *C* API to the *tk* canvas widget.

Using R for systems understanding – a dynamic approach

Thomas Petzoldt[1], Karline Soetaert[2]

1. Technische Universiotät Dresden, Institut für Hydrobiologie, 01062 Dresden
2. Centre for Estuarine and Marine Ecologoy (CEME), Netherlands Institute of Ecology (NIOO)
*Contact author: thomas.petzoldt@tu-dresden.de

Keywords: Simulation Models, Differential Equations, Stoichiometric Matrix, Ecology

The *R* system is not only a statistics and graphics system. It is a general-purpose high-level programming language that can be used for scientific computing in general. It is increasingly superseding conventional spreadsheet computing and became one of the standard environments for data analysis and modeling in ecology. An increasing collection of packages explicitly developed for dynamic modeling (**deSolve**, **simecol**, **FME, ReacTran,** cf. Soetaert et al 2010) and a growing number of textbooks teaching systems understanding by using R examples (Bolker, 2008; Soetaert and Herman, 2009; Stevens, 2009) are just an indicator for this trend.

The contribution will focus on practical experience with implementing and using dynamic models in R from an ecological modeler's perspective who works together with field and lab ecologists. Here, dynamic models play an essential role for improving qualitative and quantitative systems understanding. Aspects of two different case studies with increasing level of complexity are presented to demonstrate different application scenarios and modeling techniques:

- A model of semi-continuous laboratory cultures using package **deSolve** and the event mechanism,
- A water quality model for a polluted river using package **ReacTran** for transport and a compact representation using stoichiometric matrices (Peterson matrices) for matter turnover (cf. Reichert and Schuwirth, 2010).

The examples are organized in package **simecolModels**, a collection of simulation models that cover the range from teaching demos up to the ecosystem level. It is discussed how model implementations can be organized in a readable and computation-efficient way and how the model outcome can be visualized.

References

Bolker, B. (2008). *Ecological Models and Data in R.* Princeton University Press.

Reichert, P. and Schuwirth, N. (2010) A generic framework for deriving process stoichiometry in environmental models. *Environmental modelling and Software* 25, 1241-1251.

Soetaert, K. and Herman, P. M. J. (2009). *A Practical Guide to Ecological Modelling Using R as a Simulation Platform.* Springer.

Soetaert, K., Petzoldt, T. and Setzer, R. W. (2010) Solving differential equations in R. *The R Journal* 2/2, 5-15.

Stevens, M. H. H. (2009). *A Primer of Theoretical Population Ecology with R.* Springer.

Using multidimensional scaling with Duchon splines for reliable finite area smoothing

David L. Miller[1⋆], **Simon N. Wood**[1]

1. Mathematical Sciences, University of Bath, UK
⋆Contact author: dave@ninepointeightone.net

Keywords: spatial smoothing, generalised additive models, within-area distances, splines

Splines are often used to perform smoothing over geographical regions. However when boundary features intrude into the study region, splines may smooth across the features in a way that makes little sense and may be misleading. Other smoothers also suffer form this problem of inappropriately linking parts of the domain (for example either side of a river or peninsula). One view is that the problem is caused by the smoother using of an unsuitable metric to measure the distance between points. A simple substitution of within-region distances in place of Euclidean distances has been proposed previously to combat this problem, with some success, but at the price of some loss of clarity about what the resulting smoothing objective means, and quite high computational cost.

Our new approach takes within-area distances and uses them to project the location data into a new (often high-dimensional) space using multidimensional scaling. Conventional smoothing can then take place in this new space in which the euclidean distances between points now approximates the original within-region distances. We show that Duchon's 1977 generalisation of thin plate splines has particular advantages for smoothing in the MDS space.

As well as finite area smoothing, the new method provides a means for smoothing with respect to general distance measures, and examples of both spatial smoothing and more general distance-based smoothing are given.

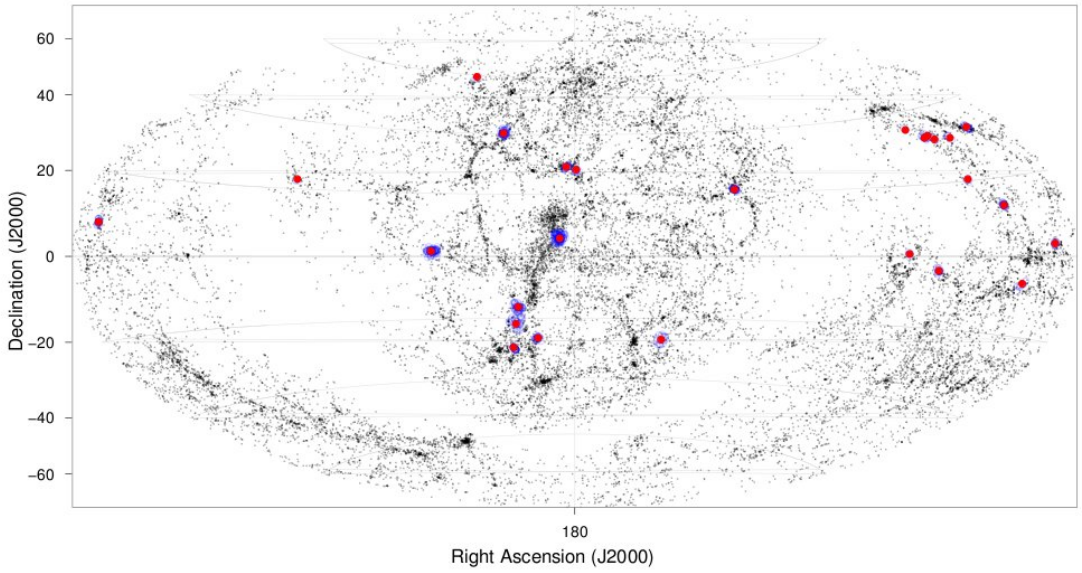
Studying galaxies in the nearby Universe, using R and ggplot2

Alastair Sanderson*

School of Physics & Astronomy, University of Birmingham, UK
*Contact author: ajrs@star.sr.bham.ac.uk (http://www.sr.bham.ac.uk/~ajrs)

Keywords: Astronomy, galaxies, multivariate data visualization, ggplot2

Galaxies are cosmic building blocks whose properties tell us about the origins and fate of the Universe. These objects are drawn together under the influence of gravity to form a complex pattern of filaments, and at the intersection of these filaments lie groups and clusters of galaxies– the largest gravitationally bound structures in the Universe. Here, large concentrations of dark matter tightly hold the fast-moving galaxies together, enabling close interactions and mergers to take place, which transform their properties. The location, luminosity and morphology of $\sim 10^5$ galaxies in the local Universe represents a 5 dimensional multivariate dataset which is ideally suited to exploration with *R*, and **ggplot2** in particular. I will show how *R* is being used to develop new insights into our understanding of galaxy groups and the process of cosmic feedback in the Universe, based on results from the CLoGS (Complete Local-Volume [galaxy] Groups Sample[1]) project, using the HyperLeda[2] galaxy database.



[1] http://www.sr.bham.ac.uk/~ejos/CLoGS.html
[2] http://leda.univ-lyon1.fr

Panel discussion: Challenges Bringing R into Commercial Environments

Louis Bajuk-Yorgan[1,*]

1. TIBCO Software Inc.
*Contact author: lbajuk@tibco.com

Keywords: Commercial, Applications, Panel

Champions of wider R usage within commercial environments often face multiple challenges, such as a lack of IT knowledge or acceptance, concerns about technical support, questions around validation or other regulatory compliance, difficulties in maintaining multiple versions of R, etc.

In this panel discussion, R champions and commercial vendors will share their views and experiences. Questions from the audience are highly encouraged.

NOTE:

I organized a panel discussion on this topic for useR 2010. The discussion was well-attended, with lots of questions from the audience, and several attendees approached me afterwards to suggest this be a regular feature of useR conferences. The members of the panel from last year are listed below. If my proposal is accepted, I would invite participation from Revolution and Mango again, and seek out new R champions to broaden the discussion.

Panel composition last year:

- Revolution Analytics (Norman Nie, CEO)
- TIBCO (Lou Bajuk-Yorgan, Sr. Director Product Management)
- Bret Musser, Director, Clinical Biostatistics at Merck
- Mango Solutions (Richard Pugh, CTO)
- Jim Porzak, Senior Director of Marketing Analytics at Ancestry.com
- Thomas G Filloon, Principal Statistician, Procter & Gamble

Suggested logistics:

- Each speaker sends one slide to moderator (TBD), moderator controls the slides
- Each speaker has 3 minutes to talk
- Rest of the time is devoted to Q&A and discussion

microbenchmark: A package to accurately benchmark *R* expressions

Olaf Mersmann[1*], **Sebastian Krey**[1]

1. TU Dortmund, Department of Statistics
*Contact author: olafm@datensplitter.net

Keywords: Benchmarking, Timing, High Performance Computing

We present the *R* package **microbenchmark**. It provides functions to accurately measure the execution time of *R* expressions. This enables the user to benchmark and compare different implementation strategies for performance critical functions, whose execution time may be small, but which will be executed many times in his program. In contrast to the often used `system.time` and `replicate` combination, our package offers the following advantages: firstly it attempts to use the most accurate method of measuring time made available by the underlying operating system, secondly it estimates to overhead of the timing routines and subtracts it from all measurements automatically and lastly it provides utility functions to quickly compare the timing results for different expressions.

In our presentation we will first describe the implementation of the timing routines for each platform (Windows, MacOS X, Linux and Solaris) and highlight some of the difficulties and limitations when measuring sub-millisecond execution times. Topics covered include strategies for dealing with CPU frequency scaling and multi-core system, why elapsed time is measured and not CPU time, why time is measured and not clock cycles as well as ways to estimate the granularity of the underlying timing routine. We conclude with a practical guideline for benchmark experiments on each of the operating systems mentioned.

After the viability of timing very short running *R* expressions has been established, some examples are provided to show why it is useful to study the performance characteristics of such expressions at all. First off, a baseline for the "speed" of the *R* interpreter is established by measuring the time it takes to execute the simplest possible function[1]. We consider this to be the equivalent of the cycle time of a microprocessor. Next we extend this idea to S3 and S4 methods to see how much overhead one incurs for the method dispatch in comparison to a simple function call. Afterwards we investigate the time it takes to generate a vector containing the numbers from 1 to n (for different n) using the different methods available in base *R*. Time permitting we conclude with a real world example where using high-level vectorized *R* functions actually hurt performance when compared to a naive implementation using low level operations.

[1] The simplest possible function being `function() NULL`.

Vector Image Processing

Paul Murrell[1*]

1. Department of Statistics, The University of Auckland
*Contact author: paul@stat.auckland.ac.nz

Keywords: image processing, vector graphics, importing graphics, SVG, javascript

This talk will describe a project to convert a static PDF map of a university campus into an interactive SVG map, using *R*.

The conversion involves three steps: importing the original map into *R*, using the **grImport** package; processing the map contents to identify important features; and using the **gridSVG** package to add interactivity to the map and export it in an SVG format.

Both the import and export steps depend on some recent improvements in the handling of lines and text in the **grImport** and **gridSVG** packages, plus the introduction of a new "path" primitive in the **grid** package. These new graphics features and enhancements will be described and demonstrated.

There will also be a discussion of the image processing step and the addition of interactivity to the map. The former will demonstrate the usefulness of treating an image as a source of data, which can be manipulated using the standard data processing tools in *R*, and the latter will demonstrate a way to add simple interactivity to the SVG output that is produced by the **gridSVG** package.

References

Murrell, P. *gridSVG: Export grid graphics as SVG*. R package version 0.6-0.

Murrell, P. (2009). Importing vector graphics: The grImport package for R. *Journal of Statistical Software 30*(4), 1–37.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Adding direct labels to plots

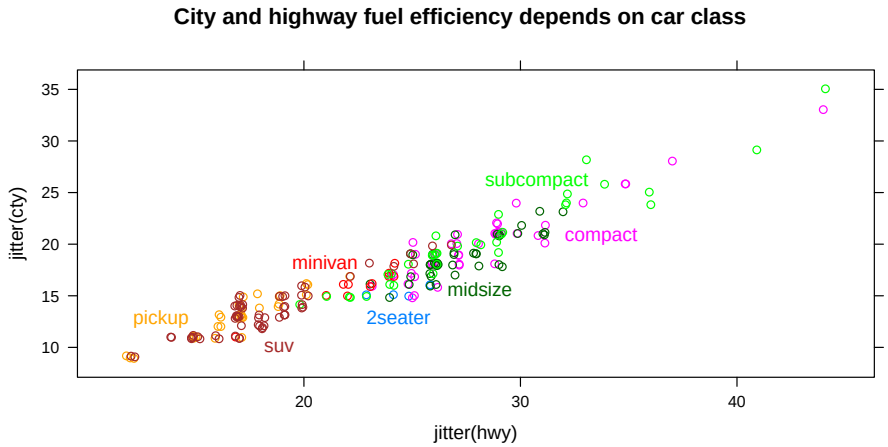
Toby Dylan Hocking[1,2,⋆]

1. INRIA Sierra team for machine learning research, 23 avenue d'Italie, Paris, France
2. Institut Curie/INSERM U900/Mines ParisTech cancer bioinformatics group, 26 rue d'Ulm, Paris, France
⋆Contact author: toby.hocking@inria.fr

Keywords: Visualization, graphics, legends, direct labels

High-level plotting systems such as **lattice** and **ggplot2** provide automatic legends for decoding color labels in *R* plots (Sarkar, 2008; Wickham, 2009). However, with many colors, legends become difficult to read, and direct labels are a more suitable decoding method:

```
> library(lattice)
> data(mpg, package = "ggplot2")
> p <- xyplot(jitter(cty) ~ jitter(hwy), mpg, groups = class,
+     main = "City and highway fuel efficiency depends on car class")
> library(directlabels)
> print(direct.label(p))
```



Direct labels are inherently more intuitive to decode than legends, since they are placed near the related data. However, direct labels are not widely used because they are often much more difficult to implement than legends, and their implementation varies between plotting systems.

