Contents

Family-based analysis of genome-wide gene × gene interactions

Marit Ackermann[1,*] **, Andreas Beyer**[1]

1. Cellular Networks & Systems Biology, Biotechnology Center, TU Dresden

* Contact author: marit.ackermann@biotec.tu-dresden.de

Keywords: epistasis, gene × gene interaction, biostatistics

Complex diseases are caused by an interplay of several genetic alterations and environmental factors such as life style [1]. Recent advances in genomics and biotechnology have opened the gate to the genome-wide genotyping of thousands of possibly related individuals. Such data can now be used for studying epistatic genetic interactions at a genomic level.

While traditional family-based association or linkage studies are restricted to either a small number of markers or very specific pedigree structures, new methods for high-throughput data often disregard the inherent population structure leading to spurious findings of gene-gene interactions [2].

We propose an approach to infer genome-wide genetic interactions by using the genotype information of parent-child trios. Our method is applicable to very large data samples and a large number of markers. Instead of using the marginal frequencies of the observed alleles at two markers, we make use of inheritance patterns to infer the expected allele frequencies. Since the approach works conditional on ancestral genotype information it drastically reduces the number of false positive findings due to population effects. Moreover, we correct for the selection pressure against certain alleles that can also confound the results.

The approach is illustrated using a pedigree of almost 2300 mice that have been genotyped at more than $10,000$ SNPs. Results of our analysis and their biological significance will be discussed.

References

[1] Christopher A. Maxwell, Vctor Moreno, Xavier Sol, Laia Gmez, Pilar Hernndez, Ander Urruticoechea and Miguel Angel Pujana (2008). Genetic interactions: the missing links for a better understanding of cancer susceptibility, progression and treatment. *Molecular Cancer*, 7:4.

[2] Solomon K. Musani, Daniel Shriner, Nianjun Liu, Rui Feng, Christopher S. Coffey, Nengjun Yi, Hemant K. Tiwari, and David B. Allison (2007). Detection of Gene × Gene Interactions in Genome-Wide Association Studies of Human Population Data. *Human Heredity*, 63, 67–84.

The *rdyncall* package: An improved foreign function interface for R.

Daniel Adler[1,*], **Tassilo Philipp**[2]

1. Institut for Statistics and Econometrics, Georg-August University of Göttingen
2. Potion Studios

* Contact author: dadler@uni-goettingen.de

R provides a foreign function interface (`.Call()`, `.C()` and `.External()`) to invoke function calls to precompiled library code. The interface supports a very limited subset of possible argument and return types for a foreign function. For instance, there is no direct support for passing scalar arguments types from R to C functions. Hence, it is often necessary to write C wrapper functions to make a binding work, which can be a cumbersome process.

We present the R package *rdyncall* which provides an enhanced foreign function call interface to precompiled code, support for wrapping R functions into C callback objects, and R helper functions to work with C data structures. It can handle most C argument and return types, and performs automatic type conversions between R and C during calls and callbacks.

The package is implemented using the *dyncall* library that encapsulates architecture-, OS- and compiler-specific function-call semantics. For each class of function-call semantics — the so called *calling convention* — the library uses a small *call kernel* written in assembly. It has been ported to several architectures (currently x86, x86_64, ppc32, arm and mips) and multiple calling conventions (e.g. on x86: 'cdecl', 'stdcall', and gnu/microsoft 'fastcall' and 'this call').

A key concept in this package is the use of signature strings; type information is encoded as a compact text string and specifies the full semantics for calls and callbacks. This data format is easy to use, open for extensions and is also very efficient for low-level processing. As a neat side effect, signature strings can be regarded as portable representations for binding information across programming languages.

We show how *rdyncall* can be used to bind R with precompiled code without the need for additional C wrapper code. Examples include bindings to libSDL (a portable multimedia library), to OpenGL (3D graphics rendering) and to the R shared library itself (e.g. access to low-level R memory mangement and in-place sorting of atomic R vectors).

References

Adler, D. and Philipp, T. (2008). The *dyncall* library,
    http://www.dyncall.org/.

Lantinga, S. (1998). The Simple DirectMedia Layer library,
    http://www.libsdl.org/.

Segal, M. and K. Akeley. (1992). The OpenGL Graphics System. A Specification, Version 1.0.
    http://www.opengl.org.

sda: an R package for shrinkage discriminant analysis

Miika Ahdesmäki[1,2,*] **, Korbinian Strimmer**[1]

1. Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Härtelstr. 16-18, D-04107 Leipzig, Germany

2. Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland

* Contact author: miika.ahdesmaki@imise.uni-leipzig.de

Keywords: James-Stein shrinkage, linear classification, feature selection, prediction

We have developed a package ("sda") for R that implements multi-class linear discriminant analysis especially for high dimensional problems in omics-data. The package implements feature selection in a natural way without resorting to Monte Carlo or cross validation methods, thus achieving computational efficiency and also comparable or better performance than competing state-of-the-art high dimensional classifiers. The implementation relies on James-Stein-type shrinkage and uses local false nondiscovery rate (fndr) methodology to control the number of important features. The performance of our approach has been shown favourable in several applications including gene expression microarrays, metabolomic data and mass spec data - all of which are noisy and high dimensional in nature.

The package is easy to use and readily available in CRAN `http://cran.r-project.org/`. The implementation performs the data analyses and visualises the results automatically in an intuitive way, so that it is approachable also for the nonstatisticians.

References

Strimmer, K. (2009). R-package "sda",
   `http://cran.r-project.org/web/packages/sda/`.

Ahdesmäki, M. and Strimmer, K. (2009). A natural feature selection criterion for high-dimensional multi-class linear discriminant analysis. *Submitted.*

Empirical Transition Matrix of Multistate Models: The etm Package

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

1. Freiburg Centre for Data Analysis and Modelling, University of Freiburg, Germany
2. Institute for Medical Biometry and Informatics, University Medical Centre, Freiburg, Germany

* Contact author: `arthur.allignol@fdm.uni-freiburg.de`

Keywords: Aalen-Johansen estimator, competing risks, covariance matrix, current leukaemia free survival

When dealing with complex event history data in which individuals may experience more than one single event type, multistate models provide a relevant modelling framework. Well known examples include the competing risks model in which subjects may die from several possible causes, and the illness-death model that permits to study the impact of an intermediate event on a terminal event. Quantities of interest in this framework are the transition probabilities that can be estimated by the empirical transition matrix, that is also referred to as the Aalen-Johansen estimator. In this talk we present the R-package **etm** that computes and displays these transition probabilities. **etm** also features a Greenwood-type estimator of the covariance matrix, which has recently been found to be the preferable estimator in the competing risks situation. The use of the package is illustrated through a prominent example in bone marrow transplant for leukaemia patients.

References

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Klein, J., Szydlo, R., Craddock, C. and GoldmanJ. (2000). Estimation of current Leukaemia-Free Survival Following Donor Lymphocyte Infusion Therapy for Patients with Leukaemia who Relapse after Allografting: Application of a Multistate Model. *Statistics in Medicine*, 19, 3005–3016.

Putter, H., Fiocco, M. and Geskus, R. B. (2007). Tutorial in Biostatistics: Competing Risks and Multi-State Models. *Statistics in Medicine*, 26, 2389–2430.

Towards a R-centric architecture for multi purpose geographical analysis on heterogeneous multi-source data

Arlette Antoni[1*] , Thierry Dhorne[1] and Yann Le Guyadec[2]

1. Université Européenne de Bretagne, Université de Bretagne-Sud, CNRS, Lab-STICC, Centre de Recherche Yves Coppens BP 573, F-56017 Vannes cedex, France
2. Université Européenne de Bretagne, Université de Bretagne-Sud, Valoria, Centre de Recherche Yves Coppens BP 573, F-56017 Vannes cedex, France

* Contact author: arlette.antoni@univ-ubs.fr

Keywords: R-centric platform, Geographical Information Systems, R-software, Components Integration.

R provides an elegant and widely accepted platform with a rich function set for statistical data analysis, mathematical computations and flexible interactive graphics. The number of available R packages keeps growing at amazing speed, making it increasingly challenging for developers to deal with complex projects.

In particular, there is a big amount of packages related to spatial analysis in a broad sense encompassing spatial statistics, geometrical tools, maps management, ...

On the other hand, there has been some attempts in coupling/intergrating Statistical Systems with/in GIS. A thorough study showing the interest and the difficulties of integration of statistical and geographical information systems has been proposed in [1].

A complete and up to date reference adressing all this topics is provided by [2].

It is important to point out that some types of analysis require heavy and complex computations alternating between intensive use of data in the statistical and graphical analysis under the geographical information system.

We concentrate on a R-centric strategy and discuss the following requirements:

- Abstract data types that support the creation, access, and sharing of heterogeneous multi-source data
- Data coherency over coupled tools
- Management spatial or spatio-temporal data
- Database connectivity and web interface
- Integration with other softwares
- Robustness to data evolution
- Efficient vizualisation strategies

Some of this topics are adressed (conception and development of new R -types, possibility to extend the built-in functions,...) are adressed in the presentation.

References

[1] **de Andrade Neto P.R., Ribeiro P.J. and Fook K.D** (2005) Integration of Statistics and Geographic Information Systems VII Simpósio Brasileiro de Geoinformática, Campos de Jordão, Brasil, 20-23 novembro 2005, INPE,pp 139-155.

[2]**Bivand R.S., Pebesma E.J. and Gmez-Rubio V.** (2008) Applied Spatial Data Analysis with R. Springer, New York, 2008.

Ideas for Introducing Power in the Service Statistics Course

Alan T. Arnholt and Jose Almer T. Sanqui, Appalachian State University, Boone, NC, USA

Introduction

Power is one of the more important and least covered topics in an introductory Statistics course. This poster shows how power, even with non-central distributions, can be covered for students in a basic statistics course.

Prerequisite Concepts

Once a student understands just a few concepts, the *power of a test* can be introduced.

- $\alpha = \mathrm{P}(\text{type I error}) = \text{level of significance} = \mathrm{P}(\text{reject } H_0 | H_0 \text{ is true}) = \mathrm{P}(\text{accept } H_1 | H_0 \text{ is true}).$
- $\beta = \mathrm{P}(\text{type II error}) = \mathrm{P}(\text{fail to reject } H_0 | H_0 \text{ is false}) = \mathrm{P}(\text{accept } H_0 | H_1 \text{ is true}).$
- Given a composite alternative hypothesis $H_1 : \theta \in \Theta_1$, $Power(\theta) = \mathrm{P}(\text{reject } H_0 | H_0 \text{ is false}) = \mathrm{P}(\text{accept } H_1 | H_1) = 1 - \beta(\theta)$, where $\beta(\theta)$ is the probability of a type II error at a given θ.

Normal Distribution and Power

Computing the power for a particular alternative or finding the power function when working with normal distributions is covered in most texts and is easily done with R.

Problem: Given a normal distribution with unknown mean μ and known standard deviation $\sigma = 3$, for a test of the null hypothesis $H_0 : \mu = 48$ versus the alternative hypothesis $H_1 : \mu = 48$ using an α level of 0.05

A. With a sample of size one, compute the probability of a type II error.

B. Graph the $Power(\mu)$ for values of μ from 25 to 55 for testing a two tailed alternative hypothesis using samples of size one and nine.

Answers:

A. To find the the probability of a type II error, use the R commands

```
> cv <- qnorm(0.95, 40, 3)
> typeIIerror <- pnorm(cv, 48, 3)
> typeIIerror
[1] 0.1534347
```

B. To graph the $Power(\mu)$ for values of μ from 25 to 55 for testing a two tailed alternative hypothesis using samples of size one and nine, use the R commands:

```
> mu <- seq(25, 55, 0.01)
> powerONE <- pnorm(qnorm(.025, 40, 3), mu, 3) +
+ pnorm(qnorm(0.975, 40, 3), mu, 3, lower=FALSE) +
> powerNINE <- pnorm(qnorm(.025, 40, 1), mu, 1) +
+ pnorm(qnorm(0.975, 40, 1), mu, 1, lower=FALSE) +
> plot(mu, powerONE, type="l", lwd=2, col="blue",
+ ylim=c(0,1.1))
> lines(mu, powerNINE, col="blue", lwd=2)
```


Binomial Distribution and Power

So that students do not think that all power can be computed by shifting a **central** sampling distribution either to the right or left, introduce power using the binomial distribution.

Problem: Suppose $X \sim Bin(n = 14, \theta = 0.5)$. Determine the $Power(\theta = 0.85)$ when testing $H_0 : \theta = 0.5$ versus $H_1 : \theta > 0.5$ with $\alpha = 0.0897$ (a decision to reject the null hypothesis when $X > 9$).

Answer: By graphing a $X \sim Bin(n = 14, \theta = 0.5)$ and a $X \sim Bin(n = 14, \theta = 0.85)$, as shown below, it is very easy for the students to visualize the $Power(\theta = 0.85)$ with an asymmetric distribution.

To find the $Power(\theta = 0.85)$, use the R commands

```
> POWER <- sum(dbinom(10:14, 14, 0.85))
> POWER
[1] 0.9532597
```


Central and Non-central t Distribution

Simulation is an effective way to reinforce the concept of a sampling distribution.

Central t Distribution

Example: Have students simulate the quantity

$$tc = \frac{\overline{X} - \mu_1}{S/\sqrt{n}}$$

when sampling from a normal distribution. Compare the quantiles of the simulated sampling distribution versus the theoretical quantiles of a t_{n-1}.

The R simulation of 50,000 samples of size 16 from a normal distribution with mean of 100 and standard deviation of 20 is in the online script.

A density histogram of the quantity tc with a superimposed density of a t_{15} along with the theoretical and simulated quantiles suggest the simulation is a quite accurate representation of a t_{15} distribution:

Non-central t Distribution

To introduce the non-central t distribution with non-centrality parameter γ ($t^*_{n,\gamma}$), have the students simulate the quantity

$$tnc = \frac{(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet})}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (1)$$

The online script takes $m = 50{,}000$ samples from $N(\mu_1 = 120, \sigma_1 = 20)$ of size $n_1 = 16$ and from $N(\mu_2 = 100, \sigma_2 = 20)$ of size $n_2 = 25$.

The simulated values are displayed in a density histogram and a non-central t with non-centrality parameter $\gamma = 3.123475$ is superimposed over the simulated values.

The simulated values are counted to compute $Power(\mu_1 - \mu_2 = 20)$ for $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$ at the $\alpha = 0.05$ level (critical values shown in red above).

$$\widehat{Power}(\mu_1 - \mu_2 = 20) = \frac{\#\{tnc > t_{0.975,39}\} \cup \{tnc < t_{0.025,39}\}}{m} = 0.862$$

This agrees well with the theoretical power of the test (0.861).

Power could also be approximated using the simulation approach when the variances for the two populations are unknown and unequal (Behrens-Fisher problem).

The Non-centrality Parameter

For the non-centrality parameter, γ, and the t statistic, t,

$$\gamma = \frac{(\mu_1 - \mu_2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2}}{\sigma} \quad \text{and} \quad t = \frac{(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2}}{S_p} \quad (2)$$

t measures the statistical differences between the **sample** means and γ is used to measure the statistical differences between the **population** means.

Squaring both quantities in (2), gives

$$F = t^2 = \frac{(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet})^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{S_p^2} = \frac{MS_{\text{Treatment}}}{MS_{\text{Error}}}$$

and

$$\lambda = \gamma^2 = \frac{(\mu_1 - \mu_2)^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1}}{\sigma^2} = \frac{SS_{\text{Hypothesis}}(\text{population})}{\sigma^2}$$

where $SS_{\text{Hypothesis}}(\text{population})$ is the sum of squares for treatments obtained by replacing $\overline{Y}_{i\bullet}$ with μ_i, $\overline{Y}_{\bullet\bullet}$ with $\mu_{\bullet\bullet}$ and $\overline{Y}_{\bullet\bullet}$ with $\frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2}$.

When λ is the ratio of $SS_{\text{Hypothesis}}(\text{population})$ to σ^2, the calculation of λ is straightforward. The $SS_{\text{Hypothesis}}(\text{population})$ will always be the sum of squares formula for the H_0 being tested.

> This method of computing λ extends to any hypothesis the user would like to test. It is not limited merely to the equality of treatment means nor to equal sample sizes.

To compute the power of the test when $\mu_1 - \mu_2 = 20$, $\sigma_1 = \sigma_2 = 20$, $n_1 = 16$, and $n_2 = 25$ using a two-sided alternative with $\alpha = 0.05$, compute the non-centrality parameter to be

$$\gamma = \frac{(\mu_1 - \mu_2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2}}{\sigma} = \frac{(120 - 100)\left(\frac{1}{16} + \frac{1}{25}\right)^{-1/2}}{20} = 3.123475,$$

The power of the test is then

$$Power(\mu_1 - \mu_2 = 20) = \mathrm{P}\left(\text{Reject } H_0|H_1\right)$$
$$= \mathrm{P}\left((T < t^*_{n,(2|n_1+n_2-2)})\big|T \sim t^*_{n_1+n_2-2,\gamma}\right) +$$
$$\mathrm{P}\left((T > t_{(1-\alpha/2|n_1+n_2-2)})\big|T \sim t^*_{n_1+n_2-2,\gamma}\right)$$
$$= \mathrm{P}((t^*_{39,3.123475} < 0.025;39)) + \mathrm{P}((t^*_{39,3.123475} > t_{975,39}))$$
$$= \mathrm{P}((t^*_{39,3.123475} < -2.022691)) + \mathrm{P}((t^*_{39,3.123475} > 2.022611)) = 0.8612027$$

R commands

To find $Power(\mu_1 - \mu_2 = 20)$ with R, one can use the standard commands pt() and qt() as follows:

```
> cv1 <- qt(0.025, 39)
> cvu <- qt(0.975, 39)
> Power <- pt(cv1, 39, 3.123475) +
+ pt(cvu, 39, 3.123475, lower.tail=FALSE)
> Power
[1] 0.8612027
```

The function power.t.test() will return power for one- and two-sample t tests (when each sample is the same size) and power.anova.test() will return the power for one-way analysis of variance problems when the sample sizes are equal. It would not be hard to modify the current code to either function to accommodate unequal sample sizes as special cases.

Non-central F Distribution

Suppose the true mean grade for University A's students using teaching methods 1, 2, 3, and 4 have means of 71, 73, 75, and 80 with a common standard deviation of $\sigma = 12$. If $n_1 = 11$, $n_2 = 13$, $n_3 = 10$, and $n_4 = 12$, determine the probability a difference among the means will be detected using $\alpha = 0.05$.

$$\lambda = \frac{SS_{\text{Hypothesis}}}{\sigma^2} = \frac{\sum_{i=1}^{4} n_i(\mu_{i\bullet} - \mu_{\bullet\bullet} - \bar\mu_{\bullet\bullet})^2}{\sigma^2}$$
$$= \frac{11(71 - 74.78)^2 + 13(73 - 74.78)^2 + 10(75 - 74.78)^2 + 12(80 - 74.78)^2}{12^2}$$
$$= 3.65157$$


```
> Power <- 1 - pf(qf(0.95, 3, 42), 4, 42, 3.65157)
> Power
[1] 0.2204405
```

Conclusion

With appropriate examples, power can be covered with R in an introductory statistics course. When students use simulation, they gain an intuitive understanding of power, even for non-central distributions.

References

Ugarte, M. D., Militino, A. F., and Arnholt A. T. 2008. *Probability and Statistics with R*. Chapman & Hall/CRC. Boca Raton, FL.

[An R script for all graphs in the poster is available at http://www1.appstate.edu/~arnholta/UseR2009]

logi.DIAG

High-Volumn Real-time Data

Thomas Baier[1,*] **,Erich Neuwirth**[2]

1. logi.cals
2. University of Vienna
* Contact author: thomas.baier@logicals.com

Keywords: automation, industry, data mining, real-time

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

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

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

As automation systems are getting more and more complex, the size of data sets is increasing far more than the size of the applications. New approaches requiring statistical methods to handle the large amounts of data will have to be established on the market in just a few years. Partners from both industry and academia are working on the funded research project "logi.DIAG" to find solutions for systems with such increasing complexity. See logi.DIAG (2008) for more information.

In contrast to typical applications of data-mining (or even data-warehouse systems), PLCs are very limited in both computing speed and memory (typically a few hunded kilobytes of RAM and no persistent storage at all). Therefore one of the issues is finding ways for data compression and storage which takes these kinds of resources into account. The approach described by Chambers et al. (2006) has been considered especially useful for this case and its adoption for the specific problem domain is one of the first analysis steps in this project.

R is been used for prototyping implementations and during analysis and adoption of algorithms and if possible for further analysis.

References

logi.DIAG (2008). logi.DIAG. Test-Driven Automation.
    http://www.logidiag.at/.

Chambers et al. (2006). Monitoring Networked Applications With Incremental Quantile Estimation. *Statistical Science, 2006, vol. 21, p. 463,*

Mayday RLink – The best of two worlds

Florian Battke[†], Stephan Symons, and Kay Nieselt

Center for Bioinformatics Tübingen, University of Tübingen, Sand 14, 72076 Tübingen, Germany
† Corresponding author: battke@informatik.uni-tuebingen.de

Keywords: microarray analysis, visual analytics, R integration

DNA Microarrays are the standard method for large scale analyses of gene expression and epigenomics. Analysis software must keep pace with the increasing complexity of generated data. MAYDAY [1] is a free and flexible graphical workbench for visualization and analysis of microarray data. It is written in Java and can be used as fully-functional WebStart application on every major computing platform without any installation. New challenges can swiftly be met due to MAYDAY's plugin interface. Currently, MAYDAY includes a large variety of plugins for visual data exploration, clustering, machine learning and classification, as well as Gene Set Enrichment Analysis. MAYDAY can import data from several file formats, database connectivity is included for efficient data organization. Numerous interactive visualization tools, including box plots, profile plots, principal component plots, our enhanced heatmap [2], the use of metadata to enhance plots as well as the possibility to create publication quality images make MAYDAY a power analysis tool for microarray data.

MAYDAY offers an intuitive interface to work on experimental data and many methods for interactive *visual* data exploration. R, on the other hand, can be used to quickly create unconventional visualizations, it offers an extremely versatile shell allowing e.g. fast filtering of data by arbitrarily complex criteria as well as methods for interactive *computational* exploration of data. The wealth of R packages available, such as those from the BioConductor project, is an immensely valuable resource and one of the reasons R is used throughout the microarray community.

With RLink we provide an interactive and integrated approach to harness the power of R within the framework provided by MAYDAY, further increasing the power of our visual analytics platform. Using the `rJava` package [3] for R, we have integrated an interactive R shell into MAYDAY. Within this shell, users can directly work on MAYDAY's core data structures, and apply methods provided by R or R packages such as LIMMA. Results can either be passed back to MAYDAY in the form of new datasets or attached to the original data as meta-information. Plugins offered by MAYDAY can also be called from the R shell allowing sophisticated analyses by combining the methods both programs offer, manual as well as scripted. The implementation hides the syntactic complexities involved in using native Java objects within R, all interaction is done via R objects that behave like normal R vectors resp. matrices.

The combination of both programs integrates the best of both worlds providing researchers with a background in R the opportunity to quickly test new hypotheses or find out details about their data that MAYDAY's user interface does not provide. While for these needs new plugins could of course be developed for MAYDAY, this is often not desirable for one-time analyses specific to only one particular dataset. Thus, using the RLink shell alongside MAYDAY's graphical user interface, scientists can quickly gain a deeper understanding of their data.

MAYDAY and RLink are available at `http://www.zbit.uni-tuebingen.de/pas/mayday/`

References

[1] J Dietzsch, et al. (2006) MAYDAY – a microarray data analysis workbench. Bioinformatics 22:1010.

[2] N Gehlenborg, et al. (2005) A framework for visualization of microarray data and integrated meta information. Information Visualization 4:164.

[3] S Urbanek. rJava: Low-level R to Java interface

The R package `nlstools`: a toolbox for nonlinear regression

Florent Baty[1,*] , Sandrine Charles[2,3], Jean-Pierre Flandrois[2,3], Marie-Laure Delignette-Muller[2,3],

1. COPSAC, University Hospital Copenhagen, Gentofte, Denmark
2. University of Lyon, Lyon, France
3. CNRS UMR5558, Villeurbanne, France

* Contact author: florent.baty@gmail.com

Keywords: Nonlinear regression, diagnostics, confidence regions, bootstrap, jackknife

There is an increasing interest in the use of nonlinear regression models in a broad diversity of scientific fields (incl. chemistry, agricultural science, pharmacology, and microbiology). Various **R** functions are already dedicated to the fit of nonlinear models (Ritz & Streibig, 2008). The basic routine that provides nonlinear least squares estimates is the function `nls` from the `stat` package. The relative complexity of use of nonlinear optimization algorithms may prevent non-statisticians of using these models.

Unlike in linear regression, the fit of nonlinear models requires a great deal of attention regarding the definition of the starting values from which the algorithm will start its least-squares minimization procedure. Important issues associated with nonlinear regression relate, for example, to the assessment of the validity of the error model, the estimation of the parameters' confidence regions and confidence intervals, the identification of influencing observations (Bates & Watts, 1988; Huet et al., 2004). The available nonlinear regression modules lack some of these diagnostic functionalities, and there is a need to provide users with an extended toolbox of functions.

We developed the package `nlstools` which helps users to fit nonlinear regression models and provides a unified framework to test the error model assumptions and assess the quality of fit of the model. `nlstools` is designed to work directly with `nls` objects. This package includes a set of functions and graphical tools that will assist the user in creating `nls` objects and carrying out various diagnostics tests. These functions are organized as follows:

- Preparation of the fit of nonlinear regression models with `nls`: `preview`

- Summary of fit: `plotfit`, `overview`

- Validity of the error model assumptions: `nlsResiduals`

- Parameter's estimates confidence regions: `nlsConfRegions`, `nlsContourRSS`

- Confidence intervals and influencing observations using resampling techniques: `nlsBoot`, `nlsJack`

- Various examples of nonlinear regression models and illustrative datasets

Overall, the **R** package `nlstools` constitutes a useful add-on toolbox for nonlinear regression diagnostics. It is available on CRAN. Future developments are currently under way.

References

Bates DM & Watts DG (1988). Nonlinear Regression Analysis and its Applications, John Wiley & Sons, New York.

Huet S, Bouvier A, Gruet MA, Jolivet E (2004). Statistical Tools for Nonlinear Regression: A Practival Guide with S-plus and R Examples. Springer-Verlag, New York.

Ritz C & Streibig JC (2008). Nonlinear Regression with R. Springer, New York.

Population designs evaluation and optimization in R: the PFIM function

Caroline Bazzoli[1,*], Sylvie Retout [1,2], Emanuelle Comets[1], Hervé Le Nagard[1], France Mentré[1, 2]

1. INSERM U738 and Université Paris Diderot, Paris, France;
2. AP-HP, UF de Biostatistiques, Hôpital Bichat, Paris, France;
* caroline.bazzoli@inserm.fr

Keywords: Nonlinear mixed effects models; Fisher information matrix; Population design; Pharmacokinetics;

Nonlinear mixed effect model or population approach has been developed for the analyses of biological processes described by longitudinal data. It allows estimation of the mean value of the parameters in the population studies and their interindividual variability.

Population analyses often involve a limited sampling strategy in the data collection, mainly due to ethic or financial concerns. To face with the risk of unreliable results in such limited samples studies, efforts have been given since a decade to the development of a methodology for population designs evaluation and optimization based on the expression of the Fisher information matrix for nonlinear mixed effects models (Mentré et al. 1997; Retout et al. 2002; Bazzoli et al. 2007). In this context, we have proposed PFIM (Retout et al. 2003), a R function for population design evaluation and optimization.

PFIM evaluates population designs for single or multiple response models and thus returns the expected standard errors, defined as the square roots of the diagonal elements of the inverse of the Fisher information matrix, on the population parameters with the design evaluated. To use PFIM, some prior information has to be supplied by the user such as the structural model, its parameterization and a priori values of the parameters. PFIM can also optimize population designs with different optimization options, based on the D_optimality criterion, i.e. to maximize the determinant of the Fisher information matrix or minimize its inverse.

Since 2003, several releases of PFIM have been proposed. Currently, two main versions now are implemented in parallel: a graphical user interface package using the R software (PFIM Interface) and a direct R version. The latter requires knowledge in R programming but benefits of the latest methodological developments performed in our research team.

Examples of the use of both versions of PFIM will be presented. They are performed in the context of pharmacokinetics and pharmacodynamics data. Pharmacokinetics deals with the time-course of drug concentration, whereas pharmacodynamics refers to the time-course of drug action in the body (Retout et al. 2007).

PFIM versions and their extensive documentation are freely available on the PFIM website (Retout et al. 2007).

References

Bazzoli et al. (2009). Fisher information matrix for nonlinear mixed effects multiple response models: evaluation of the first order linearization using a pharmacokinetic/pharmacodynamics model. *Statistics in Medicine*. In press.

Mentré et al. (1997). Optimal design in random-effects regression models. *Biometrika*, 84(2), 429-442.

Retout et al. (2002). Fisher information matrix for non-linear mixed-effects models: evaluation and application for optimal design of enoxaparin population pharmacokinetics. *Statistics in Medicine*, 21(18), 2623-39.

Retout et al. (2003). Optimisation of individual and population designs using Splus. *Journal of Pharmacokinetics and Pharmacodynamics*, 30(6), 417-443.

Variance estimation in second cameroonian households survey

BEM Justin[1,2,*]

1. Central African Banking Commission
2. AERC Researcher
* Contact author: justin.bem@gmail.com

The lack of knowledge about variance estimation in complex survey and the lack resources to get of computing package usually lead sub-saharian statistic offices to abandon the task of variance estimation after point estimation in surveys. This article treats of variance estimation in complex survey and for complex statistics. It is our contribution to help researchers in poor countries to avoid difficulties in variance estimation. We use linearisation and replication methods to estimate variance of interest statistics in Second Cameroonian Households Survey (CHS 2). Finally, our computations carry out with R survey package show a good accuracy for survey estimators. The results are used to propose and optimal sampling design for the CHS 3 and to proceed to hypothesis tests. We show that CHS 3 can be realized with a smaller sample size without lost of accuracy.

References

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*,9(1),13–17.

Binder, D. A. and Kovacevic, M. (1993). Estimating some measures of income inequality from survey data : An application of estimating equation approach, Proceedings of the ASA Survey Research Methods 550-555.

The PTW package: Global Parametric Time Warping in R

Tom Bloemberg, Jan Gerretzen, Hans Wouters,
Lutgarde Buydens and Ron Wehrens

Radboud University Nijmegen, The Netherlands

Keywords: Alignment, Chemometrics, Time Warping

Chemometric analyses of chromatograms and spectra are often hampered by misalignments due to small changes in experimental parameters (column ageing in chromatography, pH differences in NMR, etc.). Several computational techniques have been proposed to correct for such shifts, notably Correlation Optimized Warping (COW[1]) and Parametric Time Warping (PTW[2]) which have become popular during the last years.

The widespread use of multivariate detection methods in chromatography and the development of new 'hybrid' or 'hyphenated' techniques like GC-MS and LC-NMR demand the development of global alignment methods, that use the multivariate nature of the detector to their advantage. An example of such a technique is COW-CODA[3], an expansion of COW including the selection of high-quality chromatographic traces by the COmponent Detection Algorithm (CODA[4]).

PTW can also be modified to make it capable of performing global alignments. Here we present the **PTW** package, containing an R impletation of the Parametric Time Warping algorithm, based on the original implementation by Eilers[2]. The algorithm has been expanded to include:

- an optional global alignment making full use of multivariate detection methods;

- the use of optimization methods from the **stats** package, enhancing the search for the optimal alignment beyond the original steepest descent method;

- a number of measures for assessing the alignment quality;

- a method for selecting the best traces to use for alignment based on an enhanced version of CODA;

- a number of visualization options.

References

1. Niels-Peter Vest Nielsen, Jens Michael Carstensen and Jørn Smedsgaard (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, 805, 17–35.

2. Paul H. C. Eilers (2004). Parametric Time Warping. *Analytical Chemistry*, 76, 404–411.

3. Christin Christin, Age K. Smilde, Huub C. J. Hoefsloot, Frank Suits, Rainer Bischoff and Peter L. Horvatovich (2008). Optimized time alignment algorithm for LC-MS data: Correlation optimized warping using component detection algorithm-selected mass chromatograms. *Analytical Chemistry*, 80, 7012–7021.

4. Willem Windig, J. Martin Phalp and Alan W. Payne (1996). A noise and background reduction method for component detection in liquid chromatography/ mass spectrometry. *Analytical Chemistry*, 68, 3602–3606.

cran2deb: A system to automatically provide 1500+ CRAN packages as Debian binaries

Charles Blundell[1], **Dirk Eddelbuettel**[2]

1. Gatsby Computational Neuroscience Unit, UCL
2. Debian Project

* Contact author: edd@debian.org

Keywords: R, CRAN, packages, Debian

This paper introduces the cran2deb system that provides automated builds of binary Debian packages from (essentially all) available CRAN packages for the R statistical environment and language.

Part of the growing popularity of R is due to the availability of over 1500 source packages at the CRAN repositories alone. These packages provide anything from R extension via new statistical methodologies, add approaches specific to given scientific disciplines, connect R to various backends such as databases, or provide specific user-interfaces. The BioConductor repository provides another few hundred packages specifically for bioinformatics research bringing the total to almost 2000 packages.

Users, however, are frequently stymied by the system administration task of building and compiling these source packages as this may entail obtaining and installing other toolchains or libraries from unknown third parties. This is often a multi-step process with manual intervention to resolve problems: source packages may not compile, dependencies may require further dependencies (that in themselves do not compile). The whole process may, after several hours, prove entirely fruitless. Linux distributions have shown how a single package management system, for all installed software, can help: universal control over the intricate details inter-package dependencies makes the system administration task significantly less arduous. Packages are already compiled and as the exact same package is used by many people, the user can have a higher expectation of an installation working first time. It would therefore be helpful to provide binary packages that are fully integrated into an existing Linux distribution such as Debian. Users could then use the existing package management tools to install, upgrade, query or remove packages. This becomes even more useful in large installations such as departments, work groups or computing clusters.

The cran2deb system fits into the infrastructure for the Debian GNU/Linux distribution, a Linux distribution with over 20,000 packages. cran2deb utilises the Debian toolchains for package building, in particular the pbuilder program to facilitate unsupervised building of packages in pristine build environments augmented with finely-grained build-dependencies. It uses a simple database backend (currently provided by SQLite) for stateful information and logging.

This work extends prior work by Eddelbuettel et al (2007). It is however a reimplementation using R as a scripting language, provided primarily by the first author. This was part of the Google Summer of Code (GSoC) program 2008, with the second author acting the mentor for this GSoC project. It has been extended further since the GSoC 2008 program finished.

The approach taken here is bottom-up: information about individual packages is analysed and aggregated into a dependency graph—the dependency graph is not just restricted to R packages; any Debian package may be included. If necessary, the minimal set of packages needed to fulfil the dependencies for a given package are then built alongside with the target package. The set of available packages is updated by querying the CRAN mirror network using R's internal tools, and new packages are compiled at each update pulse. Current versions of available packages are then provided in a downloadable repository.

We discuss some of the lessons learned in building this service, and the next steps that are needed to make cran2deb part of the CRAN network.

References

Eddelbuettel D, Vernazobres D, Gebhard A and Moeller S (2007), `apt-get install cran bioc:` On automated builds of 1700 R packages for Debian. Presentation at useR! 2007, Iowa State University, Ames, Iowa, August 8-10, 2007.

Network Text Analysis of R Mailing Lists

Angela Bohn[1,*]**, Ingo Feinerer**[2]**, Kurt Hornik**[1]**, Patrick Mair**[1]**, Stefan Theußl**[1]

1. Wirtschaftsuniversität Wien, 1090 Wien, Austria
2. Technische Universität Wien, 1040 Wien, Austria

* Contact author: angela.bohn@gmail.com

Keywords: Social Network Analysis, Text Mining, Network Text Mining, R Mailing Lists

In the worldwide R community, users and developers typically coming from a multitude of professions have one thing in common, namely the great interest in R. Due to the nature of shared interests they feel the need to discuss open questions, help other R users, and share their experiences. Hence, a large amount expert talks between developers on the one hand and feedback from users on the other hand assists the development of successful open source software. In the R world, these decentralized and geographically dispersed processes are supported by the R mailing lists R-help and R-devel. Thousands of authors write dozens of e-mails daily and their findings and information is shared not only with the subscribers, but also with even more internet users. Social network analysis (SNA) is able to reveal the writers' communication structure and find behavioral patterns. What's more, text mining (TM) allows examining the content of the great amount of e-mails. Despite the great potential, only few approaches to combine SNA and TM exist so far. In our poster, we show how such a combination, more precisely a "Network Text Analysis" of the R-help and R-devel mailing list, can help to gain even more insights into the process of open source development.

References

[1] Ingo Feinerer, Kurt Hornik, and David Meyer, *Text Mining Infrastructure in R*, Journal of Statistical Software **25** (2008), no. 5, 1–54.

[2] Andrew Mccallum, Andres Corrada-Emmanuel, and Xuerui Wang, *Topic and Role Discovery in Social Networks*, IJCAI '05: Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, 2005, pp. 786–791.

[3] Stanley Wasserman and Katherine Faust, *Social Network Analysis, Methods and Applications*, Cambridge University Press, 1997.

The Forecast of the Export Quantity of Thai Frozen Sea Food

Rujirek Boosarawongse

KMITL, Department of Applied Statistics, Faculty of Science, Bangkok 10520, Thailand.

rujirek@mozart.inet.co.th

Keywords: Forecasting, Validation Error, Decomposition, Exponential Smoothing, Box and Jenkins

The objective of this study is to determine the method and model which is appropriate to forecast the export quantity of four products of the frozen sea food the frozen giant tiger shrimps, the frozen shrimps, the frozen fish and the frozen cuttlefish. The 82 monthly data of each products from January 2002 to October 2008 are collected from the Office of Agricultural Economics, Ministry of Agriculture and Cooperatives. The following four forecasting techniques are used to analyze : Decomposition, Double Exponential Smoothing, Triple Exponential Smoothing, Box and Jenkins Method. The MSE MAPE and MAD are used to evaluate the fitted models. The best model is the one that gives the lowest values of MSE MAPEand MAD. The results show that the Double Exponential Smoothing method give the best forecasting model for all considered products.