The **directlabels** package solves these problems by providing a simple, unified interface for direct labeling in *R*. Given a **lattice** or **ggplot2** plot saved in the variable `p`, direct labels can be added by calling `direct.label(p,f)` where `f` is a Positioning Method that describes where labels should be placed as a function of the data. The power of this system lies in the fact that you can write your own Positioning Methods, and that any Positioning Method can be used with any plot. So once you have a library of Positioning Methods, direct labeling becomes trivial and so can more easily used as a visualization technique in everyday statistical practice.

References

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York: Springer. ISBN 978-0-387-75968-5.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

iRegression: a regression library to symbolic interval-valued variables

Lima Neto Eufrásio[1*], Souza Filho Cláudio[1], Anjos Ulisses[1]

1. Departamento de Estatística, Universidade Federal da Paraíba, Cidade Universitária s/n, João Pessoa, PB, Brazil
*Contact author: eufrasio@de.ufpb.br

Keywords: Regression, interval variable, symbolic data analysis.

Symbolic data analysis (SDA) has been introduced as a domain related to multivariate analysis, pattern recognition and artificial intelligence in order to introduce new methods and to extend classical data analysis techniques and statistical methods to symbolic data (Billard and Diday 2006). In SDA, a variable can assume as a value an interval from, a set of real numbers, a set of categories, an ordered list of categories or even a histogram. These new variables take into account the variability and/or uncertainty presented in the data. Interval variables have been studied in the area of SDA, where very often an object represents a group of individuals and the variables used to describe it need to assume a value which express the variability inherent to the description of a group. Moreover, interval-valued data arise in practical situations such as recording monthly interval temperatures in meteorological stations, daily interval stock prices, among others. Nowadays, different approaches have been introduced to analyze interval-valued variables. In the field of SDA, approaches to fit a regression model to interval-valued data have been discussed in the literature. However, the access to such methods still is restricted, being necessary to request to the authors. Billard and Diday (2000) were first to propose an approach to fit a linear regression model to symbolic interval-valued data sets. Lima Neto and De Carvalho (2008) improved the previous approach presenting a new method based on two linear regression models, the first regression model over the midpoints of the intervals and the second one over the ranges, which reconstruct the boundaries of the interval-values of the dependent variable in a more efficient way when compared with the Billard and Diday's method. The aim of this work is to development a R library, called iRegression, that includes some regression methods for interval-valued variables. This new library will be the first one developed to treat symbolic data in the regression context. Thus, some regression methods for symbolic interval-valued variables will be accessible to students, teachers and professionals.

Acknowledgments : The authors would like to thank CNPq (Brazilian Agency) for their financial support.

References

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley, New York.

Lima Neto, E.A. and De Carvalho, F.A.T.. (2008), Centre and range method to fitting a linear regression model on symbolic interval data. *Computational Statistics and Data Analysis 52*, pp. 1500-1515

Mixtures of Unimodal Distributions

Carlos E. Rodríguez[1,2,⋆] **, Stephen G. Walker**[1]

We present a new Bayesian mixture model. The main idea of our proposal is to change the components distribution of the mixture. Whereas the normal distribution is typically used as the kernel distribution, it does have some serious issues for the modeling of clusters, see for example Escobar & West (1995) or Richardson & Green (1997). If a cluster is skewed or heavy tailed, then the normal will be inefficient and many may be needed to model a single cluster. Our intention is to use as kernel a family of distribution functions for which the only constraint is that they are unimodal. Hence, we define a cluster as a set of data which can be adequately modeled via a unimodal distribution. To construct unimodal distributions we use a mixture of unform distributions with the Dirichlet Process, (Ferguson (1973) and Sethuraman (1994)), as the mixing distribution. To sample from the correct posterior distribution we use the ideas of Kalli, Griffin & Walker (2009), and Damien & Walker (2001), Tierney (1994), Green (1995) and Godsill (2001).

Keywords: Unimodal Distribution, Dirichlet Process, Markov chain Monte Carlo and Slice Sampler

1. International Society for Bayesian Analysis (ISBA)
2. Mexican Statistical Association (AME)
⋆Contact author: cerh2@kent.ac.uk

References

[1] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. Annals of Statistics, **1**, 209-230.

[2] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica, **4**, 639-650.

[3] Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Computational Graphical Statistics*, **10**, 230-248.

[4] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

[5] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society*, Series B **59**, 731-792.

[6] Escobar, M.D. and West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*. **90**, 577-588.

[7] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics* **22**, 1701-1762.

[8] Damien, P. and Walker, S.G. (2001). Sampling truncated normal, beta and gamma densities. *Journal of Computational and Graphical Statis- tics* **10**, 206-215.

[9] Kalli, M., Griffin, J.E. and Walker, S.G. (2009). Slice sampling mixture models. *Statistics and Computing*. **21**, 93-105.

Accelerating Simulations in R using Automatically Generated GPGPU-Code

Frank Kramer[1*], Andreas Leha[1], Tim Beissbarth[1]

1. Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany
*Contact author: frank.kramer@med.uni-goettingen.de

Keywords: GPGPU, CUDA, Simulation, simR2CUDA

Newly developed classifiers, estimators, and other algorithms are often tested in simulations to assess power, control of alpha-level, accuracy and related quality criteria.

Depending on the required number of simulation runs and the complexity of algorithms, especially the estimation of parameters and the drawing of random numbers under certain distributions, these simulations can run for several hours or even days. Such simulations are embarrassingly parallel and, therefore, benefit massively from parallelization. Not everyone has access to clusters or grids, though, but highly parallel graphics cards suitable for general purpose computing are installed in many computers. While there is a distinct number of maintained *R*-packages available (Eddelbuettel2011) that are able to interface with GPUs and allow the user to speed up computations, integrating these methods in simulation runs is often complicated and mechanisms are not easily understood.

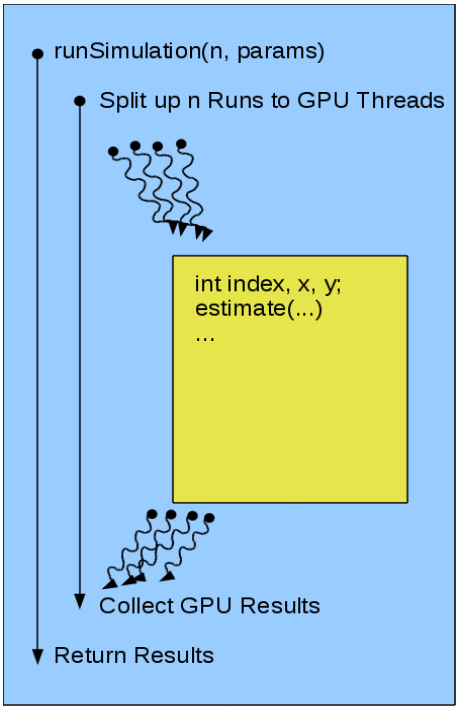


Figure 1: Illustration of generic process flow in generated code

We present a new package that acts as a wrapper for *CUDA*-implementations and offers a systematic way for a statistician to accelerate simulations written in *R*. After including the package **simR2CUDA** the user defines the algorithm for a single simulation run in an *R* function and passes the function, other required parameters and the number of simulation runs to a compiler function. The allowed syntax concerning flow-control for the function defining a single simulation run is restricted to comply with GPU and implementation limitations. Given that a *CUDA* environment is correctly installed and configured, a wrapper function, representing the whole simulation, and a *CUDA* function, representing an individual simulation step, are generated and integrated into *R* using `dyn.load`. Upon calling the generated code the simulation wrapper, a simple *C* function, splits up the independent simulation runs for parallel execution on GPU threads and collects and returns the results, after execution of all n runs on the GPU has finished (see Figure 1).

References

Eddelbuettel (2011). CRAN Task View: High-Performance and Parallel Computing with R, http://cran.r-project.org/web/views/HighPerformanceComputing.html.

The State of StatET

Stephan Wahlbrink[1], **Tobias Verbeke**[2,*]

1. WalWare / wahlbrink.eu
2. OpenAnalytics BVBA
*Contact author: tobias.verbeke@openanalytics.eu

Keywords: R, IDE, Eclipse, StatET, integrated development environment

The StatET plug-ins for Eclipse offer the professional R developer and statistician an integrated development environment since 2004. This presentation will look back on an interesting year of development and demonstrate some of the hidden and less hidden new functionality that was introduced between UseR!2010 and UseR!2011.

WalWare et al. (2004–2011). Eclipse Plug-In for R: StatET. http://www.walware.de/goto/statet.

The R Service Bus

Tobias Verbeke[1,⋆]

1. OpenAnalytics BVBA
⋆Contact author: tobias.verbeke@openanalytics.eu

Keywords: R, SOA, ESB, enterprise service bus, software integration

R is a unique computational and visualization engine that can be useful to power data-driven decision making in about any analytics problem. The complexity of many integration tasks, though, requires dedicated tools. The R Service Bus is a swiss army knife that allows you to plug R into your processes independently of the technology used by other software applications involved in the workflow. It has been in development for over two years and released under open source license to the R community in April 2011.

In this presentation we will provide an overview of the architecture of the R Service Bus and show how the architectural decisions guarantee reliability, flexibility, high availability and scalability of R-based analytics applications. Next we will demonstrate how applications can connect to RSB using RESTful and SOAP-based web services, e-mail protocols, messaging protocols (JMS, STOMP) etc. and which different payloads can be used to unleash to power of R.

OpenAnalytics (2010–2011). The R Service Bus. http://www.openanalytics.eu/r-service-bus.

Teaching Measurement Systems Analysis to Engineers Using *R*

Thomas Roth[1*]

The Department of Quality Science - Technical University of Berlin

*Contact author: Thomas Roth

Keywords: MSA, Capability, Gage R&R, ANOVA, DoE

Measurement Systems Analysis is an important aspect within the statistical education of engineers. While classical design of experiments (DoE) is dealt with in statistical courses for engineers, more specific designs and their analysis to identify the components of variation of measurement systems are rarely discussed. These Gage Repeatability & Reproducibility studies as well as procedures relating to terms such as Bias, Linearity and Gage Capability are subject of national and international standards and obligatory within but not restricted to automotive industry.

Methods utilizing the comprehensive **qualityTools** package, summaries and graphs as well as an example for teaching Measurement Systems Analysis to engineers as part of an obligatory statistics course for engineers are illustrated. The conceptual design of the methods and the relation to national and international standards are presented.

References

A.I.A.G. (2010). *Measurement systems analysis: Reference manual* (4 ed.). Detroit, Mich: DaimlerChrysler and Ford Motor and General Motors.

Burdick, R. K., C. M. Borror, and D. C. Montgomery (2005). *Design and analysis of gauge R&R studies: Making decisions with confidence intervals in random and mixed ANOVA models*. Philadelphia, Pa, Alexandria, Va: Society for Industrial Applied Mathematics and American Statistical Association.

Herrmann, J. and T. Roth (2010). Qualitätsmanagement als Pflichtfach für Bachelor an der TU-Berlin. In R. Schmitt (Ed.), *GQW 2010 , (Aachen, Germany), // Unternehmerisches Qualitätsmanagement*, pp. 205–217. Aachen: Apprimus-Verl.

ISO (2008). Quality management systems – Requirements (ISO 9001:2008).

ISO (2010). Statistical methods in process management – Capability and performance – part 7: Capability of measurement processes (ISO 22514-7).

Roth, T. (2010). qualityTools: Statistical Methods for Quality Science.

Verband der Automobilindustrie e.V. (2010). *Prüfprozesseignung: Eignung von Messsystemen, Mess- und Prüfprozessen, Erweiterte Messunsicherheit, Konformitätsbewertung*.

The *cards* Package

Jason Waddell[1,⋆]

1. OpenAnalytics BVBA
⋆Contact author: jason.waddell@openanalytics.eu

Keywords: R, cards, card games, teaching, probability

Many statistics courses use cards as examples for presenting probability theory. The **cards** package provides tools for plotting of cards and card games, along with adaptable interactive examples for card-based probability exercises and games.

In this presentation we will demonstrate functions for generating randomized decks and graphically displaying cards. In addition we will demonstrate how the package can be used to create fully interactive card games (such as casino-solitaire) that record and report play statistics.

Using R in Insurance: Examples from Lloyd's

Markus Gesmann[1,*], Viren Patel[1], Gao Yu[1]

1. Lloyd's, One Lime Street, London EC3M 7HA
*Contact author: markus.gesmann@lloyds.com

Keywords: Statistical Modelling, Data Management, Financial Analysis, Graphics, R packages

We propose to present about how the Analysis team of Lloyd's is using R in its operations. The poster focuses on 4 areas: statistical modelling, visualisation, usage of *R* and *LaTeX* to create management information, and collaboration via R packages.

Statistical modelling: Understanding the volatility of the risks insured at Lloyd's is crucial for pricing, business planning and capital setting. The Analysis team uses R to fit loss distribution to the historical data and Monte Carlo simulation to review the performance of syndicates and the market.

Visualisation: Data visualisation can help to provide early insight into favourable or unfavourable development of the business environment. The Analysis team uses **lattice** and **googleVis** to visualize the results of our work in a form that is easy to disseminate. As an example, we introduce the *Statistics Relating to Lloyd's* document, and how to use the **googleVis** package to display some of its information.

Using *R* and *LaTeX* to create management information: In a market with over 80 syndicates it is necessary to automatise the creation of performance reports for management information**.** Our team plays back Benchmarking reports which allows management and underwriters to view their own performance in the context of competitors. These reports are being produced using *R* and *LaTex*, among other software tools.

Creating internal R packages: For easier availability and dissemination teams in Lloyd's create packages of their R functions. This includes both packages tailored to a specific task and general packages to be used to comply with internal branding guidelines.

References

Lloyd's (2010). Statistics Relating to Lloyd's, http://www.lloyds.com/stats/.

Solving Norm Constrained Portfolio Optimizations via Coordinate-Wise Descent Algorithms

Yu-Min Yen[1] **and Tso-Jung Yen**[1]

1. Department of Finance, LSE and Institute of Statistical Science, Academia Sinica
*Contact author: Y.YEN@lse.ac.uk

Keywords: Portfolio optimization, norm constraint, coordinate-wise descent algorithm

In this paper we demonstrate that coordinate-wise descent algorithms can be used to solve portfolio selection problems in which asset weights are constrained by l_q norms for $1 \leq q \leq 2$. A special case of the such problems is when $q = 1$. The l_1 norm constraint promotes zero values for the weight vector, leading to an automatic asset selection for the portfolio. We first consider the case of minimum (global) variance portfolio (mvp) in which the asset weights are constrained by weighted l_1 and squared l_2 norms. We use two benchmark data sets to examine performances of the norm constrained portfolio. When the sample size is not large in comparison with the number of assets, the norm constrained portfolio tends to have a lower out-of-sample portfolio variance, lower turnover rate, less numbers of active assets and short-sale positions, but higher Sharpe ratio than the one without such norm constraints. We then show some extensions; particularly we derive an efficient algorithm for solving an mvp problem in which assets are allowed to be chosen grouply. All of the program codes for the algorithms are written by *R*.

Using *R* for air quality data analysis: A tool for designing improved large-scale air pollution prevention programs

A. Alija[1⋆], M. N. Gonzalez[1], A. Junquera-Perez[1], A. Ayesta[1], E. Setien[1], L. Garcia[1] and J. Blanco[1]

1. Ingenieros Asesores S.A. Parque Tecnologico de Asturias. Parcela 39, 33428. Principado de Asturias

⋆Contact author: aab@ingenierosasesores-sa.es

Keywords: Air quality, Air pollution, **openair**, data mining.

Emissions of air pollutants derive from almost all economic and societal activities. The majority of greenhouse gas emissions (GHG), acidifying substances, tropospheric ozone precursor emissions and material input caused by the life-cycles of activities related to consumption can be allocated to the main consumption areas of eating and drinking, housing and infrastructures, and mobility [Age11].

In this work, we present an intensive use of *R* and more concretely package **openair** [CR11]. Package **openair** lead out to analyse long historical time series of air pollution data coming from a large urban region somewhere [pri]. Long historical series of data have the inconvenience of they are very difficult to analyze for several reasons. Perhaps, some of these reasons are: first, air pollution data is a data extremely correlated in time and second, air pollution data is a complex data which require a very optimized visualization tools to carry out even the simplest analysis. Therefore, much of the data available is only briefly analysed; perhaps with the aim of comparing pollutant concentrations with national and international air quality limits. Package **openair**, thanks to the specific functions, helps us to overcome some of these barriers.

We would like to thanks PhD. Lucas Fernandez Seivane for our helpfull discussions.

References

[Age11] EEA (European Environment Agency). Soer2010. http://www.eea.europa.eu/soer, 2011.

[CR11] David Carslaw and Karl Ropkins. *openair: Open-source tools for the analysis of air pollution data*, 2011. R package version 0.4-0.

[pri] private. To preserve the confindenciality of the data source we will keep the location anonymous.

Invasions by polymorphic species

Elizabeth Elliott[1,*] and Stephen Cornell[1]

1. University of Leeds
*Contact author: jhs5ece@leeds.ac.uk

Keywords: Range-expansion; Evolution; Reaction-diffusion model; Numerical simulations

Two issues of current importance in ecology are the range expansion of species and invasion of exotic organisms. Climate change is affecting the environments that species are currently adapted to, and to survive species will either need to adapt to the new conditions or shift with their current environment. Empirical observations and simulation models have shown that higher dispersal ability may lead to increased rates of spread but little is known about the effect of having a community of dispersal phenotypes on the rate of range expansion. We use a spatially explicit analytical model based on partial-differential equations to investigate the invasion of a species with two dispersal phenotypes into a previously unoccupied landscape. These phenotypes differ in both their dispersal and population growth rate. Using analytical techniques and carrying out numerical simulations in *R* (R Development Core Team, 2010) using the **deSolve** package (Soetaert *et al.*, 2010), we find that the presence of both phenotypes can result in faster range expansions than if only a single phenotype were present in the landscape, and that typically the invasion can occur up to twice as fast as a result of this polymorphism. This has implications for predicting species invasion speeds, suggesting that speeds cannot just be predicted from looking at a single phenotype and that the presence of a community of phenotypes needs to be taken into consideration.

References

R Development Core Team (2010). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Soetaert, K., Petzdolt, T. and Setzer, R. W (2010). Solving differential equations in R: Package deSolve. *Journal of Statistical Software 33*, 1-25.

Using *R* to Empower a New Plant Biology

Naim Matasci[1,*], Matthew Vaughn[2], Nirav Merchant[1,3]

1. The iPlant Collaborative, BIO5 Institute, University of Arizona
2. The iPlant Collaborative, Texas Advanced Computing Center, University of Texas at Austin.
3. Biotechnology Computing Facility, BIO5 Institute, University of Arizona
*Contact author: nmatasci@iplantcollaborative.org

Keywords: cyberinfrastructure, plant sciences, bioinformatics, high performance computing

The iPlant Collaborative is a U.S. National Science Foundation program to build a cyberinfrastructure for the plant sciences that will enable new conceptual advances through integrative, computational thinking. Teams composed of plant biologists and computer scientists work on grand challenges that address fundamental questions in plant biology, in particular focusing on elucidating the link between genotype and phenotype and on building a tree of life to represent the evolutionary history of all green plants.
Countless computer programs address individual aspects of these grand challenges but the complexity and sheer magnitude of the problem makes it almost impossible for individual researchers or research groups to tackle such problems alone.
iPlant provides a cyberinfrastructure that allows researchers in the plant science community to collaborate and share and integrate data and algorithms. Furthermore, by providing access to High Performance Computing resources it permits the generation and analysis of very large datasets as well as the usage of computationally intensive algorithms.
Given the growing importance of *R* in virtually every field of biology and bioinformatics, iPlant's cyberinfrastructure has been enriched by an interface to facilitate the development, execution and distribution of *R* scripts.
Developers can work on constructing new tools in Atmosphere, iPlant's virtual environment, through an IDE and access data and workspaces located in their iPlant home directory on in shared locations. All I/O operations are handled through iRODS, the Integrated Rule-Oriented Data System which permits the sharing of data and provides direct read access to plant genomes and other large files. The use of iRODS also ensures consistency of data across the different access point to iPlant's cyberinfrastructure (Atmosphere, API, Discovery Environment). Through iPlant's API users and developers can execute computationally intensive scripts on Texas Advanced Computing Center's HPC resources. Finally, by providing a simple metadata description, developers can create GUIs for their applications and make them available to colleagues and collaborators only or to the community at large through iPlant's Discovery Environment.

References

Atmosphere: http://atmo.iplantcollaborative.org/
iRODS: https://www.irods.org/
iPlant's Discovery Environment: http://preview.iplantcollaborative.org/de/

Rknots, an *R* package for the topological analysis of knotted biological polymers

Federico Comoglio[1⋆], **Maurizio Rinaldi**[2]

1. Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology (ETH) Zurich, Basel, Switzerland
2. Department of Chemical, Food, Pharmaceutical and Pharmacological Sciences (DiSCAFF), University of Piemonte Orientale "A. Avogadro", Novara, Italy
⋆Contact author: federico.comoglio@bsse.ethz.ch

Keywords: **Rknots**, knotted polymers, computational knot theory, topology, proteins

The topological study of biological polymers has led to important insights into their structural properties and evolution Virnau et al. (2006). From a topological point of view polymers can be naturally modeled as sequences of 3D points, i.e. open polygonal paths. Their closure generates classical objects in topology called knots. Boosted by Taylor's work Taylor (2000), the characterization of knotted proteins is a subject of increasing interest in both experimental and computational biology due to the close structure-function relationship and reproducible entangled folding. The exponential growth of the total number of structures deposited into the Protein Data Bank (PDB) requires dedicated computational high-throughput methods able to deal with a large amount of data. These methods combine a structure reduction scheme of a protein backbone model with the computation of a knot invariant. Being easy to compute, the Alexander polynomial represents the current default choice Kolesov et al. (2007). Although this choice is primarily supported by the fact that the knots so far identified in proteins are of the simplest types, the Alexander polynomial is not able to discern knots chirality and its limited power is not optimal to develop a generalized framework aiming to describe topological properties of 3D structures.
Recently, we developed a topological framework for the computation of the HOMFLY polynomial Freyd et al. (1985), a more powerful invariant able to define knots chirality. By screening the entire PDB, we obtained an up-to-date table of knotted proteins that also includes two newly detected right-handed trefoil knots in recently deposited protein structures Comoglio and Rinaldi (2011).
However, the application of our framework is not limited to proteins and in order to provide an open-source package to the scientific community working in the field, we developed and here we present **Rknots**, an *R* package that includes functions and utilities to process and topologically explore polymers along with dedicated utilities for working with protein structures. Among the implemented functions, this first release includes PDB entry import or fetching (exploiting the flexibility of the *R* package **bio3d**), structure reduction via the Alexander-Briggs algorithm Alexander and Briggs (1926) and the MSR (Minimal Structure Reduction algorithm) Comoglio and Rinaldi (2011), topological invariants and linking number computation, 3D and knot diagram graphics function along with a knots and links dataset repository and utilities.
We welcome external contributions to **Rknots** in order to further extend the package functionalities and provide a general purpose tool to identify knots in three-dimensional structures.

References

Alexander, J.-W. and G.-B. Briggs (1926). On types of knotted curves. *Ann of Math 28*, 562–586.

Comoglio, F. and M. Rinaldi (2011). A topological framework for the computation of the homfly polynomial and its application to proteins. *PLoS ONE in the press*.

Freyd, P., D. Yetter, J. Hoste, W.-B.-R. Lickorish, K. Millett, and A. Ocneanu (1985). A new polynomial invariant of knots and links. *Bull Amer Math Soc (NS) 12*, 239–246.

Kolesov, G., P. Virnau, M. Kardar, and L.-A. Mirny (2007). Protein knot server: detection of knots in protein structures. *Nucleic Acid Res 35*, W425–W428.

A specialised software for statistical applications in macromolecular crystallography

James Foadi[1,*], **Gwyndaf Evans**[2], **David Waterman**[3]

1. MPL, Imperial College, London SW7 2AZ
2. Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0DE
3. CCP4 - Research Complex at Harwell (RCaH) - Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0FA

*Contact author: j.foadi@imperial.ac.uk

Keywords: Macromolecular crystallography

A new package for applications in macromolecular structural crystallography, **cRy**, has been developed in the renowned *R* statistical platform. This is the first software of its kind and it is supposed to provide a bridge between the large communities of crystallographers and professional statisticians. At present macromolecular crystallographers make heavy use of large systems of programs, like CCP4 [1] or PHENIX [2], mainly written in Fortran, C/C++ or Python. These programs handle the several statistical operations, normally carried out in crystallography, with *ad hoc* routines, usually developed by different authors, often not sharing a common statistical platform. Data and results are exchanged through files with well-defined formats. The **cRy** package reads and writes files in the most commonly used crystallographic formats, carries out all major crystallographic calculations and provides an interface between the crystallographic data structure and the statistical objects and tools offered by *R*. **cRy** provides, thus, that common statistical platform that, at present, is still lacking in structural crystallography. The code has been developed using S4 classes.