References

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C.(1994). *Time Series Analysis, Forecasting and control,* 3rd ed. Englewood Cliffs, N. J., Prentice-Hall,.

Ljung, G. M. and Box, G.E.P. (1978). "On a measure of lack of fit in time series models", *Biometrika*, 65, 297-303.

Wei, W.S. (1990). *Time Series Analysis*, New York: Addison-Wesley,

Integration of R environment with the Grid

Marcelina Borcz[1,*], Piotr Bała[1,2]

1. Faculty of Mathematics and Computer Science, Nicolaus Copernicus University
   ul. Chopina12/18, 87-100 Toruń, Poland
2. Interdisciplinary Center for Mathematical and Computational Modeling, Warsaw University
   ul. Pawińskiego 5a 02-106 Warsaw, Poland
* Contact author: marbor@mat.umk.pl

Keywords: Grid, R, UNICORE

R, an environment for statistical calculations, is widely used in many fields by mathematicians, physicians, biologists and others. They usually have to handle huge amount of data. Therefore there is a need for significant computing power, larger then can be provided by a typical laboratory. Possible solution is concept of grid computing introduced in several years ago. From that time much progress in grid technology have been made. UNICORE (Uniform Interface to Computing Resources) is one of grid middlewares that have been successfully used in research and production. It makes distributed resources available in a seamless and secure way. UNICORE 6 provides a graphical client - which allows to load gridbeans – a graphical interface to applications. Gridbeans can be used to build simple jobs or can be treated as building blocks for workflows consisting of different tasks and operations. Here authors introduce gridbean for R enabling an integration of R environment with the grid middleware.

R gridbean is easy to use for both people who are used to work with R and beginners. The user interface contains a panel which makes possible writing commands or open previously saved scripts. There is also a field to input appropriate script arguments. Additional files with data can be upload by browsing and attaching them using dedicated graphical interface. Before a job is executed user can choose an appropriate target system or he can leave this to workflow service. Results are visible in an output panel as text and graphics. Based on the PDF renderer R gridbean enables to view plots and save them to the file as an image All of the features make R gridbean a convenient tool for every R user who need significant computing power which is made available by distributed computing centers. As additional benefit user receives simple environment to handle data and statistical simulations.

References

Borcz, M., Kluszczyński, R. Bała, P. (2007). BLAST Application on the GPE/UnicoreGS Grid. *LNCS*, 4375, 244–252.

Foster, I., Kesselman, C. (1999). The Grid: Blueprint for a New Computing Infrastructure. *Morgan Kaufmann Publishers*.

PDF Renderer website: *https://pdf-renderer.dev.java.net*

UNICORE website: *http://www.unicore.eu*

AnalyzeFMRI: an R package to perform statistical analysis on FMRI datasets

Cecile Bordier[1] , Michel Dojat[1] & Pierre Lafaye de Micheaux[1,2,*]

1. Grenoble Institut des Neurosciences (GIN) - INSERM U836 / Equipe 5
   Neuroimagerie Fonctionnelle et Métabolique
   Université Joseph Fourier - Site Santé BP 170
   38042 Grenoble Cedex 9
   France

2. Laboratoire Jean Kuntzmann / Équipe SAGAG
   Université de Grenoble
   BSHM, 1251 avenue centrale BP 47
   38040 Grenoble Cedex 09
   France

* Contact author: Pierre.Lafaye-de-Micheaux@upmf-grenoble.fr

Keywords: Functional Imaging, Neuroimaging, Brain, Independent Component Analysis (ICA), TclTk

AnalyzeFMRI is a developing package, initiated by J. Marchini, for the processing and analysis of large structural Magnetic Resonance Imaging (MRI) and Functional MRI (FMRI) datasets. In this presentation, we first introduce MRI and fMRI to enlight the data specificities and the main image processing steps. We then describe the current package version and the functionnalities we have recently added, mainly NFTI format management, cross-platform visualization based on Tcl/TK components and temporal and spatial IC analysis. We illustrate our presentation with examples coming from human visual experiments [1,2], especially demonstrating the interest of spatial and temporal IC analysis [3] compared to standard general linear model [4]. We conclude about the interest of the AnalyzeFMRI package for the exploration of MRI data and outline our plans for future extensions.

References   [1] Dojat M, Piettre L, Delon-Martin C, Pachot-Clouard M, Segebarth C and Knoblauch K. Global integration of local color differences in transparency perception: an fMRI study, Visual Neuroscience. 2006;23:357-64.

Warnking J, Dojat M, Guérin-Dugué A, Delon-Martin C, Olympieff S, Richard N, Chéhikian A and Segebarth C. fMRI retinotopic mapping - step by step, Neuroimage. 2002;17:1665-83.

Calhoun VD, Adali T, Pearlson GD and Pekar JJ. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms, Hum Brain Mapp. 2001;13:43-53.

Friston KJ, Holmes AP, Poline JB, Frith CD and Frackowiak RSJ. Statistical Parametric Maps in Functional Imaging: a general linear approach, Human Brain Mapping. 1995;2:189-210.

The impact of Water Supply and Sanitation interventions on child health: evidence from DHS surveys

Ron Bose[1,*] ,

1. International Initiative for Impact Evaluation (3ie)

* Contact author: rbose@3ieimpact.org

Keywords: quantile regressions, matching models, infant mortality, piped water

In this paper I examine the impacts on child health, using diarrhoea as the health outcome, (amongst children living in households) with access to different types of water and sanitation facilities, and from socio-economic and child specific factors. Using cross-sectional health DHS survey data, I employ the propensity score method to match children belonging to different treatment groups, defined by water types and sanitation facilities, with children in a control group. I also employ quantile regression techniques to compare my results and to check for their robustness. Results indicate that disease-specific awareness has strong marginal effects on reducing the predicted probabilities of diarrhoeal outcomes in young children, which are consistent across the models utilised. I also find disease-specific awareness to have the largest impact on reducing the burden of disease from diarrhoea across a select group of predictors

References

Jalan, Jyotsna and Martin Ravallion (2003). Does Piped Water Reduce Diarrhoea for Children in Rural India? *Journal of Econometrics*, 112, 153-173.

Rosenbaum, Paul R and Donald B. Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1), 41-55.

Integrating R into the InfoVis System Visplore

Roland Boubela[1,2,*], Peter Filzmoser[1] and Harald Piringer[2]

1. Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria
2. VRVis research company, Donau-City-Str. 1, A-1220 Vienna, Austria
email: roland.boubela@gmail.com, P.Filzmoser@tuwien.ac.at, hp@vrvis.at

Visplore (`http://www.visplore.at`) is a state-of-the-art InfoVis system which allows interactive visual analysis of large heterogeneous datasets. Various views for numerical and categorical data are available. To enable the system being responsive and interactive even with large data sets it takes advantage of the extensive use of multithreading.

Aim of this project is to integrate R into the InfoVis system *Visplore*. This integration allows for easily synchronizing data between the R workspace and the internal data representation of visplore. The linking-up of the interactive data selection and a representation in R discloses the ability to produce interactive R graphics or to use these selections for modelling. For the user's convenience the R console is implemented as a view and an R package supplies functions to control features of *Visplore* directly. When starting a *Visplore* session one can use the importer for R workspace files to get the data into the system. At any time during the analysis process a snapshot of the current state of the system including loaded data, open views, active selections, etc. can be saved in a session file which includes the actual R workspace.

The tight integration of R into the InfoVis system *Visplore* shows considerable advantages in finding patterns in the data. Visualizing results of modelling approaches helps to find reasonable models and to get a picture of the structure of even massive data sets.

EnQuireR: exploration of questionnaires with R

J. Bouche[1], M. Cadoret[1,*], G. Fournier[1], O. Fournier[1], S. Lê[1], F. Le Poder[1]

1. AGROCAMPUS OUEST, 65 rue de Saint-Brieuc, CS 84215, F-35042 Rennes cedex

* Contact author: marine.cadoret@agrocampus-ouest.fr

Keywords: categorical variables, univariate data analysis, multivariate data analysis, clustering, semantic markup

The use of categorical variables is commonplace in many fields (consumer market studies, politics, health, food science and so on), and in particular in making surveys. As surveys are becoming increasingly popular, there is a growing need for statistical methods including categorical variables.

The main objective of the **EnQuireR** package is to automate the survey process. This package will perform univariate and multivariate data analyses. Those two levels of analysis provide the user a range of functions to improve decision-making aid. Until now, multivariate analysis of categorical variables was performed for instance by the R package **FactoMineR**. Unlike **FactoMineR**, the **EnQuireR** package focuses on one type of applications, *i.e.* the statistical analysis of questionnaires and hence on categorical variables mainly, and allows:

- a faster way to perform the survey process, or any dataset including categorical variables;

- the display of many different outputs including both numerical results and graphs which are precious tools for decision-making aid;

- an easier view of the results by the automatic generation of a *.pdf* report and of a *Beamer* type presentation via the use of **Sweave**.

This package targets a wide range of users from students to scientists and is designed to be accessible to anyone with a basic knowledge of statistics. During the talk we will first present the univariate analysis methods then some methodologies dedicated to multivariate analysis.

The **EnQuireR** package contains the following functions:

- Bar plots: the function **ENbarplot()** can be used to obtain bar plots either sorted by alphabetical order or by bar sizes for each variable with the percentage of missing values. The function **XvsYbarplot()** allows to obtain a bar plot for a variable depending on another variable.

- Distance between variables: the function **chisq.desc()** can be used to measure the statistical relationship between categorical variables. The χ^2 test is used to measure this relationship.

- Multiple Correspondence Analysis (MCA): **missmca()** performs MCA with missing values, **ENlisib()** can be used to improve the graph readability by suppressing the display of objects with a poor quality of representation and **ENMCA()** is used to do cluster analysis following MCA.

- **ENellipse()** draws confidence ellipses around the categories of a variable of interest.

- The function **ENmark()** performs a semantic markup with one, two or three levels.

References

S. Lê, J. Josse and F. Husson (2008). FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software*, 25 (1). pp. 1-18.

F. Leisch (2002). Sweave, Part I: Mixing R and LaTeX: A short introduction to the Sweave file format and corresponding R functions. *R News*, 2 (3). pp. 28-31.

Providing R Reporting Capabilities to a Web Application from a Version
Controlled R Code Database

Dylan Browne1*, Feng Zhu1, Fan Shao1, Richard Pugh1, Oliver Kimberlin1

1.   Mango Solutions, UK
*   Contact author: dbrowne@mango-solutions.com

   Mango were approached to produce a web-based modeling & reporting system for
a pharmaceutical customer. The primary goal of the system would be to generate
graphical, tabular and textual report items in order to evaluate and compare
various pharmacokinetic models. The outputs from the system had to include
reports in various formats containing these report items.

   A key requirements within the application included the ability to store
version-controlled R reporting code which could be readily converted to report
items when associated with data sources. Another requirement was the ability to
link the output reports to a full audit trail, to ensure any item in any output
document can be recreated if requested.

   Mango created a reporting mechanism that could be made available to the
(Java) web application, while allowing users to interact with reporting code
managed in an Oracle database. Communication within the system between the code
storage and the application was achieved using an XML-based messaging system ,
with the result audit trail logging in the database and (optionally) stored
within the output documents.

   This paper discusses the design and use of R in this distributed manner, and
will focus on innovative solutions to the technical challenges this project
presented. This paper will conclude by demonstrating firstly the reporting
capabilities of the resulting system, then showing the same process as a "behind
the scenes" walkthrough.

Abstract submission for:

Benjamin Bryant
PhD Student in Policy Analysis
Pardee RAND Graduate School
The RAND Corporation
Santa Monica, California, USA.
bryant@prgs.edu

Supporting Robust Decisions with Classification and Data-Mining Algorithms

This paper illustrates the concept of scenario discovery and its implementation in a new R package, 'sdtoolkit'. Scenario discovery is a recently developed and useful tool for decisionmaking under uncertainty, in which algorithms are used to identify key sets of plausible future conditions that may lead to unacceptable policy outcomes, in turn allowing decisionmakers the opportunity to create hedging strategies and increase the robustness of their policies. While scenarios have been used for many years to support thinking about the future, scenario discovery is unusual in its quantitative approach to scenario generation, relying on mathematical models of the relevant system to interact policy decisions with exogenous uncertainties and then generate a database of quantitative outcome measures under hundreds to millions of uncertainty combinations. In this context, scenarios are defined as regions of the uncertainty space exhibiting a behavior of interest (typically, poor policy performance). To usefully search through this database and identify relevant scenarios in a form that are still highly interpretable, we have developed a modified version of Friedman and Fisher's Patient Rule Induction Algorithm and implemented it as an R package. Both the algorithm and the package itself are highly interactive, with features for automated data checking, enhanced graphics beyond those of the original PRIM algorithm, and some diagnostic statistics to help the user better judge the confidence they may place in the scenario descriptions.

Local Classification of Discrete Variables by Latent Class Models

Michael Bücker[1]

1. Fakultät Statistik, Technische Universität Dortmund, D-44221 Dortmund, Germany, `buecker@statistik.tu-dortmund.de`

Keywords: Local classification, Latent Class Analysis, Categorical data, Finite mixtures, Naive Bayes

Global classifiers may fail to distinguish classes adequately in discrimination problems with inhomogeneous groups. Instead, local methods that consider latent subclasses can be adopted in this case. Three different models for local discrimination of categorical variables are are implemented in the `lcda` (latent class discriminant analysis) package. They are based on Latent Class Models (cf. [2]), which are discrete finite mixture distributions. Therefore, they can be estimated via the EM algorithm. One model is constructed analogously to the Mixture Discriminant Analysis (cf. [1]) by class conditional Latent Class Models. Two other techniques are based on the idea of Common Components Models (cf. [3]). Applicable model selection criteria and measures for the classification capability are suggested. In a simulation study, discriminative performance of the methods is compared to that of decision trees and the Naive Bayes classifier. It turns out that the MDA-type classifier can be seen as a localization of the Naive Bayes method. Additionally the procedures have been applied to a SNP (single nucleotide polymorphism) data set.

References

[1] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society B*, 58:155-176, 1996.

[2] P.F. Lazarsfeld and N.W. Henry. *Latent structure analysis*. Houghton Miflin, Boston, 1968.

[3] M.K. Titsias and A.C. Likas. Shared kernel models for class conditional density estimation. *IEEE Transactions on Neural Networks*, 12:987-997, 2001.

A Graphical Tool for the Detection of Modes in Continuous Data

Thomas Burger[1*] , **Thierry Dhorne**[1]

1. Université Européenne de Bretagne, Université de Bretagne-Sud, CNRS, Lab-STICC, Centre de Recherche Yves Coppens BP 573, F-56017 Vannes cedex, France

* Contact author: thomas.burger@univ-ubs.fr

Keywords: mode estimation, kernel transform, multi-scale mean shift, dendrogram

In (Bickel, 2003) is presented a robust parametric estimator for the mode of a monomodal continuous distribution. Therefore, it is necessary that the distribution is monomodal. On the other hand, there have been some non-parametric methods for the estimation of the local modes of multimodal distributions. Here, we present a graphical tool that conveniently helps deciding on visual bases, the number of modes of a distribution.

To do so, the distribution is convoluted by a kernel of various scales to let local maxima of the density appear. Conceptually, the approach is similar to time-frequency analysis or wavelet analysis, but in order to best describe the shape of the distribution, Gaussian kernels are used. They are known to be more efficient in computer vision and pattern classification, and the corresponding representation fits the theoretical expectations (Mokhtarian, 1992).

Some other works have explored this connection between pattern classification and descriptive statistics. Hence, a work with ideas similar to ours has already been proposed to publication (Griffin, unpublished), but to our knowledge, in spite of its quality, it remains unpublished. It is based on a multi-scale mean shift algorithm, and the approach is once again rather formal: the point is more to find the various modes, than to provide a convenient way to represent them. Hence, in spite of a common theoretical framework (the similarity with time-frequency analysis in computer vision), the objective is somewhat different.

In addition to this work, we propose a dendrogram-like representation that helps the expert to describe the datasets and/or to propose an adapted mixture model. From an experimental point of view, the method is validated on real and simulated datasets. Finally, an efficient implementation is given.

References

BICKEL, D. (2003). Robust and efficient estimation of the mode of continuous data: The mode as a viable measure of central tendency, Journal of statistical computation and simulation, vol. 73, Issue 12, pp. 899-912.

GRIFFIN, L. D., LILHOLM, M. (unpublished). A Multiscale Mean Shift Algorithm for Mode Estimation. Submitted in 2005 to IEEE Transatction on Pattern Analysis Machine Intelligence.

MOKHTARIAN, F. and MACKWORTH, A. K.(1992). A Theory of Multiscale, Curvature-Based Shape Representation for Planar Curves, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 14, Issue 8, pp. 789-805.

Multiple hurdles models in R: the mhurdle package

Fabrizio Carlevaro[1], Yves Croissant[2,*] , Stéphane Hoareau[3]

1. Département d'économetrie, University of Genève
2. Laboratoire d'économie des transports, Université Lumière Lyon II
3. Universitde la Réunion

* Contact author: yves.croissant@let.ish-lyon.cnrs.fr

In applied econometric studies, the dependent variable often exhibits limited variation, *e.g.*:

- the number of hours of work supplied is non-negative,

- the expenditure in the consumption of a particularly good is non-negative,

In these circumstances, ordinary least squares estimation is biased and inconsistent. However, the model can be estimated consistently using maximum likelihood methods that take into account the censored nature of the dependent variable.

This problem has been treated for a long time in the statistic literature dealing with survival models which are implemented in `R` with the `survival` package. It has also close links with the problem of selection bias, for which some methods are implemented in the `sampleSelection`.

`mhurdle` deals specifically with models where the dependent variable is zero-left censored and may present a large proportion of 0, which is typically the case in household expenditure surveys.

Since the seminal paper of Tobin, a large literature in the econometric field has been developed to deal correctly with this problem of zero observations. More specifically, zero observations may appear for the following three reasons:

- budget constraint: the household would like to consume the good, but his consumer problem has a corner solution because the good is too expensive and/or his income is to low,

- selection: the good is not selected by the household, *i.e.* it's not an argument of its utility function,

- infrequency: the good is bought by the household, but with a low frequency so that zero expenditure may be observed during the survey.

The original *Tobin* takes only the first source of zero into account. With `mhurdle`, the three sources of zero may be introduced in the model.

References

Cragg JG (1971). Some statistical models for limited dependent variables with applications for the demand for durable goods. *Econometrica*, 39(5), pp.829-44.

Deaton A, Irish M (1984). A statistical model for zero expenditures in household budgets. Journal of public economics, 23, pp.59-80.

Tobin J (1958). Estimation of relationships for limited dependent variables. Econometrica, 26, pp.24-36.

Managing chronological objects with **timeDate** and **timeSeries**

Yohan Chalabi[1,2,3] **, Diethelm Würtz**[1,2,3]

1. ITP ETH, Zurich
2. Rmetrics Association, Zurich
3. Finance Online, Zurich

* Contact author: chalabi@phys.ethz.ch

Keywords: calendar management, chronological objects, time series

Time and date management plays an essential role in financial applications with data recorded in different time zones.