References

[1] Collaborative Computational Project 4 (1994). The CCP4 Suite of Programs for Protein Crystallography. *Acta Cryst. D*, **50**, 760–763.

[2] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echoo Is, J. J. Headd, L. -W. Hung, G. J. Kapral, R. W. Grosse-Kuntsleve, A. J. Terwilliger and P. H. Zwart (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst. D*, **66**, 213–221.

EMA - A R package for Easy Microarray data Analysis

Nicolas Servant[1,2,3,⋆], Eleonore Gravier[1,2,3,4,], Pierre Gestraud[1,2,3], Cecile Laurent[1,2,3,6,7,8], Caroline Paccard[1,2,3], Anne Biton[1,2,3,5], Isabel Brito[1,2,3], Jonas Mandel[1,2,3], Bernard Asselain[1,2,3], Emmanuel Barillot[1,2,3], Philippe Hupé[1,2,3,5]

1. Institut Curie, Paris F-75248, France
2. INSERM, U900, Paris F-75248, France
3. Ecole des Mines ParisTech, Fontainebleau, F-77300 France
4. Institut Curie, Departement de Transfert, Paris F-75248, France
5. CNRS, UMR144, Paris F-75248, France
6. CNRS, UMR3347, Orsay F-91405, France
7. INSERM, U1021, Orsay F-91405, France
8. Université Paris-Sud 11, Orsay F-91405, France

⋆Contact author: ema-support@curie.fr

Keywords: Microarray analysis

The increasing number of methodologies and tools currently available to analyse gene expression microarray data can be confusing for non specialist users. Based on the experience of biostatisticians of Institut Curie, we propose both a clear analysis strategy and a selection of tools to investigate microarray gene expression data. The most usual and relevant existing *R* functions were discussed, validated and gathered in an easy-to-use *R* package (**EMA**) devoted to gene expression microarray analysis.
Removing noise and systematic biases is performed using the most famous techniques for Affymetrix GeneChip normalisation. The data are then filtered to both reduce the noise and increase the statistical power of the subsequent analysis. Exploratory approaches based on R packages such as **FactoMineR**, or **mostclust** and classically used to find clusters of genes (or samples) with similar profiles are also offered. Supervised approaches, as Significance Analysis of Microarrays (**siggenes** package) approach or ANOVA functions, are proposed to identify differentially expressed genes (DEG) and functional enrichment of the DEG list is assessed based on **GOstat** package.
The package includes a vignette which describes the detailed biological/clinical analysis strategy used at Institut Curie. Most of the functions were improved for ease of use (fewer command lines, default parameters tested and chosen to be optimal). Relevant, enhanced and easy-to-interpret text and graphic outputs are offered. The package is available on The Comprehensive R Archive Network repository.

References

Bertoni, A. and G. Valentini (2007). Model order selection for bio-molecular data clustering. *BMC Bioinformatics 8 Suppl 2*, S7.

Falcon, S. and R. Gentleman (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics 23*, 257–258.

Le, S., J. Josse, and F. Husson (2008). Factominer: an R package for multivariate analysis. *Journal of statistical software 25*, 1–18.

Servant, N., G. Eleonore, P. Gestraud, C. Laurent, C. Paccard, A. Biton, I. Brito, J. Mandel, B. Asselain, E. Barillot, and P. Hupé (2010). Ema - a R package for easy microarray data analysis. *BMC Research Notes 3*, 277.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A 98*, 5116–21.

Classification of Enzymes via Machine Learning Approaches

Neetika Nath*, John B. O. Mitchell

Biomedical Sciences Research Complex and School of Chemistry, University of St Andrews, KY16 9ST
*nn223@st-andrews.ac.uk

Keywords: Enzyme Classification, Machine Learning, R, Protein Function Prediction.

Abstract: We compare enzyme mechanistic descriptors derived from the MACiE (Mechanism, Annotation and Classification in Enzymes) database [Holliday et al., 2006] and use multivariate statistical analysis for assessment of enzyme classification. Each enzyme has an Enzyme Commission (EC) number, a numerical code designed to classify enzymes by describing the overall chemistry of the enzymatic reaction. The EC number system was devised five decades ago, in a pre-bioinformatics age. As the volume of available information is increasing, a large number of informatics groups have tried to use protein sequence and structural information to understand and reproduce the classification, some of which have been successful [Cai et al., 2004]. Other groups have effected automatic EC classification using chemoinformatics descriptions of the underlying reactions. Our objective is to develop a computational protocol using the *R* package **CARET** [Kuhun *et al.*, 2007] to predict EC number from MACiE-derived descriptors. We evaluate 260 well annotated chemical reaction mechanisms of enzymes using machine learning methods, placing them into the six top level EC classes. Moreover, we compare the classification performances of three supervised learning techniques, Support Vector Machine (SVM) [Vapnik, 1998], Random Forest (RF) [Breiman, 2001] and K Nearest Neighbour (kNN), for the reaction mechanism classification task using five different descriptor sets from MACiE data. **Results:** We found that all classifiers performed similarly in terms of overall accuracy with the exception of K Nearest Neighbour analysis, which has the lowest performance. The best performance was achieved by the Random Forest classifier.

References

Holliday, G.L., et al.. (2006) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms http://www.ebi.ac.uk/thornton-srv/databases/MACiE/

Cai, C., L. Han, et al. (2004). "Enzyme family classification by support vector machines." *Proteins* 55: 66 - 76.

Kuhn, M., Wing, J., Weston, S, Williams A., Keefer, C. & Engelhardt, A. (2007) caret: Classification and Regression Training, http://cran.r-project.org/web/packages/caret/

Vapnik, V. N. (1998) Statistical Learning Theory. New York: John Wiley and Sons.

Breiman, L. (2001) “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5– 32.

BeadDataPackR: A Compression Tool For Raw Illumina Microarray Data

Mike L. Smith[1,⋆], **Andy G. Lynch**[1]

1. Department of Oncology, University of Cambridge, Cancer Research UK Cambridge Research Institute, Robinson Way, Cambridge CB2 0RE, UK
⋆Contact author: mike.l.smith@cancer.org.uk

Keywords: Bioinformatics, Microarray, Illumina, Compression

Microarray technologies have been an increasingly important tool in biological research in the last decade, and a number of initiatives have sought to stress the importance of the provision and sharing of raw microarray data. Illumina BeadArrays provide a particular problem in this regard, as their random construction and high number of replicate probes simultaneously adds value to analysis of the raw data and obstructs the sharing of those data, with many of the standard repositories having no facilities for storing the raw data. This leaves the burden of distributing such data with the individual researcher. The **BeadDataPackR** package provides an interface to a novel compression scheme for raw Illumina BeadArray data, designed to ease the storage and bandwith concerns that providing access to such data brings. It offers two key advantages over off-the-peg compression tools. First it uses knowledge of the data formats to achieve greater compression than other approaches, and second it does not need to be decompressed for analysis, but rather the values held within can be directly accessed by existing tools, such as the popular **beadarray** package.

References

M.L.Smith and A.G.Lynch (2010). BeadDataPackR: A Tool to Facilitate the Sharing of Raw Data from Illumina BeadArray Studies. *Cancer Informatics 9*, 217–227.

GWAtoolbox: An R Package for Time Efficient Quality Control of Multiple GWAS Data Files

Daniel Taliun[1,2,⋆], Christian Fuchsberger[3], Peter P. Pramstaller[1], Cristian Pattaro[1] on behalf of the CKDGen consortium

1. Institute of Genetic Medicine, European Academy Bozen-Bolzano (EURAC), Bolzano, Italy
2. Free University of Bozen-Bolzano, Bolzano, Italy
3. Department of Biostatistics, University of Michigan, Ann Arbor, MI
⋆Contact author: daniel.taliun@eurac.edu

Keywords: genome-wide association study, quality control, visualization

In the recent years, Genome-Wide Association Studies (GWASs) have been proven to be a very powerful approach to uncover common genetic variants affecting human disease risk and quantitative outcome levels. To date, 1212 genetic loci were reported to be significantly associated with at least one of 210 traits (Hindorff et al. (2011)). To allow sufficient power to identify variants with small effects, GWAS sample size has been augmented by pooling results from dozens of individual GWASs into large meta-analyses efforts. However, combining results from a large number of GWASs, which differ in terms of study design, population structure, data management, and statistical analysis, poses several challenges regarding the consistency and quality of data which are usually difficult to be addressed systematically. This is mainly due to the GWAS file size, which typically includes 2.5 to 7 million rows (corresponding to genetic variants) and > 9 columns (attributes). Consequently, the data harmonization across studies usually takes several weeks or months.

While working in the CKDGen Consortium, aim to detect renal function genes (Köttgen et al. (2010)), we have been performing meta-analyses of several GWASs. To standardize and speed up the quality control (QC) process we developed the **GWAtoolbox**, an *R* package which lightens and accelerates the handling of huge amounts of data from GWASs. **GWAtoolbox** provides time efficient QC of data stored in dozens of files. Based on a simple configuration script, **GWAtoolbox** can process any number of files and produce QC reports in a matter of minutes. QC reports consist of an extensive list of quality statistics and graphical output presented using DHTML, which allow fast and easy inspection of individual data files. Additional statistics and graphs allow quick identification of studies that are systematically different from the other (outliers). The high time efficiency was achieved through the data reduction technique integrated into the visualization pipeline. In particular, instead of passing all the million data points to the *R* plotting routines, only a small part of data points is selected in a such way that preserves the quality of the final graphical output. Additionally, the implementation of all computationally intensive steps was transferred to *C++*.

Through its extensive use in several current GWAS meta-analyses, **GWAtoolbox** has been proven to significantly speed up the data management and to improve the overall meta-analysis quality. **GWAtoolbox** is open source and available for MacOS, Linux, and Windows OS, at http://www.eurac.edu/gwatoolbox.

References

Hindorff, L. et al. (2011). A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed March 23, 2011.

Köttgen, A. et al. (2010). New loci associated with kidney function and chronic kidney disease. *Nat Genet 42*(5), 376–84.

Is the seasonal variation in hospitalisations rates of atrial fibrillation induced strokes in Denmark and New Zealand dynamic?

Anette Luther Christensen[1,*], Simon Hales[2], Sren Lundbye-Christensen[1], Lars HVilsted Rasmussen[1], Kim Overvad[1], Claus Dethlefsen[1]

1. Department of Cardiology, Center of Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital, Aalborg, Denmark
2. Department of Public Health, University of Otago, Wellington, New Zealand
*Contact author: anluc@rn.dk

Keywords: State space model, Model selection, Seasonal variation, Atrial Fibrillation, Stroke

Atrial Fibrilation (AF) is the most common cardiac arrythmia Friberg et al. (2003) and the hospitalisations rates of AF have increased during the last two decades; in fact AF is considered an epidemic. The frequency of AF increases with age and considering the population demographic it is expected that the number of people with AF will be increasing remarkably. Having AF often leads to palpitations, respiratory distress, and fatigue. Furthermore, AF is considered an independent risk factor for stroke Wolf et al. (1991). The consequences of stroke on patients are crucial and may be a considerable burden on society regarding rehabilitation of stroke patients. It has been reported that hospitalisations with stroke exhibit seasonal variation during the calendar year, however seasonal variations in AF induced strokes have not been investigated. Furthermore, it has not been investigated whether there has been changes in the seasonal variation of stroke hospitalisations. Knowledge of the seasonal variation of AF induced stroke, and possible changes over time, may contribute to knowledge of the etiology of AF and may improve prophylaxis treatment in AF patients, which may lead to improved prognoses for AF patients.

Using a state space model to fit daily incidence rates of AF induced strokes, accounting for a secular trend and modelling the seasonal variation as a sum of sinusoids with different frequencies, it is possible to investigate whether hospitalisations of AF induced strokes exhibit seasonal variation. Furthermore, performing model selection we may be able to investigate the dynamic nature of the seasonal variation over time. However, modelling the daily incidence rates of AF induced strokes as being Poisson distributed using a state space model, as proposed by Lundbye-Christensen et al. Lundbye-Christensen et al. (2009), is not trivial in regards to estimation algorithms and especially performing model selection.

We identified daily hospitalisations in Denmark using the Danish National Patients Registry, and in New Zealand using the National Minimum Data Set. Both registries hold records of all hospitalisations, each record includes information on date of hospitalisation and primary diagnosis as well as secondary diagnoses. Analyses will be performed on data sets from each country separately, to make comparisons of seasonal variation in incidence rates of AF induced strokes between the two countries. All analyses will be performed in *R* using the package **sspir**.

References

Friberg, J., P. Buch, H. Scharling, N. Gadsbll, and G. B. Jensen (2003). Rates of Hospital Admissions for Atrial Fibrillation. *Epidemiology 13*(6), 666–672.

Lundbye-Christensen, S., C. Dethlefsen, A. Gorst-Rasmussen, T. Fischer, H. C. Schønheyder, K. J. Rothman, and H. T. Sørensen (2009). Examing Secular Trends and Seasonality in Count Data Using Dynamic Generalized Linear Models: a New Methodological Approach Illustrated with Hospital Discharge Data on Myocardial Infarction. *European Journal of Epidemiology 24*(5), 225–230.

Wolf, P. A., R. D. Abbott, and W. B. Kannel (1991). Atrial Fibrillation as an Independent Risk Factor for Stroke: the Framingham Study. *Stroke 22*(8), 983–988.

Predictive Tool Development
A Time-to-event study of Diabetes Complications in Ontario

Kelvin Lam[1*], Douglas Manuel[1,2]

1. Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada
2. The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
* Contact author : kelvin.lam@ices.on.ca

Keywords: Diabetes Complications, Competing Risk, Data Visualization, Administrative Database, Diagnostic Tools

For physicians and people with diabetes, a simple and accurate predictive tool for complications such as acute myocardial infarction (AMI) and cardiovascular disease (CVD) will be invaluable for clinical decision making. Our study looked at how age and previous medical history affects duration to these complications as well as death using the competing risk methodology, where the latter served as the main competing event. *R* package **cmprsk** was used to derive model parameters and baseline cumulative hazards. To determine the model validity, diagnostics including c-index, deciles plots and quantile risk ratio plots will be presented using **ggplot2**.

References

[1] Fine JP and Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. JASA 94:496-509.

[2] Putter H, Fiocco M, Geskus RB (2007). Tutorial in biostatistics: Competing risks and multi-state models. Statistics in Medicine 26, 2389–2430.

[3] Bob Gray <gray@jimmy.harvard.edu> (2010). **cmprsk**: Subdistribution Analysis of Competing Risks. R package version 2.2-1. http://CRAN.R-project.org/package=cmprsk

Analyzing the American Community Survey with *R*

Anthony Damico[1,*] , Rachel Licata[1]

1. Kaiser Family Foundation
*Contact author: adamico@kff.org

Topic Area: Statistics in the Social and Political Sciences

Keywords: American Community Survey, Health Policy, Big Data, SQL, County-Level Data

Research Objective: Designed to replace the information gathered by the decennial Census, the American Community Survey (ACS) is likely to play an increasingly important role in the future of health policy research, especially for regional and state-level studies. Despite powerful statistical analysis, data manipulation, and presentation capabilities, the open-source R Statistical Software Package has not become widely adopted in the fields of health economics, health policy, or healthcare management and evaluation. User guidelines and other technical documents for government-funded and publicly-available data sets rarely provide appropriate syntax examples to R users. The objective of this presentation will be to describe the steps required to import ACS data into R, to prepare that data for analysis using the replicate weight variance and generalized variance formula calculation techniques, and to produce the principal set of statistical estimates sought by health policy researchers.

Study Design: This presentation reviews the step-by-step method explanations and syntax needed to analyze the ACS with the R Statistical Software package. This includes importation instructions, estimate calculations, variance and error term calculations, as well as linkages to other data sets. In order to equip healthcare researchers with the tools needed to analyze this large dataset on their personal computers, each of these steps include a brief discussion of absolute minimum computing requirements, as well as detailed workarounds and shortcuts for the more memory-intensive processes.

Population Studied: The ACS represents all civilian, noninstitutionalized Americans. The examples used in this presentation include state and regional estimates; however, all instructions and syntax are presented with the intention of allowing the researcher to re-define a population of interest with minimal effort.

Principal Findings: Depending on research budget, computing resources, and level of programming skill, conducting analyses of the ACS with R often presents a viable alternative to other statistical analysis packages.

Conclusions: Given the large file size of the ACS, interested health policy researchers may be limited in their ability to analyze this survey with off-the-shelf statistical software packages due to memory overload issues. Although R users often experience similar memory limits and problems, the flexibility of the core R programming language and its integration with both parallel processing engines and Structured Query Language (SQL) allows for the relatively straightforward analysis of large, complex-sample survey data such as the ACS.

Implications for Policy, Practice or Delivery: Providing health policy researchers and statisticians with a strategy to work with the American Community Survey using freely available software will increase their ability to examine a variety of health and demographic indicators at the sub-national level. By outlining the technical steps to analyze this data, researchers could study topics such as regional characteristics of Health Professional Shortage Areas, state demographic factors associated with Medicare Advantage plan premiums, or county-level demographics of uninsured populations. An understanding of the methods needed to work with the ACS will open up the field of geographic region-based analyses to health policy researchers.

Nonparametric estimation of a heaping mechanism for precise and heaped self-report data

Sandra D. Griffith[1,⋆], **Saul Shiffman**[2], **Daniel F. Heitjan**[1]

1. Department of Biostatistics & Epidemiology, University of Pennsylvania
2. Department of Psychology, University of Pittsburgh
⋆Contact author: sgrif@upenn.edu

Keywords: Digit preference, Coarse data, Rounded data, Smoking cessation, Measurement error

Open-ended numerical measures, often used in self-report to assess quantities or frequencies, exhibit a form of measurement error termed heaping. Heaping occurs when quantities are reported with varying levels of precision. Digit preference is a special case of heaping where the preferred values are round numbers. Daily cigarette counts, for example, commonly exhibit heaps at multiples of 20, and to a lesser extent, 2, 5, and 10, when measured by retrospective recall methods. Because heaping can introduce substantial bias to estimates, conclusions drawn from data subject to heaping are suspect. Several methods have been proposed to estimate the true underlying distribution from heaped data, but all depend on unverifiable assumptions about the heaping mechanism. A data set in which subjects reported cigarette consumption by both a precise method (ecological momentary assessment as implemented with a hand-held electronic device) and a more traditional, imprecise method (timeline followback, or periodic retrospective recall) motivates our method. We propose a nonparametric method to estimate the conditional distribution of the heaping mechanism given the precise measurement. We measure uncertainty in the heaping mechanism with a bootstrap approach. We describe our implementation of the method using *R* and graphically illustrate our results with the **ggplot2** package. Application to our data suggests that recall errors are a more important source of bias than actual heaping.

References

Griffith, S. D., S. Shiffman, and D. F. Heitjan (2009, November). A method comparison study of timeline followback and ecological momentary assessment of daily cigarette consumption. *Nicotine & tobacco research 11*(11), 1368–73.

Shiffman, S. (2009, September). How many cigarettes did you smoke? Assessing cigarette consumption by global report, Time-Line Follow-Back, and ecological momentary assessment. *Health psychology 28*(5), 519–26.

Wang, H. and D. F. Heitjan (2008, August). Modeling heaping in self-reported cigarette counts. *Statistics in medicine 27*(19), 3789–804.

Forecasting multivariate time series using the DLM package: An application to road traffic networks

Osvaldo Anacleto-Junior

Dept. of Mathematics and Statistics, The Open University, Milton Keynes, UK
Contact author: o.anacleto-junior@open.ac.uk

Keywords: **DLM** package, traffic networks, time series, graphical models

Traffic data have some characteristics that can be quite challenging to deal with from a statistical modelling perspective. To have a broad view of the traffic network under analysis, information have to be collected from a series of sites, which can generate a high-dimensional multivariate time series. Also, as management decisions have to be taken in real time during periods like rush hours, traffic systems require forecasts in a on-line environment.

Traffic flows in a road network can be modelled by a class of models known as graphical dynamic models. These models use a directed acyclic graph (DAG) in which the nodes represent the time series of traffic flows at the various data collection sites in a network, and the links between nodes represent the conditional independence and causal structure between flows at different sites. The DAG breaks the multivariate model into simpler univariate components, each one being a dynamic linear model (DLM). This makes the model computationally simple, no matter how complex the traffic network is, and allows the forecasting model to work in real-time. Also, graphical dynamic models can accommodate changes that may happen in a network due to external events like traffic accidents and roadworks, which can heavily affect traffic flows and also the structure of the network, leading to sudden changes in the data being analysed.

This poster will report an application of this class of model using the **DLM** package in a busy motorway junction in the UK. It will be shown the advantage of some package features over standard DLM model fitting approaches, focusing in particular on its numerically stable singular value decomposition-based algorithms implemented for filtering.

Who's in the Waiting Room? Modelling Multivariate Time Series of Counts of Patients to Hospital Emergency Departments

Sarah Bolt[1,⋆], **Ross Sparks**[1], **James Lind**[2]

1. CSIRO Mathematics, Informatics and Statistics, North Ryde, NSW, Australia
2. Gold Coast Hospital, Southport, QLD, Australia
⋆Contact author: sarah.bolt@csiro.au

Keywords: Multivariate, Times Series, Health Informatics.

While predictive tools are already being implemented to assist in forecasting the total volume of patients to Emergency Departments [Jessup et al. (2010)], early detection of any changes in the types of patients presenting would help authorities to manage limited health resources and communicate effectively about risk, both in a timely fashion. But before we are able to detect changes we must understand and model the expected counts of presentations for all possible subgroups. For example, we need to forecast the expected number of patients on a given day at a particular hospital, with a particular disease, of a particular age, sex, etc.

So our objective was to model this large collection of interdependent time series. This problem presented a number of issues including the sheer number of them and the fact that many of these time series had very low counts and displayed overdispersion relative to a Poisson model. Furthermore, since the goal of these models was to model away known behaviours, we had to incorporate effects including significant seasonality, day of the week effects and some strongly interacting variables.

The method we present here divided the problem into two different components. The first used a regression approach to model time series of aggregated counts. The second allocated proportions of these counts to subgroups using a binary regression tree analysis. This method thus drew together *R* functionality from two different areas:

- Analysis of times series of counts using `glm` **{stats}** [R Development Core Team (2010)], `glm.nb` **{MASS}** [Venables and Ripley (2002)] and `gamlss` **{gamlss}** [Rigby and Stasinopoulos (2005)]. The latter function allowed us to model counts with a negative binomial distribution and a modelled dispersion parameter.
- Allocation of counts to subgroups using the poisson regression tree approach of `rpart` **{rpart}** [Therneau et al. (2010)].

References

Jessup, M., M. Wallis, J. Boyle, J. Crilly, J. Lind, D. Green, P. Miller, and G. Fitzgerald (2010). Implementing an emergency department patient admission predictive tool. *Journal of Health Organization and Management 24*, 306–318.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics 54*, 507–554.

Therneau, T. M., B. Atkinson, and B. D. Ripley. (2010). *rpart: Recursive Partitioning*. R package version 3.1-48.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.

Algorithm for defining hospital stays

Katrine Damgaard[1*], Jon Helgeland[1],

1. Norwegian Knowledge Centre for the Health Services
*Contact author: kda@nokc.no

Keywords: Healthcare, hospitals, Multicore

Objective: The objective was to link hospital stays which constituted a complete chain of admissions for patients transferred between wards/department within/between hospitals, in order to compare mortality between hospitals

Material and methods: The Norwegian Knowledge Centre for the Health Services has developed a system for retrieving data from the patient administrative systems (PAS) at all Norwegian hospitals. By use of this system, we retrieved PAS data from each hospital for patients discharged during 2005-2009.

Each record corresponds to one admission at one ward within a department in a hospital. To define a complete patient stay, we have to aggregate the admissions; stays at one department, hospital stays and stays involving more than one hospital. The latter is important for patients transferred between hospitals to obtain their complete medical treatment history. All permanent residents in Norway have a personal identification number (PIN) which enables linking between hospitals.

We have developed an algorithm that concatenates the ward admissions to a chain of admissions. The function gives you all the levels of stays described above. The input to the function is a serial number, PIN, date of admission, date of discharge, hospital, department and further optional parameters. The function uses the package **multicore** in *R*, for the ability to use more than one processor for large datasets. One of the properties of the algorithm is that you can choose different time windows, or tolerances, for concatenating the ward admissions. We chose 24 hours as tolerance for our purpose.

Results: With the choice of 24 hours time window for difference between time of discharge and next admission, we found 10 485 022 hospital stays involving one or more hospitals out of 16 370 163 ward admissions from 3 304 546 patients.

Conclusion: The algorithm can be used to aggregate large patient administrative data sets, with acceptable running times. The tolerance limit can be adjusted to suit the purpose and the data at hand.

Using merror 2.0 to Determine Measurement Bias and Imprecision

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, Calibration, Bias, Imprecision, Grubbs' Estimator

Researchers often need to evaluate the agreement between $n=2$ or more devices (instruments, methods) based on measurements made jointly on *N* items (subjects). In the simplest case, each measurement X_{ij} for device *i* and item *j* is described as:

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

where α_i and β_i describe how device *i* distorts (biases) the unknown true values μ_j, and ϵ_{ij} is a Normally distributed random error with mean 0 and standard deviation (SD) σ_i for device *i*. μ could be distributed Normally with mean $\bar{\mu}$ and SD σ in some applications. For uniqueness, the geometric average of all of the β_i is constrained to equal one. Jaech (1985) describes many approaches to estimating the parameters including method of moments (MOM) and maximum likelihood (ML). Estimates of these parameters will provide calibration curves relating any two devices, for example.