In this talk, we describe the concepts and methods behind the S4 classes `timeDate` and `timeSeries` used in the Rmetrics packages for financial data with time and holiday management. In particular, we will explore how timeDate (Würtz and Chalabi, 2009a) allows mixing data collected in different time zones and with different daylight saving time rules (DST). Selected examples will be provided to show its functionality and advantages compared to the other time and date classes. Typical use cases are the management of business holidays or selection of the last working day in a month.

The second part of the presentation will focus on the enhanced implementation of the `timeSeries` class in the package timeSeries (Würtz and Chalabi, 2009b), which is extremely fast. It represents a good example of how to implement an efficient S4 class. Examples will be provided to demonstrate how to use the `timeSeries` class.

Finally, we will also discuss recent and future development directions of the timeDate and timeSeries packages.

References

Würtz, D. and Y. Chalabi (2009a). *The timeDate Package*. cran.r-project.org.

Würtz, D. and Y. Chalabi (2009b). *The timeSeries Package*. cran.r-project.org.

Using R to provide a reporting plug-in
for an Eclipse application

Jonathan Chard1*, Geoff Gibbs1, Andy Dunn1, Fan Shao1, Oliver Kimberlin1,
Richard Pugh1

1.   Mango Solutions, UK
*   Contact author: jchard@mango-solutions.com

   Mango Consultants have produced and delivered many web-based Java
applications using R as the reporting platform. When Mango were asked to design
and implement a desktop modeling & reporting application for a major
pharmaceutical company, an Eclipse RCP platform was selected. In order to
provide reporting capabilities to the application, a flexible plug"R" in was
created.

   This paper discusses the design and use of the plug-in, and will demonstrate
the resulting capabilities of the RCP system. This paper will also discuss how
the design of the R plug-in ensures other Eclipse system can readily take
advantage of R's reporting capabilities.

Fitting Models for the Iowa Gambling Task with R

Chung-Ping Cheng[1,*] , Ching-Fan Sheu[2]

1. Department of Psychology and Research Center of Mind, Brain, and Learning, National Chengchi University, Taiwan
2. Institute of Education, National Cheng Kung University, Taiwan

* Contact author: cpcheng@nccu.edu.tw

Keywords: learning models, Iowa gambling task, mixed-effects

The Iowa gambling task (IGT) is a procedure to diagnose decision-making deficits in people with neurological problems(Bechara, Damasio, Damasio, & Anderson, 1994). Several learning models have been proposed to account for subject's sequential choices during the task. Recently, Ahn, Busemeyer, Wagenmakers & Stout (2008) considered a list of eight such decision-learning models for performance in the IGT. This paper presents a general framework for these IGT models and provides a mixed-effects formulation in fitting models to data using R. In addition to estimating model parameters from individual performance in IGT, our R package can also fit data of individuals from different populations simultaneously by incorporating possible random effects. A number of diagnostic indices are implemented in the routine to facilitate model comparisons and to check model adequecy.

References

Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E. J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, 32, 1376-1402.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, 50, 7-15.

R on Amazon EC2

Karim Chine[1]

1.Cloud Era Ltd, Cambridge, UK

* Contact author: karim.chine@m4x.org

Keywords: biocep, cloud computing, EC2, distributed computing, R workbench, collaborative data analysis

Biocep[1] builds on top of R an open platform for computing and data analysis. The Amazon Elastic Compute Cloud (Amazon EC2) web service provides users with the ability to execute applications in Amazon's computing environment. Using a rich workbench within the browser, the statistician can now work with an R server running on EC2 as if it was local to his machine. The platform hides the complexity of Amazon's cloud computing infrastructure and the R server is abstracted with a simple URL. multiple statisticians can connect simultaneously to the same EC2-R server and analyze data   collaboratively via a set of broadcasted views. For example, the console log is sent in   real time to all users. Chatting is enabled and a graphic device is synchronously updated for all. Biocep includes an editable R-enabled collaborative spreadsheet that retains data on the server, removing limits on client machines. Distributed and linked statistical graphics based on a refactored iplots[2] package enable the collaborative highlighting and color brushing of various linked plots.

Biocep makes distributed computing using R and EC2 accessible to a larger number of statisticians. Easy-to-use functions enable the control from within an R session of several EC2-R workers individually or as a cluster to solve embarrassingly parallel problems. The SOAP-R and RESTful-R Biocep's frontends enable the use of pools of EC2-R workers from Perl,Python, C, C#.. A provided web application uses EC2-R pools to expose an API similar to the Google charts API that returns R Graphics in any format in response to a URL.

Once connected to an R server running at any location, the Biocep's workbench enables data analysis applications wrapped as plugins to access to that server. The workbench can be used as a RESTful Web Service bridge between the R Server and various desktop applications (Excel, OpenOffice,..).

The presentation will include demos of some of the described use cases using publicly available Biocep-R Amazon Machine Instances.

References

[1] Karim Chine(2008), "Biocep, Towards a Federative, Collaborative, User-Centric, Grid-Enabled
     and Cloud- Ready Computational Open Platform,"
      escience,pp.321-322, 2008 Fourth IEEE International Conference on eScience, 2008
      www.biocep.net

[2] Simon Urbanek, Martin Theus (2003). iPLOTS, high interaction graphics for R. Proceedings of the 3rd
      International Workshop on Distributed Statistical Computing (DSC 2003)
      www.iplots.org

Fitting Multidimensional IRT Models with R

Mei-Chen Chu[1*], Ching-Fan Sheu[2]

1. Institute of Cognitive Science, National Cheng Kung University, Taiwan
2. Institute of Education, National Cheng Kung University, Taiwan

* Contact author: u7695105@mail.ncku.edu.tw

Keywords: Multidimensional IRT model, nonlinear mixed model.

Item response theory (IRT) is widely used in assessment and evaluation research to explain how participants respond to questions. IRT assumes that people respond to a test item according to their ability and the difficulty of the item. From the statistical point of view, IRT models are mixed-effects models because the difficulty of the items is a fixed effect, whereas the person ability is considered as a random effect. Currently, several R packages can be used to fit various IRT models. For example, the ltm package (Rizopoulos, 2006) can handle the Rasch model, the latent trait model, the three-parameter model, and the graded response model; and the eRm package (Mair & Hatzinger, 2007) can fit the rating scale model and the partial credit model. These packages, however, can only analyze unidimensional IRT models.

Multidimensional IRT models have been proposed to account for multilevel or hierarchical data, in which subjects may be grouped into clusters and items may be nested within different dimensions. The lme4 package can be used to analyze multidimensional Rasch models for dichotomous outcomes (Doran, Bates, Bliese, & Dowling, 2007). For categorical responses, fitting multidimensional IRT models requires using commercial software such as SAS or Conquest. The paper presents an implementation of multidimensional IRT models for categorical outcomes in R and demonstrates its use with an illustrative example.

References

Briggs, D.C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, 4, 87-100.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rash model with the lme4 package. *Journal of Statistical Software*, 20(2), 1-18.

Mair, P.,& Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9), 1-20.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1-25.

Evaluation of nonlinear mixed effect models using prediction distribution errors: the npde library for R

Emmanuelle Comets [1,2,*], **France Mentré** [1,2,3]

[1] INSERM U738, Paris, France
[2] Université Paris 7, UFR de Médecine, Paris, France
[3] AP-HP, Hôpital Bichat, UF de Biostatistiques, Paris, France

* Contact author: emmanuelle.comets@inserm.fr

Keywords: Pharmacokinetics, Pharmacodynamics, nonlinear mixed effect models, model evaluation

Nonlinear mixed effect models are increasingly used to analyse dose-concentration-effect relationships in clinical studies, to study relationships in clinical studies, to help identify differences in drug safety, efficacy and pharmacokinetics among population subgroups, and to simulate clinical trials. Model evaluation is an important part of model building (EMEA 2006). Prediction discrepancies (pd), have been proposed by Mentré and Escolano (Mentré and Escolano 2006) to evaluate nonlinear mixed effect models. Brendel et al (Brendel et al 2007) developed an improved version of this metric, termed normalised prediction distribution errors (npde), taking into account repeated observations within one subject. In the present paper, we present a set of routines to compute npde.

Model evaluation consists in assessing whether a given model M (composed of a structural model and parameter estimates) adequately predicts a validation dataset V. V can be the original dataset used to build model M (internal validation) or a separate dataset (external validation). The null hypothesis H_0 is that the data in V can be described by model M. The pd for a given observation y_{ij} is defined as the percentile of this observation within the marginal predictive distribution under H_0. Prediction distribution errors (pde) are computed in a similar way after correcting for the correlation induced by repeated observations. Normalised prediction distribution errors are then obtained by transforming the pde through the inverse normal distribution. Under H_0, the distribution of the npde is that of a centered standardised normal distribution. In practice, the predictive distribution is approximated by Monte-Carlo simulations: K datasets are simulated under the null hypothesis (model M and corresponding parameters) using the design of V.

The program requires as input a file with the validation dataset V and a file containing the K simulated datasets stacked one after the other. Simulations should be performed beforehand. The program then computes the npde. Optionally, pd can be computed instead or in addition, which is less time-consuming but leads to type-I error inflation especially as the number of observations per subject increases. Graphical diagnostics are plotted to evaluate model adequacy: QQ-plots and histograms are used to compare the distribution of the npde to that of the theoretical distribution, and npde can also be plotted against predicted concentrations and independent variable to assess trends in the distribution. Tests can be performed to compare the distribution of the npde relative to the expected standard normal distribution. A global test combining a Shapiro-Wilks normality test, a Wilcoxon rank sum test for zero mean, and a Fisher test for a variance of one, with a Bonferroni correction, is reported and can be used to test the adequacy of the distribution of npde compared to the theoretical distribution.

The code is available as a library for the open-source statistical environment R. It can be downloaded from the dedicated website (http://www.npde.biostat.fr/) or from the Comprehensive R Archive Network. The package contains an example of model building followed by model evaluation using npde.

References

Brendel K, Comets E, Laffont C, Laveille C, Mentré F (2006), Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide, *Pharmaceutical Research*, 23, 2036–49.

Committee for Medicinal Products for Human Use, European Medicines Agency (2006).Draft guideline on reporting the results of population pharmacokinetic analyses, EMEA, Brussels, Belgium. `http://www.emea.eu.int/pdfs/human/ewp/18599006en.pdf`

Mentré F, Escolano S (2006). Prediction discrepancies for the evaluation of nonlinear mixed-effects models, *Journal of Pharmacokinetics and Pharmacodynamics*, 33, 345–67.

A new system for collaborative documentation for R

Danese Cooper[1,*]

1. REvolution Computing, Inc.
*   Contact author: danese@revolution-computing.com

R has an excellent set of reference manual pages for its functions, but part of the process of learning how to program in R is finding out the "tips and tricks". Questions like: What's the best way to remove missing values from a data frame?; Which time series class should I use?; or What's the fastest way to build a matrix column by column? aren't readily answered simply by looking at the help page for a function. This information exists, nebulously, in r-help posting, blog articles, and other such venues, but no simple mechanism exists today to contribute such information to the documentation of R itself.

In this talk we review some existing systems to annotate reference documentation with user-contributed content and discuss relative strengths and weaknesses of those systems. We propose a new on-line collaborative system for annotating R documentation, and demonstrate some of its planned capabilities.

ibr: Iterative Bias Reduction Multivariate Smoothing

Pierre-André Cornillon[1] , Nicolas Hengartner[2], Eric Matzner-Løber[3*]

1. Montpellier SupAgro
2. Los Alamos National Laboratory
3. University Rennes 2

* Contact author: eml@uhb.fr

In multivariate nonparametric analysis, sparseness of the covariates also called curse of dimensionality, forces one to use large smoothing parameters. This leads to biased smoother. Instead of focusing on optimally selecting the smoothing parameter, we fix it to some reasonably large value to ensure an over-smoothing of the data. The resulting base smoother has a small variance but a substantial bias. In this paper, we propose a R package named **ibr** to iteratively correct the initial bias of (base) estimator by an estimate of the bias obtained by smoothing the residuals. After a brief description of Iterated Bias Reduction smoothers, we examine the base smoothers implemented in the packages: Nadaraya-Watson kernel smoothers and thin plate splines smoothers. Then, we briefly explain the stopping rules available in the package. Finally we present the package on two examples: a toy example in $\mathbb{R}^2$ and the original Los Angeles ozone dataset.

References

Cornillon, P-A, Hengartner, N. and Matzner-Løber, E (2009) Recursive Bias Estimation for high dimensional regression smoothers, *Submitted*

Multinomial logit models in R

Yves Croissant[1,*]

1. Laboratoire d'économie des transports, Université Lumière Lyon II

* Contact author: yves.croissant@let.ish-lyon.cnrs.fr

The multinomial logit (or conditional logit) is a widely used model in econometrics to explain the choice of an alternative among a set of exclusive alternatives since the seminal works of McFadden. It is based on the hypothesis that the unobservable part of the utility functions are independently and identically distributed with the type 1 extreme value distribution. It is very easy to implement, but suffers serious drawbacks, especially the "Independence of Irrelevant Alternative Hypothesis".

Several extensions of this basic logit model has been developed in the literature.

random parameter logit in this model, the coefficients are assumed to be different among individuals: some hypothesis about the distribution of the coefficients are made and the parameters of these distributions are estimated by simulation,

heteroscedastic logit in this model, the error terms of the utility functions are still independent, but heteroscedastic,

nested logit in this model, there is a hierarchy in the choice, *i.e.* there are different nests.

Currently, a specific form of the multinomial logit model is implemented in R, with individual-specific variables, with the `multinom` function in the `nnet` package. We provide a package called `mlogit` which enables the estimation of the multinomial logit model with both individual and alternative specific variables.

Several packages currently in development that depends on `mlogit` implement some of the extensions of the multinomial logit model:

- The `rlogit` package enables the estimation of the random parameter logit model. A large set of distributions is provided (normal, log-normal, censored-normal, uniform, triangular), the correlation between coefficients may be taken into account and there is support for panel data,

- The `hlogit` the estimation of the heteroscedastic logit model.

.

References

Bhat, C. (1995). A heteroscedastic extreme value model of intercity mode choice, *Transportation Research B* 29, 471483.

Mc Fadden D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka ed., Frontiers in Econometrics, New York: Academic.

Maddala G.S. (1983). Limited dependent and qualitative variables ineconometrics. Econometric Society Monographs, Cambridge University Press.

Train K. (2003). Discrete choice modeling with simulations, Cambridge University Press.

Random Forests

Adele Cutler[1,*]

1. Utah State University
* Contact author: adele.cutler@usu.edu

Keywords: random forests, bootstrap, classification, proximity, trees

Random Forests is becoming a popular algorithm for classification and regression, in part because it does not have many tuning parameters and often works quite well "off the shelf". At the time of his death, Breiman and I were working on a paper describing the ways in which Random Forests can be used to help understand data. In this talk, I will present these ideas and examples as well as some more recent work and future plans. Topics will include variable importance, class weights, proximities and visualization.

What we wish people knew more about when working with R

Peter Dalgaard[1,*]

1. Department of Biostatistics, University of Copenhagen

* Contact author: p.dalgaard@biostat.ku.dk

Keywords: Teaching.

As every established scholar knows, the ignorance of younger people can appear to be absolutely bottomless. On second thought, you usually have to forgive them for not knowing what they were never taught. However, it does mean that we have work to do to bring e.g. PhD students to a level where they can contribute productively to R packages and to do so at a reasonable quality level.

This talk tries to establish a catalogue of items that would be essential in an introductory curriculum in statistical and scientific computing. This could include basic computer science topics, notably the theory of programming languages and object orientation, numerical analysis, and the practical toolchains involved in software development.

Mixed-effects modeling with the `lme4` package: a modern tool for the analysis of plant morphological data in R

Sebastiaan De Smedt[1,*], Katrijn Alaerts[1], Lucien Lemmens[2], Stefan Van Dongen[3], Geert Potters[1], Roeland Samson[1]

1. Department of Bioscience Engineering, University of Antwerp, Belgium
2. Department of Physics, University of Antwerp, Belgium
3. Department of Biology, University of Antwerp, Belgium
* Contact author: Sebastiaan.DeSmedt@ua.ac.be

Mixed-effects modeling is a relatively recently developed technique to describe relationships between a response variable and some covariates in data that are grouped according to one or more classification factors (Pinheiro & Bates, 2000). In traditional statistical analyses, the inside-group variability is removed by averaging measurements over the group levels, or by treating the sampling units of the same groups as independent measurements. Mixed-effects models account for the inside-group dependency by adding random effects to the model structure, and are thus lowering the risk of capitalization on chance (Type I error) (Quené & van den Bergh, 2008). Further, mixed-effects models are useful for the estimation of inside-group variation, which can be interesting in itself. The function `lmer` in the `lme4` package (Bates & Sarkar, 2007), an extended version of the earlier `nlme` package, is widely known as a powerful and flexible tool for the fitting of linear and generalized mixed-effects models in R. The `lme4` package offers fast and reliable algorithms for parameter estimation, as well as tools for evaluating the model (using Markov chain Monte Carlo sampling).

Mixed-effects modeling has strong roots in biomedical and educational research, where researchers need to acknowledge multiple random effects affecting their data: *e.g.* students' school performances are affected by the individual students, their school class, their school, etc. (Quené & van den Bergh, 2008). As such, this technique is frequently used in the aforementioned research domains. In ecological studies, the use of mixed-effects models for data analysis is, up till now, surprisingly scarce. Nevertheless, similar data structures as in social sciences are frequently encountered in ecology, especially in morphological studies: leaf characteristics are for example affected by the tree where they belong to, but also by the trees' provenance or by the trees' surrounding environment.

This poster presents a practical example of the use of mixed-effects modeling in ecological research. In this study, we use the `lme4` package for the fitting of linear as well as generalized linear mixed-effects models to baobab (*Adansonia digitata* L.) leaf morphological data. The leaf morphological parameters are clustered in trees, which are, on their turn, clustered in provenances. The aim of the study is to evaluate the effects of pruning (on tree level), as well as climate (on provenance level), on the aforementioned data. Further, an analysis of the different variance components is made. It is shown that mixed-effects modeling is an appropriate technique for data-analysis in ecological research domains, especially for research on plant or animal morphology.

References

Bates, D.M. & Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes*, R package version 0.99875-6.

Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-effects models in S and S-PLUS*. Springer, New York. 528p.

Quené, H. & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59(4), 413-425.

Extensions of CCA and PLS to unravel relationships between two data sets

Sébastien Déjean[1*]**, Ignacio González**[2]**, Kim-Anh Lê Cao** [2]

1. Institut de Mathématiques de Toulouse, UMR 5219 Université de Toulouse et CNRS
2. Plateforme Biopuces, Genopôle Toulouse Midi-Pyrénées, Institut National des Sciences Appliquées
3. ARC Centre of Excellence in Bioinformatics, Institute for Molecular Bioscience, University of Queensland, Australia

* Contact author: sebastien.dejean@math.univ-toulouse.fr

Keywords: regularization, sparse methods, graphical display, gene expression data.

In the context of systems biology and in post-genomic studies, it becomes usual to analyze simultaneously transcriptomics, proteomics and/or metabolomics data. Several approaches have been proposed to understand and to highlight the mutual interactions between two different data sets. The main challenge of the proposed methodologies relies on their ability to handle very large data sets, where the number of variables is much greater than the number of observations. Lê Cao *et al.* (2008) proposed a sparse PLS method in a canonical correlation framework which includes variables selection while integrating data. González *et al.* (2009) developed a regularized version of Canonical Correlation Analysis to deal with singular matrices occurring in the classical CCA.