For simple cases of 3 or more devices and only one measurement from each device, the **merror** package provides several useful functions. `ncb.od` (non-constant bias - original data) computes ML estimates of the bias and imprecision parameters based on the $N\times n$ rectangular array of measurements. You must have *at least n=3* devices (or independently repeat at least 1 of two devices) because otherwise there is not enough information to compute all the parameter estimates uniquely (without making very strong assumptions about the parameters). Essentially, `ncb.od` fits a measurement error structural equation model with just one latent variable (for μ). `ncb.od` will compute estimates of the α_i, β_i, σ_i for each device, and σ (with confidence intervals for the σ_i and σ). You should be aware, however, that only certain functions have meaning: 1) when the β_i do not all equal 1, then you need the ratios of the β_i to check for scale bias (devices have different sized measurement units, i.e., a non-constant bias or bias that changes with the level of the true value), 2) when the β_i all equal 1, then differences among the α_i estimate constant relative biases, and 3) the device imprecision σ_i must be divided by the corresponding β_i in order to be compared (otherwise, you are dealing with incomparable results from different measuring systems). Also included is the function `lrt` to do likelihood ratio tests of the β_i. To visualize the data, use `pairs.me` to make modified `pairs` plots where all the axis limits are the same for all plots with the line of equality (no bias) indicated. Finally, `precision.grubbs.cb.pd` and `precision.grubbs.ncb.od` are provided to compute estimates of the imprecision SD's based on, respectively, using paired data under the constant bias model, and using the original data under the non-constant bias model. To implement these functions, the Fortran code from Jaech (1985) was translated to *R*. Some minor computational mistakes in the Fortran code were fixed when re-writing in the *R* programming language.

References

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

Using Metabolomics to monitoring kidney transplanted patients: transplant from a living donor.

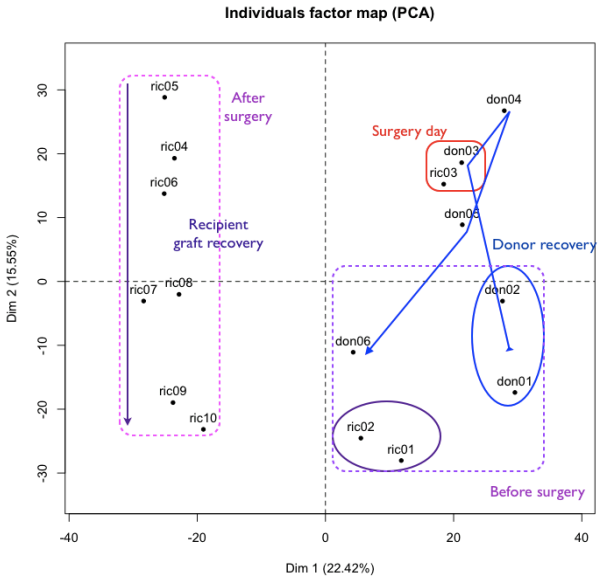
M. Calderisi[1] , A. Vivi[1] , M. Tassini[1] , M. Carmellini[2]

1. NMR Centre University of Siena, Italy
2. Department of Surgery and Bioengineering, University of Siena, Italy
*Contact author: m.calderisi@chemiometria.it
The work was partially supported by Consorzio Interuniversitario Trapianti d'Organo (www.consorziotrapianti.it)

Keywords: metabolomics, chemometrics, transplant, living donor, NMR

NMR spectroscopy is a analytical platform for biological matrices, it enables the quick generation of spectral profiles of low molecular weights metabolites. Metabonomics is a system approach used also to investigate the metabolic profile by multivariate data analysis tools. The aim of the work is to control the kidney graft recovery process in an non-invasive way, estimating the process by ^{1}H-NMR spectroscopy and chemometrics on urine samples as they contains metabolites concerning the pathological and clinical state of the patients. The goal of this statistical analysis is the identification of the dynamic biopatterns (individual patient recovery trajectories) related to the transplant. Samples have been collected every day, for both patients, along all the recovery period. The donor is a 60 years old healthy woman (6 samples), the recipient is his son, a 31 years old man (10 samples).

Data have been pre-processed (baseline correction, spectra alignment, water signal exclusion and normalization to unit length) and pre-treated (mean centering and log10 transformation). The log10 transformation has been especially useful to reduce the natural difference between high-intensity and low-intensity peaks, without boosting noise. Data have been analyzed by means of PCA and significant signal regions have been identified by selection of the highest loadings (in absolute value).



The scores plot in figure clearly shows the difference between the metabolic profile before (right side) and after (left side) the surgery (that happen during day 3) for recipient (ric) and his recovery trajectory occurring from day 5 to day 10. The metabolic profile of the donor (don) goes from a "normal" state (bottom right of the scores plot) to an altered state (top right) after the surgery day, then goes back to its original condition.

In conclusion, it is possible to state that NMR spectra plus chemometrics elaboration could be used to monitor a kidney graft recovery progress. Data elaboration has been carried out using **FactoMiner** and **ptw** and packages.

References

Francois Husson, Julie Josse, Sebastien Le and Jeremy Mazet (2010). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. R package version 1.14. http://CRAN.R-project.org/package=FactoMineR

Bloemberg, T. G. et al. (2010) "Improved Parametric Time Warping for Proteomics", Chemometrics and Intelligent Laboratory Systems, 104 (1), 65-74

hyperSpecGUI: Graphical Interaction in Spectroscopic Data Analysis

Sebastian Mellor[1], **Claudia Beleites**[2,⋆], **Colin Gillespie**[1], **Christoph Krafft**[2], **and Jürgen Popp**[2,3]

1. School of Mathematics & Statistics, Newcastle University, Newcastle/UK.
2. Institute of Photonic Technology, Jena/Germany.
3. Institute of Physical Chemistry and Abbe Center of Photonics, University Jena/Germany.

⋆Corresponding author: Claudia.Beleites@ipht-jena.de

Keywords: spectroscopy; Google Summer of Code; GUI.

Spectroscopic data analysis includes highly visual tasks that need graphical user interaction. We present **hyperSpecGUI**, a companion package enhancing **hyperSpec**. Both packages are hosted at hyperspec.r-forge.r-project.org.

The **hyperSpecGUI** packages follows the "specialise on a single thing" paradigm, as opposed to implementing an integrated development environment (IDE) or all-(spectroscopic)-purpose GUI. Visual interaction is provided by small, specialised GUIs (applets, dialogs) that are called as functions. This allows full GUI interaction with scripts and easy integration with IDEs. In particular, this allows:

Fast work-flow creation: rather than a single GUI, we use a combination of small, bespoke GUIs.

Flexibility: the package is extensible and easily adapted to new and unforeseen tasks.

Compatibility with batch processing: spectroscopic data sets are large with data analysis often including high performance computations on remote servers that do not provide graphical interaction. Swapping between GUI-based and command line based work/batch processing are smooth and easy.

Reproducibility: interactive tasks can be recorded, allowing literate programming techniques to be used, e.g. Sweave.

Interactive spike correction of Raman spectra: when cosmic rays hit the detector, spikes are observed in the Raman spectra. Several strategies to identify these artifacts exist and work well with high, sharp spikes. However, manual control and adjustment of the results is necessary as broader artifacts may be confused with sharp Raman bands and vice versa. Also the borders of broader artifacts are difficult to detect.

As an example, we present a GUI for spike filtering of Raman data, demonstrating the wrapping of this GUI with existing IDEs/editors.

Acknowledgements. SM's work is funded by the Google Summer of Code 2011. CB, CK, and JP acknowledge funding by the European Union via the Europäischer Fonds für Regionale Entwicklung (EFRE) and the "Thüringer Ministerium für Bildung, Wissenschaft und Kultur" (Project: B714-07037).

Visualisation for three-dimensional shape data

Stanislav Katina[1,⋆] **and Adrian Bowman**[1]

1. School of Mathematics and Statistics, The University of Glasgow, University Gardens, Glasgow G12 8QQ, Scotland, UK
⋆Contact author: Stanislav.Katina@glasgow.ac.uk [**Late Breaking Poster Abstract**, topic: **Visualization & Graphics**]

Keywords: Procrustes shape coordinates, thin-plate splines, 3D facial surfaces, shape index

In Late Breaking Poster we present visualisation of shape-space Principal Component Analysis (PCA) of (semi)landmarks from 3D facial surfaces. This multivariate model decomposes shape signal into low-dimensional linear combinations of high-dimensional measurements. Using Generalized Procrustes Analysis (Bookstein 1991), Procrustes Shape Coordinates (PSC) are calculated after suitable re-scaling, translation, and rotation of the data and then decomposed to affine and non-affine components. PCA is performed by singular value decomposition of the covariance matrix of centered PSC.

The data consists of 3D coordinates of 23 landmarks and 1664 semilandmarks on curves and surface patches taken from facial surfaces. Then, suitably scaled eigenvectors of the covariance matrix mentioned above were added to the symmetrised mean (reference shape) calculated by relabeling and reflecting the mean shape (semi)landmark coordinates. Using a carefully chosen mesh of 59242 points triangulated by 117386 faces and derived 1694 (semi)landmarks, we calculated standardized mesh points of the reference shape (called Thin-Plate Spline morph) by TPS interpolation model. This process is performed in an anatomically and geometrically meaningful manner, similar to the dense correspondence model of Mao et al. (2006) where the sum of principal curvatures is equivalent to bending energies. To explore the nature of particular PCs, we visualised 4D maps of signed pointwise Euclidean distances (in the sense of inwards/outwards facial direction) between shapes equivalent to extremes of the PC scores. Furthermore, these distances were decomposed to x, y, and z components of a suitably oriented face. Additionally, a shape index (SI) was calculated from principal curvatures derived from the Weingarten curvature matrix as a function of coefficients of a locally fitted parabolic regression model to facial coordinates (Goldfeather and Interrante 2004). The differences in SI between shapes equivalent to both extremes of PC scores were visualised as well.

The programs are written in R language (R Development Core Team 2011). The proposed visualisation is useful for interpretation of the effects of particular shape changes with direct practical implications in applied craniometrics and in anthropometrics more generally.

Acknowledgement

The research was supported by Wellcome Trust grant WT086901MA.

References

Bookstein, F.L., 1991: *Morphometric Tools for Landmark Data*. New York, Cambridge University Press

Goldfeather, J., Interrante, V., 2004: A novel Cubic-Order Algorithm for Approximating Principal Direction Vectors. *ACM Transactions on Graphics* **23**, **1**: 45-63

Mao, Z., Ju, X., Siebert, J.P., Cockshott, W.P., Ayoub, A., 2006: Constructing dense correspondences for the analysis of 3D facial morphology. *Pattern Recognition Letters* **27**: 597–608

R Development Core Team, 2011: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria

Bayesian Hierarchical Clustering for Microarray Time Series Data with Replicates and Outlier Measurements

Emma J Cooke[1], Paul D W Kirk[2], Richard S Savage[2*], David L Wild[2*]

1. Department of Chemistry, University of Warwick, Coventry UK
2. Systems Biology Centre, University of Warwick, Coventry UK
*Contact authors: r.s.savage@warwick.ac.uk, d.l.wild@warwick.ac.uk

Keywords: Systems biology, Bayesian hierarchical clustering

A key aim in systems biology is to link modelling of the interactions of system components with high throughput data. Time series experiments have become increasingly common, necessitating the development of novel analysis tools that capture the resulting data structure. **R/BHC** is a widely used package in *R*/Bioconductor for fast Bayesian hierarchical clustering of gene expression data (Savage *et al*, 2009). We present an extension to this package that enables time series data to be modelled and clustered, using Gaussian process regression to capture the structure of the data. Using a wide variety of experimental data sets, we show that our algorithm consistently yields higher quality and more biologically meaningful clusters than current state-of-the-art methodologies. Our approach permits a small proportion of the data to be modelled as outlier measurements, which allows noisy genes to be grouped with other genes of similar biological function. Our method exploits replicate observations to inform a prior distribution of the noise variance, which enables the discrimination of additional distinct expression profiles. These extensions will be included in the next release of **R/BHC**.

References

Savage, R. S., Heller, K., Xu, Y., Ghahramani, Z., Truman, W. M., Grant, M., Denby, K. J., and Wild, D. L. (2009). R/BHC: fast Bayesian hierarchical clustering for microarray data, *BMC Bioinformatics 10*, 242

Arbitrary Accurate Computation with R: Package 'Rmpfr'

Martin Mächler[1,2,⋆]

1. ETH Zurich (Seminar for Statistics), Switzerland
2. R Core Development Team
⋆Contact author: maechler@stat.math.ethz.ch

Keywords: Arbitrary Precision, High Accuracy, Multiple Precision Floating-Point, Rmpfr

The *R* package **Rmpfr** allows to use arbitrary high precision numbers instead of *R*'s double precision numbers in many *R* computations and functions.