On the basis of the CCA package (González *et al.*, 2008), we developped the package `CCAsPLS` to implement these two approaches (Fig. 1).

Figure 1: Schematic view of the CCAsPLS package.

The interpretation of the results is made easier with the use of various graphical display such as correlation loading plots as in factorial analysis (previously implemented in the CCA package). CCAsPLS also proposes heat maps and networks to better visualize the relationships between variables from the two data sets. These graphs are often used when dealing with high-throughput biological data.

References

I. González, S. Déjean, P. Martin, A. Baccini (2008). CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12).

K-A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse (2008). A sparse PLS for variable selection when integrating Omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 35.

I. González, S. Déjean, P.G.P. Martin, O. Gonalves, P. Besse, A. Baccini (2009). Highlighting Relationships Between Heteregeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis. *Journal of Biological Systems*, to appear.

Fitting parametric distributions using R: the `fitdistrplus` package

Marie Laure Delignette-Muller[1,2,3,*] , **Regis Pouillot**[4] , **Jean-Baptiste Denis**[5]

1. University of Lyon, Lyon, France
2. CNRS UMR5558, Villeurbanne, France
3. National veterinary school of Lyon, Marcy l'Etoile, France
4. 4515 Willard Ave., Chevy Chase, MD 20815, U.S.
5. MIA-INRA, Jouy-en-Josas, France.
* Contact author: ml.delignette@vet-lyon.fr

Fitting distributions consists of selecting the best fitting probability distribution from a predefined family of distributions. This practice is specially needed in the domain of Quantitative Risk Assessment. It requires judgment and expertise and generally needs an iterative process of distribution choice, parameters estimation, and quality of fit evaluation. The function `fitdistr` in `MASS` (Venables and Ripley, 2002) is a general-purpose maximum-likelihood fitting routine for the parameter estimation step. Other steps of the process may be developed using `R` (Ricci, 2005) but, to our knowledge, no specific package has been implemented for that purpose.

The package `fitdistrplus` provides several functions to help the fit of a univariate parametric distribution to data. Data may be continuous or discrete, and a specific approach is proposed for each of these two types of data. Continuous data may contain censored values (right-, left- and interval-censored with several upper and lower bounds) as frequently obtained as microbial or chemical analysis outputs used in risk assessment. More precisely, `fitdistrplus` is a set of integrated functions specifically written to:

- Help choose the best parametric distribution that fits a given dataset, using a skewness-kurtosis plot;

- For a given distribution, estimate the parameters using the maximum likelihood method or the method of matching moments and provide goodness-of-fit graphs (empirical and theoretical distributions plot in density and in cdf, P-P plot and Q-Q plot) and statistics (Chi-squared, Kolmogorov-Smirnov and Anderson-Darling statistics) to assess the fit;

- For a fitted distribution, simulate the uncertainty in the estimated parameters by parametric or non-parametric bootstrap resampling. This method may be used in risk assessment for describing an input by a distribution reflecting variability, conditionally to hyperparameters that are considered uncertain.

This package was first built to help the specification of distributions in quantitative risk assessment, but could be used more largely as an help to fit distributions to data, as it provides larger possibilities than the function `fitdistr`. While `fitdistrplus` is already available on the `CRAN`, new graphs for goodness-of-fit for censored data, new goodness-of-fit statistics and tests for non-censored and censored data are currently under development within the R-Forge project "Risk assessment with R"(Pouillot *et al.*, 2008).

References

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth edition. Springer, New-York, U.S.

Ricci, V. (2005). Fitting distributions with R
http://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf

Pouillot, R., Delignette-Muller M.L., Denis, J.-B. (2008). Risk Assessment with R,
http://riskassessment.r-forge.r-project.org/.

ReBaStaBa: handling Bayesian networks with R

Jean-Baptiste Denis[1,*], **Marie-Laure Delignette-Muller**[2,3,4] , **Régis Pouillot**[5]

1. Mathématiques et Informatique Appliquées - INRA - Jouy
2. University of Lyon, Lyon, France
3. CNRS UMR5558, Villeurbanne, France
4. National veterinary school of Lyon, Marcy l'Etoile, France
5. 4515 Willard Ave., Chevy Chase, MD 20815, U.S.

* Contact author: Jean-Baptiste.Denis@Jouy.Inra.Fr

Keywords: Bayesian network, R

Bayesian networks [BN] are an increasing used tool in many applications. There are several strong reasons for such a success: (1) the use of directed acyclic graph to define the structure of a BN is an attractive, efficient and easy way of formalization, (2) it exists nowdays powerful and convenient softwares to apply BN in real applications. Surprisingly enough for a statistician, BN have been mainly promoted by scientists of artificial intelligence and the main softwares dealing with BN in a statistical perspective [Plummer, 2009] do not mention them as such and do not propose specific outputs related to their underlying existence.

For some studies performed in food borne disease assessment or in human physiology, we constructed BN involving categorical and continuous variables (Figure 1). If some were attainable with OpenBugs or Jags, those based in empirical distributions required a direct programing, that we did with R. Soon, rather than doing it specifically, we undertook the writing of a collection of generic R functions, under the name of ReBaStaBa (*REseaux BAyésiens traités par STAtistique BAyésienne*) [Denis, 2009]; it could become one day an R package.

Figure 1: BN produced with rebastaba (Body Composition is a joint work with L. Mioche, Inra).

The aim of rebastaba is not to replace existing tools but (i) to give the possibility of handling (under the R environment) very general BNs, (ii) to provide from them useful outputs about their properties, (iii) to use them in a basic way and (iv) to offer interfaces with other applications (like deal, jags, grappa). Even if rebastaba is still evolving according to the points found when using it, the main concepts are stabilized and the necessary functions to facilitate inputs and to produce outputs are available.

In introduction, a distinction will be made between BN (basically considered as defining a joint probability distribution for a set of random variables) and Bayesian statistics (used to extract by means of the Bayes' theorem information from data). Then based on some cases, the main possibilities of rebastaba will be examplified. Finally some hints will be given on the retained choices and the main different S4 classes introduced to answer the challenge.

References

J.-B. Denis (2009). jbd tools,
    http://w3.jouy.inra.fr/unites/miaj/public/matrisq/jbdenis/outils/welcome.html.

M. Plummer (2009). Jags, Just Another Gibbs Sampler,
    http://www-fis.iarc.fr/~martyn/software/jags/.

Zoo/PhytoImage, a software for automatic analysis of plankton samples based on R and ImageJ

Kevin Denis[1,*], Xabier Irigoien[2], Romain François[3], Véronique Rousseau[4], Jean-Yves Parent[4], Christiane Lancelot[4] and Philippe Grosjean[1]

1. Numerical Ecology of Aquatic Systems, Mons University, 8 avenue du Champ de Mars, 7000 Mons, Belgium
2. AZTI …, Spain
3. Independent R consultant romainfrancois@free.fr, France
4. Ecology of Aquatic Systems, Université Libre de Bruxelles, Bd du Triomphe, 1050 Bruxelles, Belgium
* Contact author: Kevin.Denis@umh.ac.be

Keywords: Plankton, Image analysis, Machine learning, Ecology, Oceanography

Zoo/PhytoImage (ZooImage and PhytoImage, depending on users) is a complete solution to analyze so-called, numerically fixed plankton samples, that is, samples that were digitized usually by imaging system, either in cultures or at sea. Zoo/PhytoImage (http://www.sciviews.org/zooimage) can import and analyze images obtained with digital cameras (micro- or macrophotographies), with flatbed scanners, with the FlowCAM (http://www.fluidimaging.com), or even other digitizing devices like the Zooscan (http://www.zooscan.com), or underwater cameras, provided an adequate importation plugin is available.

Zoo/PhytoImage segments the images, localize particles and measure them (more than 30 measurements on each particle: size, shape, moments, transparencies, texture, etc.). It also extracts 'vignettes', that are little pictures of each original particle. Taxonomists can then classify manually a subset of these vignettes in different taxonomic groups (with hierarchy between these groups if relevant).

This manually classified subset constitutes the training set used to build a classification algorithm for similar samples, using machine learning algorithms. Zoo/PhytoImage can then process a series of plankton samples in batch. It counts, measures and classifies particles found in the samples and calculates ecologically meaningful variables, like relative or absolute abundances, size spectra and biomasses for each taxon.

The software also provides tools to visualize and assess the performances of the classifiers. It proposes a format to store compressed data on disk, and uses dedicated S3 objects in R for it. Metadata, including series, station, cruise, sampling method and information, digitizing technique, … are also handled by Zoo/PhytoImage.

A GUI eases the process from image importation to export of final results for those who are not familiar with R. For the others, Zoo/PhytoImage is more a toolkit of functions that y can assemble in scripts, or in their own R code for complex processing of plankton images. Zoo/PhytoImage has also proven useful in other applications, like bugs counting and classification, or in bacteriology.

Zoo/PhytoImage is partly developed for the AMORE III project (Advanced Modelling and Research on Eutrophication, http://www.ulb.ac.be/assoc/esa/AMORE/objectives.htm) funded by the Belgian Science Policy. It is also a contribution to the SCOR WG130: Automatic visual plankton identification (http://www.scor-wg130.net, Benfield et al. 2007).

References

Benfield, M.C., Ph. Grosjean, Ph. Culverhouse, X. Irigoien, M.E. Sieracki, A. Lopez-Urrutia, H.G. Dam, Q. Hu, C.S. Davis, A. Hansen, C.H. Pilskaln, E. Riseman, H. Schultz, P.E. Utgoff & G. Gorsky (2007). RAPID : Research on Automated Plankton Identification. *Oceanography*, 20(2):12-26.

A Forecasting System Developed under R, Dedicated to Temperature-Controlled Goods Hauling

Wilfried Despagne [1,2,*]

1. European University of Brittany, Lab-STICC (UMR CNRS 3192), University of Southern Brittany, Centre Yves Coppens, Campus de Tohannic, F-56017, Vannes

2. STEF-TFE, Frozen logistic made in Europe, Boulevard Malesherbes, 75008 PARIS

* Contact author: wilfried.despagne@univ-ubs.fr

Keywords: R, Forecasting, Transport, Supply Chain, Statistics, Workflow

Temperature-controlled hauling consists in transporting goods requiring a temperature defined as being between -25 °C and +15 °C. The goods are primarily perishable foodstuffs, meat products, seafood, fruit and vegetables, dairy products, frozen foods, as well as plants or medical supplies. Refrigerated transport is a vital link of the Supply Chain [1], as it is required for the handling of goods from production or manufacturing to distribution and retailing. It is subject to a great many legal constraints (maintaining the cold chain, organizing the working schedule of drivers, obtaining circulating permits, etc.). A haulier who wishes to control his operating costs must, among other constraints, master the management of human and material resources. To gain in productivity, he needs a window on the upcoming flow of goods to be transported between the various stakeholders.

Instead of remaining passive, the haulier TFE [1] has chosen to implement a forecasting system. This system must enable the company to anticipate the quantities of goods to transport and the number of bills of lading [2] to honour. A 21-day forecast of these two key points make it possible to forecast the manpower, dock equipment, and number of trucks to be made available. To achieve homogeneity in all handling and processing operations, the haulier is going for a forecasting system that is adaptable to the specific requirements of its 57 European agencies. The forecasts must be easy to consult, user-friendly, and accessible by a web interface on the intranet [2]. Finally, the company's objectives are to achieve a daily forecasting error margin lower than 5%.

The mathematical forecasting model [3] was thought out in collaboration with the Lab-STICC Laboratory team from the University of Southern Brittany. It was then developed under the R Statistics software. This software was chosen for its open source licence, its speed, its wide variety of functions, its procedure language, and for the fact that it can be executed in batch mode as well as under Windows or Linux. Moreover, after writing the forecasting algorithm, it was extended as a R-library. R gives the posibility to perform an accurate and quicky forcast for satisfying specific different needs of the haulier.

Figure 1: Forecasting System Workflow

The forecasting system was broken down into workflow [figure 1], to enable various applications to work together. Today, without human intervention, real data is updated on a daily basis, and new forecasts are calculated every week, all from a central server.

References

1. Ayadi, S. (2005). Le Supply Chain Management : Vers une optimisation globale des flux (for a global optimization of flows). Working paper, Université Catholique de Lyon.

2. Cluzel, G. (2006). Rentabilité d'un système d'information. Approche théorique (Profitability of an information system. A theoretical approach). *Revue technique de l'ingénieur*, dossier n°AG5310.

3. Despagne, W. (2008). Etude préliminaire à un modèle de prévision à court terme de l'activité d'un transporteur sous température dirigée (A preliminary study of a short-term forecasting model for a temperature-controlled hauling activity). *Modulad*, 39, 95–106.

---

[1]Subsidiary of the STEF-TFE group

[2]Document setting out the transport contract between the transporter and the sender

Using R for the design and analysis of computer experiments with the Nimrod toolkit

Neil Diamond[1,*], David Abramson[2], Tom Peachey[2]

1. Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia
2. Caulfield School of Information Technology, Monash University, Melbourne, Australia
* Contact author: neil.diamond@buseco.monash.edu.au

Keywords: Computer Experiments, Parameter Sweeps, Workflow Engines

The design and analysis of computer experiments to explore the behavior of complex systems is becoming increasingly important in science and engineering (see, for example, Santner et. al., 2003). There is much more to this than merely choosing the design and analysing the resulting data. Computer Scientists at Monash University's eScience and Grid Engineering Laboratory have developed the Nimrod suite of tools (Monash eScience and Grid Engineering Laboratory, 2009) that automates the formulation, running, and collation of the individual experiments and includes a distributed scheduling component that can manage the scheduling of individual components in a local area network. Nimrod contains tools to perform a complete parameter sweep across all possible combinations (Nimrod/G), search using non-linear optimization algorithms (Nimrod/O), or use fractional factorial design techniques (Nimrod/E).

There are a number of workflow engines which provide scientists with an environment with which they can manage data, the workflows of the various analytical steps in their investigation, and summaries of findings. Although existing workflow systems can specify arbitrary parallel programs, they are typically not effective with large and variable parallelism. Similarly, Nimrod was not designed to execute arbitrary workflows. Thus, it is difficult to run sweeps over workflows, and workflows containing sweeps. To overcome these problems, a new tool (Nimrod/K) is being developed, based on the Kepler workflow engine (Kepler Core, 2009). It leverages a number of the techniques developed in the earlier Nimrod tools for distributing tasks to the Grid.

Kepler allows the user to specify R expressions and access R objects as part of the scientific workflow. This talk will describe how existing R packages have been used and extended both to help in the design of the computer experiment, and in the analysis and display of the results.

References

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

Monash eScience and Grid Engineering Laboratory (2009). The Nimrod Toolkit., `http://messagelab.monash.edu.au/Nimrod/`.

Kepler Core (2009). Kepler Project, `http://www.kepler-project.org/`.

C++ classes to extend and embed **R**: The `Rcpp` and `RInside` packages

Dirk Eddelbuettel[1]

1. Debian Project

* Contact author: edd@debian.org

This presentation discusses the `Rcpp` and `RInside` packages that can be used to extend R in high-performance computing settings by minimising the need for data transfer, translation or serialization. `Rcpp` is more generic and can be used to extend R with both custom code, or interfaces to existing libraries. `RInside` offers to take R directly into the user-driven problem domain by embedding it into a given application.

`Rcpp` provides a number of C++ classes that facilitate extending R with compiled code in C or C++. These classes provide a more natural and 'object-oriented' interface than the relatively low-level macros provided by R and documented in the *Writing R Extensions* manual.

We discuss the following classes

`RcppParams` accepts parameters from the calling R function via a named `list` which can contain components of type `double`, `int`, `string`, `bool`, as well as in C++ types for Date and Datetime object from R;

`RcppDate` accepts R Date objects; the class `RcppDateVector` provides a vectorised variant;

`RcppDatetime` accepts R Datetime objects; the class `RcppDatetimeVector` provides a vectorised variant; both operate at a microsecond resolution;

`RcppVector` accepts numeric R vector objects that can be of either type `integer` or `double`; the class `RcppVectorView` provides a lightweight view-only form;

`RcppMatrix` accepts numeric R matrix objects that can be of either type `integer` or `double`; the class `RcppMatrixView` provides a lightweight view-only form;

`RcppStringVector` accepts R vector objects of type character; the `RcppStringVectorView` provides a lightweight view-only form;

`RcppFrame` permits construction of `date.frame` objects at the C++ level; it supports any of the atomic types listed here plus `factor` types for the columns;

`RcppResultSet` permits construction of lists of objects to be returned to R; it can accept all of the types listed here plus a STL vectors and matrices, as well as `SEXP` object common to R.

and illustrate them with examples. We briefly mention more advanced components of `Rcpp` such as function callbacks.

The more recent `RInside` package builds on these classes. It refactors code from the `littler` scripting front-end to R by Horner and Eddelbuettel (2006, 2009) as C++ classes that make it easy to embed R in arbitrary C++ applications. We illustrate the use of these classes with examples.

Why Does **Rmetrics** Need a Documentation Project?

Andrew Ellis[2,3] **, Diethelm Würtz**[1,2]**, Yohan Chalabi**[1,2]**, Martin Hanf**[3]

1. ITP ETH, Zurich
2. Rmetrics Association, Zurich
3. Finance Online, Zurich

* Contact author: ellis@itp.phys.ethz.ch

Keywords: documentation, Sweave, portfolio optimization, finance, econometrics

A common problem in open source software is documentation. Providing good documentation is a very time-consuming process, and it is often the case that software developers would rather spend their time writing code than documenting it. A further problem is that it requires a great effort to keep the documentation up-to-date, especially in a rapidly-changing software environment, since writing code and documentation seldom occur in parallel.

Of course, there are many excellent text books available, which provide documentation of R itself, or of individual R packages. A good example is the use R! series published by Springer.

In this presentation, we want to present an alternative business model for publishing documentation of open-source software. Our book (Würtz, Chalabi, Chen, and Ellis, 2009) is published on the Rmetrics website[1] as an eBook. This will allow us to quickly adapt the book to changes in the software.

We will share our experience in creating a large documentation project which aims to provide the same quality as that which can be found in certain commercial software. Furthermore, we will discuss technical aspects of writing, typesetting and publishing a book using Sweave (Leisch, 2002), LaTeX and version control. A further point we will discuss is the infrastructure which makes self-publishing possible: readily-available eCommerce software, marketing tools and on-demand publishing.

Finally, we will address the question of whether writing and publishing good documentation can improve the adoption rate of software in the finance industry, and whether this business model allows us to support the development of our open-source software.

References

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Härdle and B. Rönz (Eds.), *Compstat 2002 - Proceedings in Computational Statistics*, Heidelberg, pp. 575–580. Physica Verlag.

Würtz, D., Y. Chalabi, W. Chen, and A. Ellis (2009). *Portfolio Optimization with R/Rmetrics* (1 ed.). Zurich: Rmetrics Association & Finance Online.

[1]http://www.rmetrics.org

bigmemory: bigger, better, and platform-independent

John Emerson[1,2,*] , Michael Kane[1]

1. Department of Statistics, Yale University
*   Contact author: john.emerson@yale.edu

Keywords: high-performance computing, shared memory

The newly re-engineered package `bigmemory` uses the Boost Interprocess C++ library to provide platform-independent support for massive matrices. These matrices may be allocated to shared memory with transparent read and write locking. In addition, `bigmemory` now supports file-backed matrices, ideal for applications exceeding available RAM. Proof-of-concept applications to be discussed may include parallel algorithms, biomedical imaging, and data streaming with R.

References

Boost Interprocess Library,
   http://www.boost.org/.

R Development Core Team (2008). *R: A language and environment for statistical computing.* R foundation for Statistical Computing, Vienna Austria. ISBM 3-900051-07-0, http://www.r-project.org/.

Proximity data visualization with h-plots

Irene Epifanio[1,*]

1. Departament de Matemàtiques, Universitat Jaume I, Castelló 12071, Spain

* Contact author: epifanio@uji.es

Keywords: $h - plot$, multidimensional scaling, dissimilarity matrix, dimension reduction, human corneal endothelia

Classical multidimensional scaling methods try to preserve all pairwise proximities, whereas many of the recent nonlinear dimension reduction methods, such as Tenenbaum et al. (2000) or Roweis and Saul (2000), use only local neighborhood information to construct a global low-dimensional embedding of a hypothetical manifold near which the data fall (Hastie et al. 2009). Both approaches could become into restrictive constraints in some cases, specially if the measure between objects is not a distance.

Our motivating problem is concerned with the analysis of digital images of human corneal endothelia. In Ayala et al. (2006), different dissimilarities (non-metric measures) between these images were proposed and assessed in a simulation study and, finally, applied to the ophthalmologic problem. Note that triangle inequality is not hold by the dissimilarity considered. In order to compute these dissimilarities, the following libraries of R have been used: *Splancs*; *Spatstat* and *Survival*.

We propose a method based on $h - plot$ (Seber, 1984) for graphical exploration of dissimilarity matrices, which leads to different representations from other methods. It is a non-iterative method, very simple to implement and computationally efficient. The representation goodness can also be easily assessed. It can also be applied to asymmetric proximity data, since our methodology can handle naturally this kind of situation. It has been compared with well known methods and shown its good behavior through several examples, specially with nonmetric dissimilarities. We also propose two alternatives, depending on if the only objective is graphical representation or if cluster and pattern detection is also the goal, using the original dissimilarities or their ranks, respectively.

This work has been done by using free software, R, and specially the library *MASS* (Venables and Ripley, 2002).

An example with more illustrative results on an artificial dataset, the Swiss roll dataset, is available at the following web page: http://www3.uji.es/~epifanio/RESEARCH/hplot.pdf.

References

J. B. Tenenbaum, V. de Silva, and J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (5500), 2319–2323.

S. T. Roweis and L. K. Saul (2000). Linear embedding nonlinear dimensionality reduction by locally. *Science*, 290 (5500), 2323–2326.

T. Hastie, R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning. Data mining, inference and prediction*. Second Edition. Springer-Verlag.

G. Ayala, I. Epifanio, A. Simó, and V. Zapater (2006). Clustering of spatial point patterns. *Computational Statistics & Data Analysis*, 50(4), 1016–1032.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. ISBN 3-900051-07-0.

B. Rowlingson, P. Diggle, adapted, packaged for R by R. Bivand, pcp functions by G. Petris, and goodness of fit by S. Eglen. *splancs: Spatial and Space-Time Point Pattern Analysis*.

A. Baddeley, and R. Turner (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6), 1–42.

T. Therneau, and ported for R by T. Lumley. *survival: Survival analysis, including penalised likelihood*.

G. Seber, (1984). *Multivariate observations*. John Wiley.

W. Venables, and B. Ripley (2002). *Modern applied statistics with S-plus*. Springer.

Coalition : a simple and useful tool to distribute R-works on a set of computers

Marie-Pierre Etienne[1,2,*], **Cyril Corvazier,**[3] , **Benjamin Legros**[3]

1. AgroParisTech, UMR 518, Mathématiques et Informatique Appliquées, F-75005 Paris, France
2. INRA, UMR 518, Mathématiques et Informatique Appliquées, F-75005 Paris, France
3. Mercenaries Engineering, Paris, France

* Contact author: marie.etienne@agroparistech.fr

Keywords: Distributed Computing, Resource Manager,

Intensive simulations, Monte Carlo methods or hierarchical Bayesian estimation can be highly CPU intensive and, in order to be solved in a decent amount of time, have to be run on a set of computers. More and more computations are distributed on dedicated servers, the difficulty is then to manage the computation tasks. Most of the available tools for distributed computing with R are R specific (see for instance SNOW package, Rossini el al, 2009) and/or require strong network programming skills.

We propose a simple solution called Coalition to distribute shell commands on a set of computers. From the user point of view, it is as simple as using R (or whatever he wants) in command line and Coalition manages the computational resources very similarly to SLURM (S. M. Balle and D. Palermo, 2007).

Coalition consists of two main Python scripts : `server.py` and `worker.py`. `server.py` is run on the master computer, each computer of the set (worker) runs the `worker.py` script.

A user adds a command on the server either using a web interface or a command line through the `control.py` script. The so called 'job' is added to the server job queue. When a worker is free, it asks the server for a job. As an answer, the server attributes the next job present in the job queue to the worker. If the command is successfully achieved (exit code 0), the task is marked as finished in the job queue. To use R in this framework, one just has to submit some `Rscript` commands to the server.

A job consists of a command line but also of a priority level, some affinity flags and dependencies on other jobs. The job queue is sorted according to the priority levels of the submitted jobs. The affinity flags are used to constraint a job to be run on a subset of workers. For instance, a R job may require a particular package not available on all the workers. In this case the affinity flag may indicate this requirement and only workers with the suitable installation of R can execute the job command. The dependencies of a job, say A, consists of a set of jobs that must be finished before running A. A friendly web interface is available to submit jobs, manage the job list, watch and control the state of the workers.

Because a job command is executed on a worker, all the data required for this command must be available from any worker. This availability is straightforward with a network file-system and may be achieved through for instance FTP commands in other cases. If properly configured, the server can use a LDAP directory to authenticate the users. In this case, the workers are able to run the jobs' commands with the submitter's rights. This is a simple and safe way for the workers to access the users' files over the network.

To take advantage of multi-core CPU, one simply has to run on a single computer as many workers as cores.

A major limitation of this method leads in the necessity to split up the computing task in several independent jobs. This may be an issue for some application. But, with a suitable task splitting, this framework can achieve nearly optimal parallel performances. It has been widely tested in simulation studies context with R.

Coalition is distributed under GNU GPL license and available at `http://coalition.googlecode.com`.

References

Luke Tierney, A. J. Rossini, Na Li, H. Sevcikova (2009) SNOW Package : Simple Network of Workstations `http://cran.r-project.org/web/packages/snow/index.html`

S. M. Balle and D. Palermo (2007) Enhancing an Open Source Resource Manager with Multi-Core/Multi-threaded Support, *Job Scheduling Strategies for Parallel Processing.*

Influence Diagrams on R

J.A. Fernández del Pozo[1,*] **, C. Bielza**[1]

1. [1]Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Madrid, Spain

* Contact author: jafernandez@fi.upm.es

Keywords: Machine Learning, Probabilistic Graphical Models, Bayesian Networks, Classification, Decision Making.

We show, on the intersection of machine learning and decision analysis, a R package for development of probabilistic graphical models like influence diagrams and bayesian networks, two popular knowledge representation tools to deal with decision making and classification problems. *IdR* is a package we have developed to support some issues about tasks related to the influence diagrams building, inference,....

We represent the decision model using a R scritp which is loaded into the environment as a variable-node list. We have implemented several evaluation algorithms from literature, descriptive tools and query procedures to run the decision model. Also we export the model to other environments. First, we can export to the *GeNie* (http://genie.sis.pitt.edu/, ©Decision Systems Laboratory, 2005-2007. All Rights Reserved) application the models to use its graphical interface among others features. Second, the results we will obtain can be retrieved on the *KBM2L* (http://www.dia.fi.upm.es/ jafernan/research/kbm32class.zip in java) system to perform explanations about the reasoning and sensitivity analysis. The idea is to concentrate on the core tasks (problem representation, model building, variable description and inference) and to develop in the future packages for graphical interface, explanation and other issues.

We suggest this package as educational resource, because it is easy to define the model and all its components and to perform analysis breakdown for the evaluation process. The evaluation output is the operations sequence performed over the model, and optionally the optimal decision tables and the posteriory conditional probability tables. Generally, we need to evaluate the model and to analyse the results (correctness, sensitivity, explanation,...). We point that the *IdR* package is useful for sensitivity analysis and simulation because the representation is clear, open and it is part of a powerful statistical environment.

An overview of the *IdR* package is summarizes on: we can define models with discrete domain variables like regular influence diagrams (with one utility node) and bayesian networks; *IdR* uses the *lattice* and *cluster* packages to analyse the conditional probability tables; we can summarize the graph, the strength of probability relationships and the conditional independences between nodes; the evaluation of influence diagrams and bayesian networks can be perform numerical (exact or rough) and qualitative way; also the evaluation can be full or partial, by mean of instances on any subset of variables; *IdR* uses the vectorization on its implementation of the Bayes rule and expected utility maximization, two complex operations involved on probabilistic inference.

Finally, we will try to implement more general decision networks (continuous variables, several utility nodes, non sequence decision nodes,...), alternatives to the conditional probability tables (linear models) and utility tables (multiattibute utility functions), evaluation algorithms, and learning algorithms from data for the bayesian networks. We are interesting on paralell evaluation of huge models, using packages like *snow*, i.e. very large decision sequences. We need also high computational power when try to perform global sensitivity analysis parametrized with tens of continuous variables.

References

Fernández del Pozo, J.A. (2008). Influence diagrams on R,
`http://www.dia.upm.es/~jafernan/research/IdR.html`.

Fernández del Pozo, J.A. and Bielza, C. and Gómez, M. (2005). A List-Based Compact Representation for Large Decision Tables Management. European Journal of Operational Research, 160, Special Issue on Decision Making and AI, 638–662.

Bielza, C., Fernández del Pozo, J.A., Lucas, P. (2008) Explaining clinical decisions by extracting regularity patterns, Decision Support Systems, 44, 397408.

Analysis of Economic Data With Multiscale Spatio-temporal Models

Marco A. R. Ferreira[1,*] , Adelmo I. Bertolde[2], Scott H. Holan[1]

1. University of Missouri - Columbia
2. Universidade Federal do Espírito Santo
* Contact author: ferreiram@missouri.edu

Keywords: Areal data, Dynamic linear model, MCMC, Multiscale modeling, Spatio-temporal.

We develop a new class of multiscale spatio-temporal models for Gaussian areal data. Our framework decomposes the spatio-temporal observations and underlying process into several scales of resolution. Under this decomposition the model evolves the multiscale coefficients through time with structural state-space equations. The multiscale decomposition considered here, which includes wavelet decompositions as particular case, is able to accommodate irregular grids and heteroscedastic errors. The multiscale spatio-temporal framework we develop has several salient attributes. First, the multiscale decomposition leads to an extremely efficient divide-and-conquer estimation algorithm. Second, the multiscale coefficients have an interpretation of their own; thus, the multiscale spatio-temporal framework may offer new insight on understudied multiscale aspects of spatio-temporal observations. Finally, deterministic relationships between different resolution levels are automatically respected for both the observations, the latent process, and the estimated latent process. We illustrate the use of our multiscale framework with two examples. First, we analyze a simulated dataset of functional data with temporally evolving functions. Finally, we analyze a spatio-temporal dataset on agriculture production in the state of Espirito Santo, Brazil.

References

Marco A. R. Ferreira, Adelmo I. Bertolde and Scott Holan (2009), Analysis of economic data with multiscale spatio-temporal models, Handbook of Applied Bayesian Analysis, Editors: O'Hagan and West, Oxford University Press, to appear.

Novel method for estimating isotope incorporation into peptides using the half-decimal place rule

Ingo Fetzer[1,*] , Nico Jehmlich[2], Frank Schmidt[3]

1. Department of Environmental Microbiology, Helmholtz Center for Environmental Research - UMB, 04318 Leipzig, Germany
2. Department of Proteomics, Helmholtz Center for Environmental Research - UMB, 04318 Leipzig, Germany
3. Interfaculty Institute for Genetics and Functional Genomics, Ernst-Moritz-Arndt-University, 17487 Greifswald, Germany
* Contact author: ingo.fetzer@ufz.de

The metabolic incorporation of stable isotopes such as ^{13}C or ^{15}N into proteins has become a powerful component to disentangle substrate fluxes in bacteria and understand microbial degradation processes. Incorporation of heavy isotopes into proteins can be used to analyze process parameters such as protein turnover rates. Here we present a new method for calculating the incorporation rate of ^{13}C into peptides by using the information given in the decimal places of peptides by making full use of the information acquired by high resolution mass spectrometry. Our method for estimating ^{13}C incorporations is based on the characteristics of the so called 'half decimal place rule' (Mann 1995, Schmidt 2003). The rule follows the observation that when molecular masses of (isotopic unlabelled) peptide sequences are plotted against their digital residuals (=numbers behind the decimal point) the resulting plot will follow a linear pattern with a characteristic slope. Gradual incorporation of heavy isotopes into peptides will produce the same linear structure but the regression slope will be also altered. Thus, the steepness of the slope contains information on the relative incorporation of heavy isotopes into the proteins. Our method and all calculations have been implemented in 'R' in form of three separate scripts: Primarily molecular weights as references for unlabelled and labeled ^{13}C peptides had to be calculated. For this an existing peptide database of *Mycobacterium tuberculosis* (initially containing >900.000 protein sequences with variable lengths of 2-40 amino acids) was used. In a first script, primarily the dataset was reduced to the relevant 90,637 amino acid sequences taking advantage of R's excellent data mining abilities. In a following step the molecular weights all sequences were determined by multiplying every amino acid from each sequence with its corresponding molecular weight. In a second step, after classification of the peptides by applying k-means clustering (kmeans() in stats package) we were able to plot molecular peptide weights against their digital residuals. With help of linear curve fitting (using the lm() application of the 'stats' package) the reference slopes for unlabelled and 13C-labeled amino acid sequences could be estimated. Including our reference slopes in a third script, a user friendly approach was developed using the tcl/tk package, providing windows to guide the user easily through the procedure calculating the incorporation rates of his own data series. Since user data often contain far less measurements than we used as reference and, moreover, often contain 'outliers' standard linear curve fitting was not applicable. Here the robust linear model procedure (rlm() of the MASS package) gave more reliable results with better isotope incorporation rate estimates. With help of the previously determined slopes of fully labeled/unlabeled data the relative incorporation rates of isotopes for user data could be computed. Additionally, the applicability of this approach is demonstrated by using *Pseudomonas putida* ML2 proteins labeled uniformly via the consumption of ^{13}C$_6$-benzene. Based on several labeled peptides the incorporation of ^{13}C was calculated. As a result, the accuracy of the calculated incorporation rate depends on the number of used peptide masses, whereas only 100 peptide masses are required to get precision higher than 5 atom % of ^{13}C incorporation.

References

Mann, M. (1995). 43rd ASMS Conference on Mass Spectrometry and Applied Topics, Atlanta, GA.

Schmidt, F., Schmid, M., Mattow, J., Facius, A., Pleissner, K. P. and Jungblut, P. R. (2003) Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis. *J Am Soc Mass Spectrom.* 14: 943-956.

R to LaTeX, Univariate Analysis

Lionel Riou França[1], Christophe Genolini[1,*]

1. INSERM U669, Paris Sud Innovation Group in Adolescent Mental Health Methodology, Paris, France

* Contact author: genolini@u-paris10.fr

Abstract

The package **r2lUniv** (R to LaTeX, Univariate analysis) provides facilities to export some R analysis in a LaTeX format. Working either on a single variable or on a data.frame, its main function **rtlu** computes several statistics and graphical output for each variable, according to their nature. Four kinds of variable are considered:

- For **nominale** variables (factor, logical and characters), rtlu display frequencies and barplot.
- For **ordinal** variables (ordered), rtlu computes frequencies, quartiles and barplot.
- Numerical variables (numeric and integer) are split in two categories. Numerical variables with only few modalities are considered as **discrete**. For discrete variables, rtlu performs frequencies, mean, standard deviation, quartiles, boxplot and barplot.
- Other numerical are **continuous**. For continuous variables, rtlu calculates mean, standard deviation, quartile, boxplot and histogram.

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

Keywords: Interface to other languages, Univariate analysis, LaTeX

References

[1] Christophe M. Genolini, *EPO2007: Exam cheating at French University*, http://christophe.genolini.free.fr/EPO/EPO2007-Fraude.php.

Advanced editor for the biocep workbench

Romain François[1] , **Karim Chine**[2]

1. Independent R consultant, France. `francoisromain@free.fr`
2. Cloud Era Ltd, United Kingdom. `karim.chine@gmail.com`

* Contact author: `francoisromain@free.fr`

Keywords: : biocep workbench, jedit, user interface, text editor

The biocep project[1] is a general unified solution for integrating and virtualizing the access to R engines/servers and aims to become the federative user friendly computational e-platform for research, finance and education. The virtual workbench is part of the biocep project and brings a flexible user interface that can easily extended by the use of plugins.

The *advanced editor for the biocep workbench* is one of these plugins, providing a very powerful and customizable text editor — based on jedit[2] — directly integrated to the virtual R workbench. The plugin embeds jedit and its dockable windows as views of the workbench, and embeds a set of jedit enhancements specific to the use of R. The current feature set includes syntax highlighting, informative code completion, an object explorer, minimal support for Roxygen, detection of potential errors in R code, a code analysis tool, ...

The presentation will give a snapshot of the current features of the plugin, as well as an overview of the next phases of its development.

References

[1] Karim Chine. Biocep, Towards a Federative, Collaborative, User-Centric, Grid-Enabled and Cloud-Ready Computational Open Platform, escience,pp.321-322, 2008 Fourth IEEE International Conference on eScience, 2008

[2] Jedit developer team. JEdit, Programmer's text editor, 2009. http://www.jedit.org/

Fast Sparse Regression and Classification

Jerome H. Friedman

Stanford University

jhf@stanford.edu

Linear models are among the most used for statistical prediction. The accuracy of such models suffers when the number of observations is not large compared to the number of predictors. Regularization methods have been proposed to mitigate this problem. These methods fit a linear model to data,based on some loss criterion, subject to a constraint on the coefficient values. For large problems the general choice of loss/constraint combinations is limited by the computation required to obtain the corresponding solution estimates. This is especially the case when non convex constraints are used to induce very sparse solutions. A fast algorithm is presented that produces solutions that closely approximate those for any convex loss, and a wide variety of convex and non convex constraints, permitting application to very large problems. The benefits of this generality are illustrated by examples.

Combining Text Mining and Microarray Analysis

Christoph M. Friedrich[1,*] and Michaela Gündel[1]

1. Fraunhofer Institute for Algorithms and Scientific Computing (SCAI); Department of Bioinformatics; Schloss Birlinghoven; D-53754 Sankt Augustin; Germany
* Contact author: friedrich@scai.fraunhofer.de

Keywords: Text mining, Microarray workflow, Intracranial Aneurysms

Microarrays are well established experimental tools to measure gene expression in biological samples. The Bioconductor project [1] provides a wealth of R packages to analyse the resulting gene expression data. In [2] a semi-automatic microarray analysis workflow with limma [3] as the main analysis engine has been developed. This workflow is part of @neuLink [4], a biomedical Knowledge Discovery application suite. Another module of this application suite allows for Knowledge Discovery in biomedical text sources, namely Medline. Combining prior published knowledge with experimental data allows for example the identification of co-mentioned diseases or drugs in similar published gene expression data.