This is achieved by defining S4 classes of such numbers and vectors, matrices, and arrays thereof, where all arithmetic and mathematical functions work via the (GNU) MPFR C library, where MPFR is acronym for "***M**ultiple **P**recision **F**loating-Point **R**eliably*". MPFR is Free Software, available under the LGPL license, and itself is built on the free GNU Multiple Precision arithmetic library (GMP).

Consequently, by using **Rmpfr**, you can often call your *R* function or numerical code with mpfr–numbers instead of simple numbers, and all results will automatically be much more accurate.

```
> options(digits = 17)# to print to full "standard R" precision
> .N <- function(.) mpfr(., precBits = 200)
> exp(   1 )

[1] 2.7182818284590451

> exp(.N(1))

1 'mpfr' number of precision  200   bits
[1] 2.7182818284590452353602874713526624977572470936999595749669679
```

Applications by the package author include testing of Bessel or polylog functions and distribution computations, e.g. for stable distributions. In addition, the **Rmpfr** has been used on the `R-help` or `R-devel` mailing list for high-accuracy computations, e.g., in comparison with results from commercial software such as Maple, and in private communications with Petr Savicky about fixing *R* bug `PR#14491`.

We expect the package to be used in more situations for easy comparison studies about the accuracy of algorithms implemented in *R*, both for "standard *R*" and extension packages.

References

Fousse L, Hanrot G, Lefèvre V, Pélissier P, Zimmermann P (2007). "MPFR: A multiple-precision binary floating-point library with correct rounding." *ACM Trans. Math. Softw.*, **33**, 1–13. ISSN 0098-3500. URL http://doi.acm.org/10.1145/1236463.1236468.

Fousse L, Hanrot G, Lefèvre V, Pélissier P, Zimmermann P (2011). *MPFR: A multiple-precision binary floating-point library with correct rounding*. URL http://mpfr.org/.

Granlund T, the GMP development team (2011). *GNU MP - The GNU Multiple Precision Arithmetic Library*. URL http://gmplib.org/.

Maechler M (2011). *Rmpfr: R MPFR - Multiple Precision Floating-Point Reliable*. R package version 0.3-0, URL http://rmpfr.r-forge.r-project.org/.

A dissimilarity based on relevant statistics

Antonio Miñarro[1,⋆], **Marta Cubedo**[1], **Josep M. Oller**[1]

1. Department of Statistics. University of Barcelona
⋆Contact author: aminarro@ub.edu

Keywords: Information metric, Rao distance, Multivariate distances.

A dissimilarity index between statistical populations in the absence of model assumptions is constructed. The index is based on some properties of the information metric as can be seen in Cubedo et al. (2011).

We are interested to define a distance between p different statistical populations $\Omega_1, \ldots, \Omega_p$, obtaining, for every statistical population, a data matrix $\boldsymbol{X}_i$ of order $n_i \times m$.

We shall assume that the statistical populations may differ in some vectorial parameter of interest $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^t$, although we shall not assume a particular parametric statistical model for the distribution of X.

We define certain convenient statistics $\boldsymbol{T} = (T_1, \ldots, T_k)^t$, that we assume are unbiased and consistent estimators of the parameters of interest $\boldsymbol{\xi}$ for each population. The dissimilarity index δ_{ij}, is defined by:

$$\delta_{ij} = \sqrt{\frac{1}{n_i}(\widehat{\boldsymbol{T}}_j - \widehat{\boldsymbol{T}}_i)^t \widehat{\boldsymbol{\Psi}}_i^{-1} (\widehat{\boldsymbol{T}}_j - \widehat{\boldsymbol{T}}_i)}$$

where $\widehat{\boldsymbol{\Psi}}_i$ is an unbiased estimation of $\boldsymbol{\Psi}_i = \mathrm{cov}(\boldsymbol{T}|\Omega_i) \quad i = 1, \ldots, p$, constructed by Bootstrap methods by re-sampling the rows of the data matrix $\boldsymbol{X}_i$.

$\Delta = (\delta_{ij})$ is a $p \times p$ non symmetric dissimilarity matrix between the populations $\Omega_1, \ldots, \Omega_p$. It is possible to approach Δ by a symmetric matrix D using the trace norm $D = (\Delta + \Delta^t)/2$

We have used simulations to assess the performance of the dissimilarity defined above, which we called *Relevant Statistics Dissimilarity* (RSD), in comparison to the usual Mahalanobis distance and to the Siegel distance between multivariate normal populations, Calvo et al. (2002).

An *R* script has been developed to perform simulations, bootstrap and compute dissimilarity matrices. In addition, a Multidimensional Scaling has been realized to visualize and compare the differences between our dissimilarity and Mahalanobis and Siegel distances. Implementing the methodology in a *R* package will be discussed.

References

Cubedo, M., Miñarro, A., Oller, J.M. (2011). Some Geometrical Remarks in Statistics. In *XIIIth Spanish Biometry Conference and 3rd Ibero-American Biometry Meeting, (Barcelona, Spain)*.

Calvo, M., Villarroya, A., Oller, J.M. (2002). A biplot method for multivariate normal populations with unequal covariance matrices. *Test 11*, 143–165.

BMEA: An R package for the analysis of Affymetrix® Exon Array Data

Stephen Pederson[1*], Gary Glonek[2], Simon Barry[1]

1. School of Paediatrics & Reproductive Health, University of Adelaide, SA, Australia 5005
2. School of Mathematical Sciences, University of Adelaide, SA, Australia 5005
*Contact author: stephen.pederson@adelaide.edu.au

Keywords: Bayesian Statistics, Bioinformatics, Exon Array, Gene Expression, Alternate Splicing

Microarrays have been shown to provide consistent & accurate estimates of comparative gene expression levels[1]. Early generations of Affymetrix® arrays use probes targeting the 3' end of an mRNA transcript and were restricted to detection of changes in gene expression levels. The newer generation of arrays, such as Exon arrays, use probes which target the entire length of a transcript and are additionally capable of detecting changes in gene structure via alternate splicing.

Building on the earlier probe-level modelling[2] approach for 3' Arrays, Bayesian Modelling for Exon Arrays as implemented in the **BMEA** package, has been developed using a hierarchical Bayesian model for detection of differentially expressed genes and alternate splicing events when applied to Exon Array data. Comparison of the BMEA approach with existing methods such as FIRMA[3] shows a strong improvement in performance for detection of both differentially expressed genes & alternate splicing events, and through a lack of reliance on annotated transcripts, will serve to complement existing Bayesian approaches such as **MMBGX[4].**

References

1. Robinson, M. and Speed, T (2007) A comparison of Affymetrix gene expression arrays *BMC Bioinformatics* **8** (449)
2. Bolstad, BM (2004) Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization. PhD Dissertation. University of California, Berkeley.
3. Purdom, E., Simpson, K., Robinson, M., Conboy, J., Lapuk, A. and Speed, T (2008) FIRMA: a method for detection of alternative splicing from exon array data. *BMC Bioinformatics* **24**(15):1707–1714
4. Turro, E., Lewin, A., Rose, A., Dallman, M. and Richardson, S (2009) MMBGX: a method for estimating expression at the isoform level and detecting differential splicing using whole-transcript Affymetrix arrays *Nucleic Acids Research* **38**(1):e4

Interactive Applets for Teaching with R

Andrew Rich, Danny Kaplan, Randall Pruim, Nicholas J. Horton, Joseph J. Allaire

The perceived lack of "user friendliness" in R is sometimes cited as an obstacle to the adoption of R in teaching. Many teachers believe a mouse- and menu-driven graphical interface is essential to success in the classroom, while others think that learning a notation of computer commands is an important component of a student's education in the contemporary world shaped by powerful computation and ubiquitous data. There are elements of truth in both views, suggesting that a synthesis is appropriate: providing interactive graphical "applets" within the context of a command-driven system.

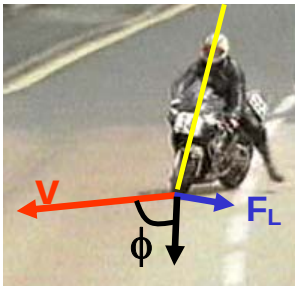
In this poster, we describe an easy-to-program facility for making easy-to-use interactive "applets" in R using the `manipulate` package in the RStudio interface for R. We have been developing a suite of such applets for teaching introductory statistics, calculus, and epidemiology. Some examples: a graphical system for selecting model terms in statistical models; a display of a mathematical function together with its derivatives and integrals to show the relationship among them; an interface for exploring parameters in SIR models to demonstrate the different possible outcomes of in the spread of infectious disease. Together with a set of command-line utilities for teaching statistics and calculus provided through the `mosaic` package, the applets should increase the attractiveness of teaching with R, and the `manipulate` package can empower instructors, even those with modest programming abilities, to create their own applets.

How engineers will learn *R* from motorcycle tires !

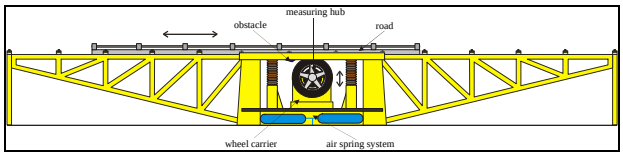
Koo Rijpkema*

Eindhoven University of Technology, Department of Mathematics and Computer Science, the Netherlands
*Contact author: j.j.m.rijpkema@tue.nl

Keywords: Statistics Education, Design of Experiments, packages **BHH2**, **DoE.base**, **DoE.wrapper**, **rsm**



For 1st and 2nd years' students from the Mechanical Engineering department at the Eindhoven University of Technology in the Netherlands the basic ideas of engineering statistics and statistical/experimental reasoning are introduced through a Design-Based Learning project. In this project about 15 groups of 7 students each are (simultaneously) introduced to the problem of instability of motorcycle road handling, such as wobble. Characteristics of the motor cycle tire play an important role in this and students have to find out through real physical experiments how pressure of the tire and vertical load influence these characteristics. For conducting the experiment they can use facilities we have available in the Automotive Lab at the Mechanical Engineering department, more specifically the flat-plank experimental setup that is available for measuring both static and dynamic characteristics of tires.



The experimental set-up used in the project.

In the project a 'distributed experimental set-up' is performed, where different groups measure characteristics at different factor-levels combinations. Data obtained by individual groups have to be combined, explored and analysed, aiming at an adequate fit of a Response Surface Model, to be used for prediction. As this project is the students' first introduction to modern computer aided statistical analysis they will primarily use the ***R Commander*** interface. However, the project serves as a start-up for more advanced courses on Engineering Statistics and Design of Experiments [1], where they will use *R* for more customized analyses, using and adapting contributions from relevant packages such as **BHH2**, **DoE.base**, **DoE.wrapper** and **rsm** [2,3,4].

References

[1] Box, G.E.P., J.S. Hunter and W.G. Hunter, (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd Edition. New York: Wiley.
[2] Groemping, U. (2009). Design of Experiments in R . Presentation at UseR! 2009 in Rennes, France
[3] Lenth, R.V. (2009). Response-Surface Methods in R, Using rsm . *Journal of Statistical Software* **32** (7), 1-17
[4] Vikneswaran (2005). An R companion to "Experimental Design". URL http://CRAN.R-project.org/doc/contrib/Vikneswaran-ED_companion.pdf

plotKML: a framework for visualization of space-time data in virtual globes

Pierre Roudier[1,*], **Tomislav Hengl**[2], **Dylan Beaudette**[3]

1. Landcare Research, New Zealand
2. ISRIC – World Soil Information, The Netherlands
3. Natural Resources Conservation Service, USDA, USA

*Contact author: roudierp@landcareresearch.co.nz

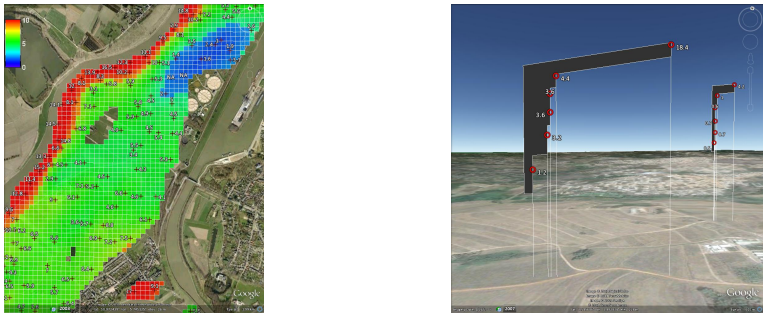
Keywords: Spatial data, data visualisation

KML (formerly Keyhole Markup Language), an XML-based format, is an OGC (Open Geospatial Consortium) standard to represent spatial data. Use of KML has been popularized with the development of virtual globe Google Earth (McGavra et al. 2009). Since it approval as an OGC standard, support for KML has been introduced in other virtual globes and geographic information systems (Craglia et al. 2008).

KML support is already available in *R* through the GDAL external library, whose bindings are provided the **rgdal** package, or directly implemented in *R* (packages **maptools** and **raster**). However, the support of the KML specifications is only partial. On the other hand, there is an increasing demand for tools to visually explore spatio-temporal patterns in a variety of environmental data (Andrienko & Andrienko, 2006).

The **plotKML** package provides the user with tools to plot spatial and spatio-temporal objects as described by **sp** classes into a KML file, and proposes an easy syntax to make complete use of the KML specifications. It aims at bringing virtual globes as a full-featured plotting canvas for spatial data, and thus provide not only methods to write various geometries in KML, but also plotting essentials, such as legend frames and metadata subsets.

Three levels of functions are provided: (i) a general, user-friendly `kml()` method, (ii) a set of intermediary functions to generate a KML file layer by layer, and (iii) and advanced templates for creating more advanced KML files.



(a) (b)

Figure 1: Examples of **plotKML** outputs. (a) Multi-layer visualisation of predictions and sampling locations. (b) Soil profile plot showing changes of measured soil organic carbon with depth.

Andrienko, N., and Andrienko, G. (2006) Exploratory analysis of spatial and temporal data: a systematic approach. Springer, 703 p.

Craglia, M., Goodchild, M., Annoni, A., Camara, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S., and Parsons, E. (2008). Next-Generation Digital Earth: A position paper from the Vespucci Initiative for the Advancement of Geographic Information Science. *International Journal of Spatial Data Infrastructures Research* **3**, 146–167.

McGavra, G., Morris, S., and Janée, G. (2009). *Technology Watch Report: Preserving Geospatial Data*. DPC Technology Watch Series Report 09-01. Digital Preservation Coalition, York, UK.

1+1=3: Combining cross-sectional multiple test data to find the historic shape of an epidemic

Gustaf Rydevik[1, 2, 3*], Mike Hutchings[2], Giles Innocent[1], Glenn Marion[1], Piran White[3]
1. Biomathematics and Statistics Scotland
2. Scottish Agricultural College
3. University of York
*Contact author: gustaf.rydevik@gmail.com

Keywords: Disease surveillance, Empirical Likelihood, EM, Bayesian Statistics, Biostatistics

We have written a set of *R* functions to simulate, estimate and visualize the historic incidence of pathogen infections, given a cross-sectional sample where two or more tests with different characteristics have been taken of each subject. We use the example of antibody and nucleic acid tests to illustrate this scenario.

In the surveillance of infectious diseases, it is often the case that resources and practical issues limit data collection to a prevalence study at a single point in time. Previous studies have used antibody titers to calculate incidence rates for various diseases, including Salmonella [Simonsen et al.], HIV [Heisterkamp SH et al.] and Pertussis[Hallander et al.]. However, to our knowledge, no published work has looked at combining antibody data with, for example, nucleic acid or culture data when calculating incidence rates.

The effect of the time interval between infection and test is different for different diagnostic tests. For example, antibodies tend to remain for a long time after infection, while tests designed to demonstrate the presence of a pathogen such as cultures or nucleic acid tests are much more sensitive to time elapsed since infection. By exploiting this difference, we can get an idea of whether the incidence is stable, on the increase, or decreasing, using only a single prevalence study.

We have developed a model for this situation which combines five different components: the incidence distribution over time, the distribution of antibodies over time after infection, the distribution of nucleic acid (or some other indicator) over time after infection, and the observed distributions of antibodies/nucleic acid in the cross-sectional sample.

Using a combined Bayesian/empirical likelihood approach, building on functions from the **mclust** package, we then use the observed distribution of antibodies and nucleic acid together with a prior for the infection incidence, to generate a shifted, posterior estimate of how the historical incidence behaved.

Because the results of these analyses differ depending on the specific relationship between antibodies and nucleic acid in the studied pathogen, simulation tools built on the **simecol** package are used for generating data to test how our approach works across various different antibody, nucleic acid and incidence distributions.

Acknowledgement: This work is being conducted within the FP7 WildTech Project (http://www.wildtechproject.com).

References

Simonsen, J., Mølbak, K., Falkenhorst, G., Krogfelt, K. A., & Linneberg, A. (2009). Estimation of incidences of infectious diseases based on antibody measurements. *Statistics in medicine*, Volume 28, Issue 14, pages 1882–1895,

Heisterkamp SH, de Vries R, Sprenger HG, Hubben GAA, Postma MJ(2008). Estimation and prediction of the hiv- aids-epidemic under conditions of Haart using mixtures of incubation time distributions. *Statistics in Medicine* 2008; 27:781–794.

Hallander, H. O., Andersson, M., Gustafsson, L., Ljungman, M. and Netterlid, E. (2009), Seroprevalence of pertussis antitoxin (anti-PT) in Sweden before and 10 years after the introduction of a universal childhood pertussis vaccination program. *APMIS*, 117: 912–922

A Package for Temporal Disaggregation of Time Series

Christoph Sax[1,2,*], **Peter Steiner**[1,3]

1. State Secretariat for Economic Affairs Seco, Switzerland
2. University of Basel, Switzerland
3. University of Bern, Switzerland
*Contact author: c.sax@seco.admin.ch

Keywords: Econometrics, Time Series, Temporal Disaggregation, Chow-Lin

Temporal disaggregation methods are used to disaggregate and interpolate a low frequency time series into a higher frequency series, while either the sum or the average of the resulting high-frequency series is consistent with the low frequency series. Disaggregation can be performed with or without the help of one or more indicator series; most indicator-based methods are variants of the method proposed by Chow and Lin (1971). Currently, there is no package in *R* to perform these tasks. **tempdisagg** offers a solution for the general task of temporally disaggregating a time series in *R*.

Use cases

Researchers in need of high frequency data can generate these data by themselves. For example, estimating a VAR (package **vars**) requires all variables to have the same frequency. **tempdisagg** allows the construction of a missing series by the researcher. In many European countries, quarterly figures of Gross Domestic Product (GDP) are estimated by statistical agencies using temporal disaggregation methods. Thanks to **tempdisagg**, at the State Secretariat for Economic Affairs, we are able to estimate Switzerland's official quarterly GDP figures almost fully automated in *R*.

Implementation

tempdisagg implements the standard methods for temporal disaggregation: Denton, Chow-Lin, Fernandez and Litterman. Denton, in its most basic form, does not require additional information besides the low-frequency series, Chow-Lin, Fernandez and Litterman, on the other hand, are using one or more indicators. **tempdisagg** is not restricted to annual or quarterly data, it works as long as the ratio of high to low frequency is an integer number (e.g. biannual to monthly data).

The selection of a temporal disaggregation model is similar to the selection of a linear regression model. Thus, `td`, the main function of **tempdisagg**, closely mirrors the working of the `lm` function (package **stats**), including taking advantage of the `formula` interface. Unlike `lm`, `td` can handle `ts` and `mts` time-series objects, as a typical use-case involves these objects. Nevertheless, it also can handle vectors and matrices in the same way as `lm`. `td` is a general function for temporally disaggregating time series and invokes particular methods (Denton, Chow-Lin, ...) based on the method argument. Internally, `td` uses the `optimize` function (package **stats**) to solve the one-dimensional optimization problem, which is at the core of some variants of the Chow-Lin procedure.

References

Chow, G. C. and A.-l. Lin (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The Review of Economics and Statistics 53*(4), 372–375.

cloudnumbers.com - Your Calculation Cloud for R

Markus Schmidberger[1,*], Moritz von Petersdorff-Campen[1], Markus Fensterer[1], Erik Muttersbach[1]

1. Cloudnumbers.com
*Contact author: markus.schmidberger@cloudnumbers.com

Keywords: R, cloud computing, computer cluster, high performance computing

Cloudnumbers.com provides scientists, as well as companies, with the resources to perform high performance calculations in the cloud. We help our customers to fight epidemics, develop highly advanced drugs and manage financial risk. Our aim is to change the way research collaboration is done today by bringing together scientists and businesses from all over the world on a single platform. *Cloudnumbers.com* is a Berlin, Germany based international high-tech startup striving for enabling everyone to benefit from the High Performance Computing (HPC) related advantages of the cloud. We regard reducing the difficulty of accessing the pooled computational power and abolishing the need for initial infrastructure investments as the key challenges in achieving our mission.

We pursue our goals by using the innovative potential of cloud computing. Leveraging the power of several pooled computer capacities and providing these on-demand via the Internet allows for a more efficient capacity utilization and highly elevated computational power. The cloud exhibits the possibility to speed up CPU and memory intensive calculations performed with applications like the well-known open-source statistics program **R** (www.r-project.org). Additionally, the cloud's characteristics predestine it for being the optimal mean for collaborative data sharing. Cooperation of project members, academics or even business networks thus is facilitated in an unprecedented way with regard to HPC calculations. User created knowledge databases as well as access to the most prominent public data resources take the collaboration and cooperation potential to yet another level and scale. All these features are integrated into the safe and efficient platform solution available on a pay-as-you-go basis at *cloudnumbers.com*.

Currently we focus on **R** (CRAN and Bioconductor packages are available). Computer clusters running with the packages **multicore**, **snow**, **Rmpi** or **snowfall** are available out of the box. At this point our main goal is to get feedback and consumer insights. That is why we will provide you with some free credit in order to allow you to test our service :

http://www.cloudnumbers.com

References

M. Schmidberger, M. Morgan, D. Eddelbuettel, H. Yu, L. Tierney, U. Mansmann; *State of the Art in Parallel Computing with R*; Journal of Statistical Software 2009: Vol. 31, Issue 1, Page 1-27;

A tentative implementation of *R* for finding the non-dominated fronts in multi-objective optimization

Ching-Shih Tsou[1,*] , Bo-Han Wu[1], Ying-Hao Lee[1], Ya-Chu Chiang[1]

1. Institute of Information and Decision Sciences, National Taipei College of Business, Taipei 10051, Taiwan, R.O.C.
*Contact author: cstsou@mail.ntcb.edu.tw

Keywords: multi-objective optimization, non-dominated solutions

Most decisions involve more than one conflicting objectives. This leads to the concept of non-dominance in multi-objective optimization that is different from the optimality concept of single objective one. Non-dominance means that the improvement of some objective could only be achieved at the expense of other objectives. Practically speaking, the effort on multi-objective optimization should be made in finding the non-dominated solutions archived as the non-dominated set by considering all objectives to be equally important. However, generating the non-dominated set could be computationally expensive and is often inefficient. This paper intends to implement three algorithms, the naive and slow, continuously updated, and an efficient method proposed by Kung et al., in *R* for finding the non-dominated fronts. Comparisons are made in order to highlight the applicability of *R* for building feasible multi-objective optimizers.

References

Deb, K. (2001). *Multi-objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, pp.33-43.

Deb, K., Agrawal, S., Pratap, A., and Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In *Proceeding of the 6th International Conference on Parallel Problem Solving from Nature*, Pages 849-858.

Kung, H., Luccio, F., and Preparata, F. (1975). On finding the maxima of a set of vectors. *Journal of the Association Computing Machinery* 22(4), 469- 476.

A hybrid machine learning algorithm for unlabeled gene expression data classification

Gokmen Zararsiz[1,*], Ahmet Ozturk[1], Erdem Karabulut[2], Ferhan Elmalı[1]

1. Department of Biostatistics and Medical Informatics, Erciyes University, Kayseri, Turkey
2. Department of Biostatistics, Hacettepe University, Ankara, Turkey
*Contact author: gokmenzararsiz@hotmail.com

Keywords: unlabeled data, clustering, classification, gene expression datasets

In gene expression data analysis, classification algorithms are widely used to classify biological samples and to predict clinical or other outcomes. But, these algorithms can not be used directly, if the data is unlabeled. Ignoring unlabeled data leads information loss and labeling them manually is a very difficult and expensive process. There are semi-supervised algorithms which were produced to label the unlabeled data. However, these algorithms are impractical in wholly unlabeled datasets. Thus, it is very significant to generate the class label for this kind of wholly unlabeled datasets and clustering algorithms can be used to obtain such labels. We proposed a hybrid machine learning algorithm for gene expression datasets to determine the best clustering with an optimum cluster number, then classify the dataset using new generated class labels. In this project, we will present the applicability and effectiveness of this algorithm by using several *R* packages: **clValid** [Brock *et al.,* 2011], **RankAggreg** [Pihur *et al.,* 2009], **e1071** [Dimitriadou *et al.,* 2011].

Acknowledgement
The project described is financially supported by Research Fund of the Erciyes University (Project Number: TSY-11-3587).

References

D. Meyer (2001). Support Vector Machines The Interface to libsvm in package e1071. *R News Volume 1/3*, 23-26.

G. Brock, V. Pihur, S. Datta, and S. Datta (2008). clValid, an R package for cluster validation. *Journal of Statistical Software*, 25(4).

J. Xie, C. Wang, Y. Zhang, S. Jiang (2009). Clustering Support Vector Machines for Unlabeled Data Classification. *2009 International Conference on Test and Measurement (Hong Kong, China)*, pp. 34-38.

V. Pihur, S. Datta, S. Datta (2009). RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics*, 10:62