The microarray analysis workflow will be presented in detail as well as "lessons learnt" during the development and use. Additionally it will be shown how text mining is combined with microarray analysis in this Knowledge Environment [5].

Acknowledgements

References

1. Gentleman, R.; Carey, V.; Bates, D.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. and Zhang, J. Bioconductor: open software development for computational biology and bioinformatics *Genome Biology,* **2004***, 5*, R80
   http://www.bioconductor.org; last accessed 2009-02-26
2. Gündel, M. *ArrayProcess: Work Flow for Microarrays*; Masters thesis, Life Science Informatics at Bonn-Aachen International Center for Information Technology (B-IT); *Germany,* **2007**
3. Smyth, G. K. *Limma: linear models for microarray data.* In Gentleman, R. et al. *(ed.) Bioinformatics and Computational Biology Solutions using R and Bioconductor, Springer,* **2005**, 397-420
4. Friedrich, C. M.; Dach, H.; Gattermayer, T.; Engelbrecht, G.; Benkner, S., and Hofmann-Apitius, M. *@neuLink: A Service-oriented Application for Biomedical Knowledge Discovery* Proceedings of the HealthGrid 2008, IOS Press*,* **2008**, 165-172
5. Hofmann-Apitius, M.; Fluck, J.; Furlong, L. I.; Fornes, O.; Kolarik, C.; Hanser, S.; Boeker, M.; Schulz, S.; Sanz, F.; Klinger, R.; Mevissen, H.-T.; Gattermayer, T.; Oliva, B. and Friedrich, C. M. *Knowledge Environments Representing Molecular Entities for the Virtual Physiological Human* Philosophical Transactions of the Royal Society A, **2008***, 366(1878)*, 3091-3110

The investigation a frequency of asthma in ECAP study in Poland

Marta Zalewska[1,*] , Konrad Furmańczyk[1,2,]**

1. Department of the Prevention of Environmental Hazards and Allergology, Medical University of Warsaw, Poland
2. Department of Applied Mathematics, Warsaw University of Life Sciences, Poland
*   Contact author: zalewska.marta@gmail.com
** Contact author: konfur@wp.pl

Keywords: Biostatistics, Correspondence Analysis, ECAP, FactoMineR

We investigate, using the correspondence analysis, the association between asthma and region in Poland according to ECAP study (Epidemiology of Allergy in Poland). The method of study was a questionnaire-based survey on ISAAC and ECRHS II .

ISAAC and ECRHS II was conducted in a total of 20449 subjects. 18617 were selected to the analysis: 50.4% adults aged 20-44 years (A), 24.2% children   6 –7 years (Ch) and 25.4% children aged 13-14 years (Te), 9998 female and 8591 male, in Poland in the years 2006 - 2008. 25,7% (n=4783) of them passed through the clinical examination, including skin prick tests and spirometry (Ch - 1329 subjects, Te - 1321 subjects, A- 2133 subjects). All study subjects were randomly selected (n=97500) from PESEL data base (identity number given to each citizen of Poland) in 8 cities and 1 rural regions (response rate 41,9%). Data acquisition was done by the Computer Assisted Personal Interviewing with GSM transmission to update the main database at the Warsaw Medical University.

All our analyses were performed in the R package FactoMineR.

References

Greenacre M (1984). Theory and Applications of Correspondence Analysis. *London: Academic Press*.

Husson F, Josse J, Le S, Mazet J (2008) The FactoMineR Package, version 1.10,
http://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf

Samoliński B (2008). Epidemiology of Allergic Diseases in Poland, Report of ECRHS II and ISAAC study*,
        Department of Environmental Hazards Prevention and Allergology, Medical University of Warsaw*

Using R for regression model selection with adaptive penalties procedures based on the False Discovery Rate (FDR) criteria

Tal Galili**, Yulia Gavrilov and Yoav Benjamini
Department of Statistics and Operations Research · Tel Aviv University
***galilita@tau.ac.il**

Keywords: model selection, Linear regression models, Forward model Selection, FDR adaptive penalties, Multiple testing

Background: Multiple hypothesis testing has become an integral component in numerous modern age statistical challenges, ranging from microarray genomic analyses to fMRI brain scans. The control the False Discovery Rate (FDR) has proven to be an effective tool in addressing the multiple testing in large problems. Similarly, recent research has introduced the FDR approach to the problem of model selection, when the number of potential variables is very large.

In our presentation we discuss an implementation in R of a set of methods applying forward variable selection to linear regression models, based on penalized FDR controlling procedures such as the procedure in Benjamini-Hochberg(1995) [**] and the adaptive one in Benjamini and Gavrilov (2009) [**].

We first review the theoretical and empirical merits of the proposed FDR based procedures.

Theoretically [*] - It is asymptotically minimax for ℓ^r loss simultaneously throughout a range of sparsity classes for an orthogonal design matrix. Empirically [***] - it showed good performance over other penalized methods in a recent comprehensive simulation study. The study was conducted with a wide range of realistic settings, including non-orthogonal explanatory variables and it compared the methods using empirical minimaxity relative to a 'random oracle' (the oracle model selection performance on data dependent forward selected family of potential models.)

We proceed with results from our ongoing research on implementing these methods to general linear model, emphasizing their implementing within other R packages (such as MASS, leaps, biglm, ff and more) with the possibility of creating a dedicated package for their use.

References
 [*] Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2000) 'Adapting to Unknown Sparsity by Controlling the False Discovery Rate', Technical Report No. 2000-19, May 2000.
[**] Benjamini, Y., Hochberg, Y., (1995) " Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing ", Journal of the Royal Statistical Society B, 57 289-300.
[***] Benjamini, Y., Gavrilov, Y., (2009) 'A SIMPLE FORWARD SELECTION PROCEDURE BASED ON FALSE DISCOVERY RATE CONTROL', Annals of Applied Statistics (in press).

Sequential Implementation of Monte Carlo Tests with Uniformly Bounded Resampling Risk

Axel Gandy[1,*]

1. Department of Mathematics, Imperial College London * Contact author: a.gandy@imperial.ac.uk

Keywords: Monte Carlo testing; p-value; R-package; Sequential test.

This talk describes an open-ended sequential algorithm for computing the p-value of a test using Monte Carlo simulation, e.g. a bootstrap test. The algorithm guarantees that the resampling risk, the probability of a different decision than the one based on the theoretical p-value, is uniformly bounded by an arbitrarily small constant. Previously suggested sequential or non-sequential algorithms, using a bounded sample size, do not have this property. Although the algorithm is open-ended, the expected number of steps is finite, except when the p-value is on the threshold between rejecting and not rejecting. The algorithm is suitable as standard for implementing tests that require (re-)sampling. It can also be used in other situations: to check whether a test is conservative, iteratively to implement double bootstrap tests, and to determine the sample size required for a certain power.

An R-package implementing the algorithm will be discussed.

References

Gandy, Axel (2008). Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. Preprint.
http://arxiv.org/abs/math.ST/0612488.

KmL: K-means for Clustering Longitudinal Data

Christophe Genolini[1,*], **Bruno Falissard**[1]

1. INSERM U669, Paris Sud Innovation Group in Adolescent Mental Health Methodology, Paris, France

* Contact author : genolini@u-paris10.fr

Abstract

The package **KmL**[1] is a generalization of the K-means algorithm for clustering Longitudinal data.

Cohort studies are becoming essential tools in epidemiological research. In these studies, measurements are no longer restricted to a single variable but can be seen as trajectories. K-means is one of the statistical methods that can be used to determine homogeneous groups of patients trajectories.

KmL is a new implementation of k-means design to work specifically with longitudinal data. It provides some facilities to deal with missing values; it also rerolls the algorithm several times, varying the starting conditions and/or the number of clusters looked for; to finish with, its graphical user interface makes this tool well suited to choose the appropriate number of clusters, when the classical critera are not efficient.

Keywords: trajectories, longitudinal data, k-means, cluster analysis, non-parametric algorithm

References

[1] Christophe Genolini, Bruno Falissard. KmL: A Non-Parametric Algorithm for Clustering Longitudinal Data, 2008
http://christophe.genolini.free.fr/kml/.

RNONMEM2: An R bundle for easy manipulation and graphing of NONMEM data.

Francisco Gochez[1,*]

1. Mango Business Solutions

* Contact author: fgochez@mango-solutions.com

Keywords: Pharmacokinetics, Reporting

NONMEM is widely regarded as the "gold standard" of nonlinear-mixed effects modelling software. However, its output is best suited for the age of paper printouts and it is notoriously difficult to use in other software. RNONMEM2, a refactoring of an earlier package produced by Mango Solutions, is a tool which aims to load and parse the output of NONMEM runs (including parameter estimates, diagnostic information, and data sets) into the R environment in a consistent, object oriented manner which is easy to use. Users can thus combine NONMEM's powerful model fitting routines with R's flexible data analysis capabilties. In addition, it provides tools for generating graphics from NONMEM runs, and also for producing standard reports and generating control files.

Risk Theory Calculations with R and actuar

Vincent Goulet[*]

* École d'actuariat, Université Laval, Québec, Canada, `vincent.goulet@act.ulaval.ca`

Keywords: Risk theory, ruin theory, compound models, phase-type distributions, **actuar**

actuar is a package providing additional Actuarial Science functionality to the R statistical system. This talk will present the features of the package targeted at risk theory calculations. Risk theory refers to a body of techniques to model and measure the risk associated with a portfolio of insurance contracts. A first approach consists in modeling the distribution of total claims over a fixed period of time using the classical collective model of risk theory. **actuar** provides functions to discretize continuous distributions and to compute the aggregate claim amount distribution using many techniques.

A second input of interest to the actuary is the evolution of the surplus of the insurance company over many periods of time. In *ruin theory*, the main quantity of interest is the probability that the surplus becomes negative, in which case technical ruin of the insurance company occurs. Function `ruin` of **actuar** computes ruin probabilities in the Cramér–Lundberg and Sparre Anderson models.

But for a few changes in terminology, the type of problems tackled by actuaries in insurance are very similar to risk measure problems in Finance or Hydrology, for example. Therefore, the talk may be of interest to a broad audience working with compound models and failure processes.

Dynamic transitions between multivariate methods

Michael Greenacre[1,*]

1. Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona

* Contact author: `michael@upf.es`

Keywords: Dynamic graphics, Multivariate analysis, Power transformations, Singular-value decomposition

The family of multivariate methods based on the singular value decomposition (SVD) is wide and varied. From the 'vanilla' option of principal component analysis (PCA) to the 'mixed fruit & chocolate sauce' options of canonical correspondence analysis (CCA) and weighted log-ratio analysis (LRA), we find a spectrum of methods applicable to continuous, categorical and compositional data, depending on certain choices made with respect to

1. transformation of the input data,

2. the weighting of the cases and variables, and

3. the metric defined between cases and between variables.

In this talk I introduce parameters into this family of methods to link them together in hybrids, making use of one or more of the above three aspects. The R environment then allows viewing the smooth transition between methods dynamically: for example, between PCA and correspondence analysis (CA), between PCA and linear discriminant analysis (LDA), between CA and LRA, between CA and CCA, between PCA and multiple correspondence analysis (MCA) via multiple factor analysis (MFA), and between the MCA of crisp and fuzzy coded data. Apart from the spin-off of hybridizing these methods and thereby enriching the family of possibilities, this approach provides striking pedagogical examples for understanding the differences in practice between the methods.

References

Greenacre, M. (2007). *Correspondence Analysis in Practice.* Chapman & Hall / CRC Press, London. Published in Spanish as *La Práctica del Análisis de Correspondencias.* Fundación BBVA, Madrid, 2008. Spanish edition available online for free download at
`http://www.fbbva.es/TLFU/tlfu/ing/publicaciones/libros/fichalibro/index.jsp?codigo=300`

Greenacre, M. (2009). Power transformations in correspondence analysis. *Computational Statistics and Data Analysis* **53**, 3107–3116.

Nenadić, O. and Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: the **ca** package. *Journal of Statistical Software* **20**(3).
`http://www.jstatsoft.org/v20/i03/`.

Design of Experiments in R

Ulrike Grömping

BHT Berlin – University of Applied Sciences, Germany
groemping@bht-berlin.de

Keywords: DoE, fractional factorial experiments, industrial experimentation, GUI, R commander plugin

R has a substantial amount of functionality for Design of Experiments (DoE) that is distributed over various R packages (cf. CRAN Task View, Grömping 2008-2009). However, so far, R has not been very successful in conquering the experimentation community outside of small expert circles. With regard to industrial experimentation, the market for DoE Software is dominated by all-round software companies like Minitab Inc. or Statsoft (Statistica), whose products are relatively simple to use. Such products are widely spread among businesses that adhere to the 6-Sigma quality management process, which involves application of DoE by many subject-matter and business process experts with only limited statistics training. Additionally, there are various specialized software products like e.g. NCSS (NCSS) or Cornerstone (Applied Materials, Inc.). The main inhibitors against usage of R for design and analysis of (industrial) experiments are

- R's steep learning curve for occasional or non-expert users
- gaps in R regarding some areas of experimental design, especially fractional factorial plans.

The talk presents a project that has two missions:

- to extend R's functionality for design and analysis of fractional factorial experiments (by extending R package **FrF2**) to fully meet state-of-the-art possibilities of benchmark software and exceed benchmark software by also incorporating newer research (e.g. Butler and Ramos 2007, Li and Lin 2003)
- to supply an interface to DoE functionality that accepts user inputs as close as possible to subject matter requirements and thus frees users from having to focus on unnecessary programming or mathematical detail. This purpose is attacked by a wrapper package for existing DoE functionality in R that acts as unifying interface and carries out translation from the subject-matter problem to the technical side as far as reasonable. The wrapper package will be usable from the command line, but also comes with a GUI that supports and guides occasional, programming-adverse, or statistically insecure R-users. The GUI will be implemented as an R commander plugin.

The project has just moved from the planning phase to the realization phase and will now be dealt with full time during my sabbatical semester. Rollout of the results to CRAN is scheduled for September 2009. The presentation will focus on the integration of existing functionality into an overall concept, selected aspects of implementing an application-driven interface, and selected aspects of extending R's DoE functionality for fractional factorial experiments. While the project will already be quite advanced, it will still be possible to accomodate useful suggestions from the R community.

References
Butler, N.A. and Ramos, V.M. (2007). Optimal additions to and deletions from two-level orthogonal arrays. *Journal of the Royal Statistical Society* **B 69**, 51-61.
Grömping, U. (2008-2009). *CRAN Task View on Design of Experiments.* http://cran.r-project.org/web/views/ExperimentalDesign.html.
Li, W. and Lin, D.K.J. (2003). Optimal Foldover Plans for Two-Level Fractional Factorial Designs. *Technometrics* **45**, 142-149.

SciViews-K and Komodo Edit, a new platform-independent GUI/IDE for R

Philippe Grosjean[1,*], Romain François[2] and Kamil Bartoń[3]

1. Numerical Ecology of Aquatic Systems, Mons University, 8 avenue du Champ de Mars, 7000 Mons, Belgium
2. Independent R consultant francoisromain@free.fr, France
3. Mammal Research Institute, Polish Academy of Sciences, Białowieża, Poland
* Contact author: phgrosjean@sciviews.org

Keywords: Code editor, Unit testing, Syntax highlighting, Object browser

SciViews was one of the first GUI available for R (Grosjean 2003), but it runs only under Windows. Still on Windows only, Tinn-R (http://www.sciviews.org/Tinn-R) has a certain success as simple, but efficient and R-aware, code editor. The new SciViews-K(http://www.sciviews.org/SciViews-K), combined with Komodo Edit merges together the main features of the old SciViews and Tinn-R, and adds some more like unit programming (live unit tests reports).

SciViews-K is a Mozilla extension for the Open Source code editor Komodo Edit (http://www.activestate.com/komodo_edit) that provides interaction between the editor and R through sockets. Komodo Edit is build on top of Mozilla and runs on Linux, Windows, Mac OS X, and a couple of other platforms.

So far, 90% of the features of both the old SciViews (object explorer, GUI similar to R Commander, Views), and of Tinn-R (code syntax highlighting, code tips, electronic R reference card, exhaustive set of functions to interact with R from the code editor) are reimplemented in SciViews-K. The plugin is still in beta stage, but version 1 will be released around October 2009.

References

Grosjean, Ph. (2003). SciViews: an object-oriented abstraction layer to design GUIs on top of various calculation kernels. *Proc. 3rd Int. Work. Dist. Stat. Comput.*, pp. 1–13.

Easy Execution of Data Mining Models through PMML

Alex Guazzelli[*], Wen-Ching Lin and Michael Zeller

Zementis, Inc., San Diego, CA, USA
*   Contact author: Alex.Guazzelli@zementis.com

Keywords: PMML, ADAPA, Model Deployment, Cloud Computing, Model Execution

PMML (Predictive Modeling Markup Language) is an XML-based language used to define data mining models. It was specified by the Data Mining Group, an independent group of leading technology companies. By providing a uniform standard to represent predictive models, PMML allows for the exchange of predictive solutions between different applications and various vendors. Many statistical packages already support the PMML standard; these include, for example, SAS and SPSS. In an effort to broaden the scientific workbench available to data mining scientists and to support the open source community, Zementis recently contributed code to the R project. In particular, we implemented the export of neural network models built with the *nnet* R package available through the *VR bundle* package (Venables and Ripley, 2002) as well as Support Vector Machines built with the *kernlab* R package (Karatzoglou et al., 2008) for objects of class *ksvm*. The same PMML exporter package (Williams et al., 2009) can also produce decision trees built with *rpart* (Therneau and port by Brian Ripley, 2008) and linear regression models as well as binary logistic regression models for objects of class *lm* and *glm* from *stats*. The PMML exporter package is currently available through CRAN (the Comprehensive R Archive Network).

All of the R exported PMML 3.2 models are readily available to be uploaded into an execution engine for scoring or classification. For example, the ADAPA engine, which can be used for testing and exploration, can be downloaded as a gadget and added to a personalized iGoogle console. This service is available free of charge and leverages the Amazon Elastic Compute Cloud (Amazon EC2).

Our aim here is to show how one can quickly build a data mining model in R, such as a Support Vector Machine, and use the PMML package to produce a model file which can be uploaded and executed in a different application. We demonstrate how one can use data containing expected results to verify correct model deployment. If all computed and expected values match, the model can be considered ready for production, i.e. available for generating predictions on incoming data as part of an overall enterprise decision management strategy. From R to ADAPA, we use PMML as an effective way to express and execute data mining models.

Our work shows how PMML can be effectively used to allow for model exchange between different applications. Also, it highlights how one can benefit from an open-source statistical package such as R to easily export models into PMML and upload them into ADAPA, a light-weight scoring engine which consumes several PMML 3.2 models and data transformations. The ease of model expression and execution allows data mining scientists to concentrate on the important tasks: data analysis and model building. Real-time, scalable execution is handled through software tools which communicate through a common language, PMML.

References

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

A. Karatzoglou, A. Smola, and K. Hornik (2008). *The kernlab package*.
    http://cran.R-project.org/web/packages/kernlab. R package version 0.9-8.

T. M. Therneau and B. A. R. port by Brian Ripley (2008). *Rpart: Recursive Partitioning*.
    http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm. R package version 3.1-42.

W. N. Venables and B. D. Ripley (2002). *Modern Applied Statistics with S*. Statistics and Computing. Springer,
    New York, 4th edition.
    http://cran.R-project.org/web/packages/VR. R package version 7.2-45.

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

Model visualisation (with ggplot2)

Hadley Wickham[1,*]

1. Rice University, Houston TX.

* Contact author: hadley@rice.edu

R has many packages for visualising data, in a huge number of flexible ways. Unfortunately the capabilities for visualising statistical models are much more limited, and most packages only provide a fixed set of visualisations with very little ability to customise the view. This talk introduces the new tools provided by ggplot2 for visualising models and discusses the philosophy behind my approach. You'll see concrete examples of how keeping apart computation and presentation pays off with the ability to construct visualisations tailored your problems. My examples will focus on mixed models, but the principles apply to all statistical models.

LMMNorm: a package for the normalization of microarrays using linear mixed models

Philippe Haldermans[1,*] **, Ziv Shkedy**[1]

1. Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Center for Statistics, Hasselt University, Agoralaan 1,B-3590 Diepenbeek,Belgium

* Contact author: philippe.haldermans@uhasselt.be

Keywords: microarrays, normalization, linear mixed model, graphical user interface

During the last years a large expansion in the research on microarrays has been observed. Many techniques for the normalization of these microarrays have been proposed and implemented, for instance nonlinear normalization (Park *et al.* (2003)) and print-tip dependent normalization (Dudoit *et al.* (2002)) . Several of these normalization methods can be written in terms of a linear mixed model as described in Haldermans *et al.* (2007) . This enables us to use an objective selection criteria, such as Akaike Information Criterion (AIC), to determine which is the best normalization method for a given array.

The `LMMNorm` package normalizes cDNA microarray data using linear mixed models (LMM) as scatterplot smoothers of the MA-plot. This can be done since the normalization models, i.e. global, linear and nonlinear can be formulated as LMM:

$$\mathbf{M}(\mathbf{A}) = \left\{ \begin{array}{ll} \beta_0 & \text{global,} \\ \beta_0 + \beta_1 \mathbf{A} & \text{linear,} \\ \beta_0 + \beta_1 \mathbf{A} + \mathbf{Z}\mathbf{b} & \text{nonlinear.} \end{array} \right. \tag{1}$$

Other normalization models, such as print-tip specific normalization (Dudoit *et al.*, 2002) or normalization with non-constant variance models can be expressed as a LMM as well. These models were implemented using the R function `lme()`. The `LMMNorm` package contains functions that automatically create the requested models, for instance the construction of the design matrix $\mathbf{Z}$ in the nonlinear model for the random effects, and compares the models using a selection criteria like AIC to produce the most appropriate model.

To facilitate the use of these functions for unexperienced users, a graphical user interface was developed. This offers the users a point and click environment, which allows them to use the statistical methodology without the need for a thorough knowledge of the model framework. It provides plots before and after normalization making it easy to see the effect of the normalization method. The GUI makes extensive use of the interface implemented in the `gWidgets` package.

References

[1] Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed, *Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments*, Statistica Sinica **12** (2002), no. 1, 111–139.

[2] Philippe Haldermans, Ziv Shkedy, Suzy Van Sanden, Tomasz Burzykowski, and Aerts Marc, *Using linear mixed models for normalization of cdna microarrays*, Statistical Applications in Genetics and Molecular Biology **6** (2007), Article 19.

[3] Taesung Park, Sung-Gon Yi, Sung-Hyun Kang, SeungYeoun Lee, Yong-Sung Lee, and Richard Simon, *Evaluation of normalization methods for microarray data*, BMC Bioinformatics **4** (2003), 33.

Using R For Flexible Modeling Of Pre-Clinical Combination Studies

Chris Harbron

1. Discovery Statistics, AstraZeneca
* Contact author:Chris.Harbron@AstraZeneca.Com

Keywords: Combinations, Synergy, Non-linear models.

There is a strong and increasing interest in the development of drug combination therapies within the pharmaceutical industry, so a robust evaluation of the potential for compounds to interact synergistically at an early stage within the pharmaceutical pipeline is clearly of great value.

I will describe a newly developed unified approach for assessing synergy which allows a number of different assessments to be made and compared under a common framework, powerfully and flexibly using all the available experimental data and giving a complete description of the studied combination space with statements of confidence. The method is applicable over wide classes of experimental design and response patterns and is backed up with informative graphical displays.

The R language, using the *nls()* function provides a powerful environment to implement these assessments, although due to the functional form of the models and the variety of data scenarios encountered, some adaptations are required. Features of this implementation ensuring that these models can be implemented reliably, robustly and efficiently will be highlighted and discussed.

mult: a Multivariate R Package
with a Dynamic Java Frontend

E. James Harner[1*] **and Dajie Luo**[1]

1. Dept. of Statistics, West Virginia University, USA

* Contact author: jharner@stat.wvu.edu

`mult` is yet another R package for doing multivariate analyses. However, `mult` has two unique features which greatly enhance its usefulness. First, `mult` is complete in the sense that it includes canonical and partial analyses. Second, these analyses can be visualized using high-dimensional dynamic graphics.

As an R package, `mult` uses S3 classes, and a function returns the output as an object whose class corresponding to the function call, e.g., the `pca` function returns an object of class `pca`. The analyses are quite general. For example, principal component analysis not only allows for robust versions, but also for partial, canonical, and partial canonical variants. Factor analysis, multiple correlation analysis, canonical correlation analysis, and discriminant analysis also have flexible implementations. Standard R graphics, e.g., biplots, are supported.

Multivariate space is most naturally explored by interactive and dynamic graphics. JavaStat (Harner and Luo, 2009) was designed to support highly interactive data analyses and dynamic graphics in a platform-independent way. The architecture of JavaStat is described elsewhere, e.g, Harner, Luo, and Tan (2008a, 2008b). JavaStat also acts as a front-end to R, but the use of R as a compute engine is transparent to the user and is optional. JavaStat has a client/server architecture to allow for high-performance computing, but the server can reside on the same machine as the client.

JavaStat has an interface to the `mult` package. Basically, certain menu selections in JavaStat trigger a series of actions. First, a Java object corresponding to the selected variables in a data table is sent from the client to the server using RMI. On the server side, the object is taken apart into a bunch of arrays, e.g., a double array represents a numeric column in the original table. Then the JRI API is used to convert these arrays into vectors in R. These vectors are created in the current workspace of R and are assigned names. Optionally a data frame can be created. After preparing data in the workspace, a modeling command, e.g., `pca`, can be issued to R through a JRI evaluation method. Since the returned object exists in the current workspace, subsequent commands are sent to extract useful results from this object. Those results are then converted into Java objects and assembled into a more complex Java object, which is sent back to the client as a whole.

At this point, the user can generate dynamic plots, e.g., a constrained grand tour or a dynamic high-dimensional biplot, or do subsequent analyses. The analyses and plots based on the returned values from R can be supplemented by the built-in functionalities of JavaStat, e.g., a dynamic parallel coordinate plot.

References

Harner, E. James and Luo, Dajie (2009). JavaStat Home,
     `http://javastat.stat.wvu.edu`

Harner, E. James, Luo, Dajie, and Tan, Jun (2008a). JavaStat: a Java/R-based statistical computing environment. *Computational Statistics* , Online SpringerLink version available in 2008.

Harner, E. James, Luo, Dajie, and Tan, Jun (2008b). JavaStat: a Java-based R Front-end, User!2008 (Dortmund, Germany), July 2008.

Indices for measuring location impact in Bayesian spatial models for agricultural field trials

Jay Harrison

Monsanto Company, St. Louis, Missouri USA
jmharri@monsanto.com

Researchers working with agricultural field trials conducted at multiple farm sites often request an assessment of the validity the data at each location. Traditionally, the coefficients of variation within the locations have been used, but alternate measures of validity based on statistics derived from frequentist linear models have been proposed by Bowman and Rawlings (1995), Bowman and Watson (1997), and Beckman, Nachtsheim, and Cook (1987). A method for assessing the impact of each location on each parameter in a Bayesian model with spatial correlation among the locations will be proposed. The method involves graphical comparisons and numerical summaries of posterior distributions from Markov Chain Monte Carlo samples obtained by using the full data set and then omitting sets of locations. The BRugs function in R is used to repeatedly call WinBUGS software to obtain the estimates for the posterior distributions. The procedure will be illustrated with an example from a hybrid comparison trial as facilitated using Darwine software on a Mac OS X operating system.

References

Bannerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Boca Raton, Florida: Chapman and Hall/CRC.

Beckman, R. J., Nachtsheim, C. J., and Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, 29, 413-426.

Bowman, D. T. and Rawlings, J. O. (1995). Establishing a rejection procedure for crop performance data. *Agronomy Journal*, 87, 147-151.

Bowman, D. T. and Watson, C. E. (1997). Measures of validity in cultivar performance trials. *Agronomy Journal*, 89, 860-866.

Carlin, B. P. (2009). Introduction to BRugs, software for calling OpenBUGS from R. http://www.biostat.umn.edu/~brad/software/BRugs/BRugs_install.html .

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical methods, second edition*. New York: Wiley.

Kronenberg.org. (2009). Darwine builds for OS X. http://www.kronenberg.org/darwine/ .

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.

Regularization Paths for Generalized Linear Models via Coordinate Descent

Jerome Friedman [1] , **Trevor Hastie** [1,2,*], **Rob Tibshirani** [2,1]

1. Statistics Department, Stanford University
2. Department of Health, Research and Policy, Stanford University

* Contact author: hastie@stanford.edu

Keywords: regularization, lasso, path, coordinate descent,logistic regression

We develop fast algorithms for estimation of generalized linear models with convex penalties. The models include linear regression, two-class logistic regression, and multinomial regression problems while the penalties include ℓ_1 (the lasso), ℓ_2 (ridge regression) and mixtures of the two (the elastic net). The algorithms use cyclical coordinate descent, computed along a regularization path. The methods can handle large problems and can also deal efficiently with sparse features. In comparative timings we find that the new algorithms are considerably faster than competing methods. The software is on CRAN in the package `glmnet`.

References

Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *Pathwise coordinate optimization*, Annals of Applied Statistics **1** (2007), no. 2, 302–332.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, Tech. report, Stanford University, 2008.

Trevor Hastie, Rob Tibshirani, and Jerome Friedman, *The elements of statistical learning: Prediction, inference and data mining (second edition)*, Springer Verlag, New York, 2009.

Computational Social Sciences using R

Ajit Kumar[1]   Neeraj Hatekar[2]

1. Department of Applied Mathematics, Institute of Chemical Technology, Mumbai, INDIA 400 019
Email: ajit72@gmail.com
2. Department of Economics, University of Mumbai, INDIA 400 098
Email: neeraj.hatekar@gmail.com

Keywords: CSS, R

The computational social science programme (CSS) at the University of Mumbai envisages using R as a large scale data mining tool as well as for statistical modelling of previously unpublished data. The first section of the paper examines the various applications of R in the interdisciplinary CSS programme. One of the objectives of the CSS programme is to computerise large unpublished data bases that are of relevance to social scientists and develop software tools that will enable practitioners to query these data on various dimensions. In particular, Western India has a large number of unpublished data sources on demographic variables like age at marriage, fertility, birth spacing, educational attainments for women that stretch over nearly a century and a half. These have been only rarely used (Hatekar, Mathur, Rege (1997)) as they are not available in a computerised form, and since appropriate tools do not exist in order to query them. The CSS programme has been digitising these data bases and using R for statistical modelling. The current paper presents one such application on modelling the fertility transition that took place in the 1920's in upper caste households in Western India. The fertility shift is modelled and shown to be associated with larger cultural shifts in the ideas of ideal households and housewives.

In the second section we deal with various issues regarding use of R in social sciences and economics in India. We also touch upon advantages and disadvantages of using R, some of challenges and difficulties in using R and how to overcome these challenges.

References

(Hatekar, Mathur, Rege (1997): "Social Reform by Legislation: The Strange Case of the Sarda Act", Economic and Political Weekly, January 1-7, 2007.

Kleiber C. and A. Zeileis (2008): Applied Econometrics with R, Springer.

R-project home page `http://www.r-project.org/`

Dynamic Control of R Graphics through RExcel

Richard M. Heiberger[1,*]

1. Temple University, Philadelphia, PA, USA
* Contact author: rmh@temple.edu

Keywords: RExcel, Dynamic Graphics, Adverse Events, Teaching

R provides powerful graphic tools. R also has a high startup cost for non-technical users. Excel is already on almost everyone's desk, provides a familiar interface, and has many control mechanisms (sliders, checkboxes, option buttons, double-clicking) with which users are comfortable. It is relatively easy to place complex R graphs into the the Excel automatic recalculation model, so the graphs are automatically updated when the data or the control mechanisms are changed on the spreadsheet. In this paper we present and discuss the behind-the-scenes details of several R graphical displays that are accessed and controlled through simple and familiar widgets.

Dynamic displays can be designed for different audience assumptions. The normal and t plot, designed for the introductory course, shows a graph of significance and power for the normal and t-tests. We adjust sliders to illustrate how the power changes as the sample mean $\bar{x}$ changes and as the location of the alternative value of the population mean μ_1 changes. The Adverse Events Dotplot, designed for the monitoring of safety data collected during clinical trials, shows the relative risk of various adverse events. We click the data to change the display characteristics of the plot, for example, to emphasize the risk or the actual frequency of occurence of the types of events.

We illustrate and discuss the technical capabilities of the interface, the characteristics of the intended audience for these displays, and design decisions we made based on these considerations.

References

Richard M. Heiberger and Erich Neuwirth (scheduled for 2009). *R through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*, Springer: UseR! series.
http://rcom.univie.ac.at.

Ohad Amit and Richard M. Heiberger and Peter W. Lane (2008). Graphical Approaches to the Analysis of Safety Data from Clinical Trials *Pharmaceutical Statistics*, 7, 1, 20–35.
http://www3.interscience.wiley.com/journal/114129388/abstract.

A new R bundle for design and analysis of computer experiments

Céline Helbert[1,*]**, Delphine Dupuy**[1] **and Yves Deville**[2]

1. École des Mines de Saint-Étienne (Dpt. G2I/3MI) 158, Cours Fauriel - 42023 Saint-Étienne Cedex 2 (France)
2. Statistical Consultant - 37, rue Lamartine 73000 Chambéry (France) (`www.alpestat.com`)
* Contact author: helbert@emse.fr

Keywords: computer experiments, space-filling designs, metamodeling, kriging, global optimization.

The DICE consortium* is a partnership between academic and industrial laboratories from various fields (e.g. petroleum, automotive, nuclear etc.). DICE aims at developing and applying statistical methods to solve problems related to computer experiments: uncertainty propagation through a flow simulator, global optimization of a car crash simulator or calculus of failure probabilities in a nuclear process.

Each methods used through the consortium applications are implemented and integrated in a numerical toolbox that allows to fully process a case study. This toolbox is actually a R bundle comprising the following 4 packages:

- `DiceDesign` for space-filling design of experiments;
- `DiceEval` for evaluation and comparison of metamodels;
- `DiceKriging` for the estimation of a response surface via Gausian Processes;
- `DiceOptim` for global optimization.

Each package addresses a specific task where classical and new methods are implemented. Among the new methods, `DiceDesign` proposes the construction of numerical designs of experiments based on stochastic processes and the computation of discrepancy and distance criteria. `DiceEval` is especially dedicated to fit usual models. A validation procedure (containing numerical criteria, graphical plots and cross-validation methods) is also proposed to validate a fitted model. `DiceKriging` implements a large panel of kriging models used for uncertainty propagation or for global optimization. The component `DiceOptim` performs different versions of EGO algorithm (Jones et al., 1998).

The different case studies proposed by the industrial partners of the DICE consortium have been processed by the bundle. One of these case studies will served as a running example to illustrate its main functionalities. In this application, the space-filling design (inputs of the simulations) is done by `DiceDesign`. `DiceEval` is then used to modelize the output of the simulator and allows to fit classical models: linear models, additive{`gam`}, PolyMARS{`polspline`} and kriging{`DiceKriging`} models. Moreover, `DiceEval` provides graphical tools in order to compare the quality of the fitted models on learning, validation and test sets.

Acknowledgements.
This work was conducted within the frame of the DICE consortium between ARMINES, Renault, EDF, IRSN, ONERA and Total S.A. The authors wish to thank J. Franco, O. Roustant, D. Ginsbourger and L. Carraro for their contributions.

References

Hastie T., Tibshirani R. and Friedman J. (2001). **The Elements of Statistical Learning: Data Mining, Inference and Prediction**, Springer.

Jones D.R., Schonlau M. and Welch W.J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global optimization*, 13, 455–492.

Welch W.J., Buck R.J., Sacks J., Wynn H.P., Mitchell T.J. and Morris M. (1992). Screening, Predicting, and Computer Experiment. *Technometrics*, 34 (1), 15–25.

*DICE stands for *Deep Inside Computer Experiments* (`http://dice-consortium.fr/`).

82

Multivariate Process Monitoring and Control with R

Elisa Henning [1,2,*] , **Custodio Cunha Alves** [1,3], **Robert Wayne Samohyl** [1]

1. Federal University of Santa Catarina
2. University of Santa Catarina State
3. University of Joinville Region
* Contact author: dma2eh@joinville.udesc.br

Simultaneously monitoring two or more quality characteristics depends on the development of more specific statistical tools to detect, identify and analyze the major causes of variability that affect the behavior of the production process. The multivariate control charts represent one of these emerging statistical techniques successfully used to monitor simultaneously several correlated characteristics that indicate the quality of a single production process. The use of graphics in the industrial environment has increased in recent years due to many resources of information technology now available to reduce the complexity of modern industrial processes. This article presents some computational routines developed in the GNU R package for the application of statistical control for multivariate processes based on the cumulative sum (MCUSUM) and exponentially weighted moving average (MEWMA). In order to reduce the number of variables Principal Components Analysis (PCA) was adopted making it possible to consider all of the original variables in only two or three dimensions. Thus, most of the variance of the process is represented by the dispersion of the points on the main components. The routines were developed in R in order to facilitate information entry to produce clear graphics and to return the maximum amount of information needed for process monitoring. The routines were applied successfully to data in the literature. While these routines can still be improved upon, we can conclude that the R environment is an important alternative for the diagnosis and monitoring of multivariate industrial processes.

References

Henning, E., Alves, C. C; Samohyl, R. W (2008). The development of graphics and control MCUSUM environment MEWMA in R as an alternative procedure for statistical analysis of multivariate processes. ENEGEP 2008 (Rio de Janeiro, Brazil), pp. 1-10.

Jackson, J.E (1980). Principal Components and factor analysis: Part I. Journal of Quality Technology, 12 (4), 201-213.

Montgomery, D. C (2004). Introduction to Statistical Quality Control. LTC Editora, 4th edition. Rio de Janeiro.

Efficiency Analysis in **R** using Parametric, Semiparametric, and Nonparametric Methods

Arne Henningsen[1,2,*]**, Subal Kumbhakar**[3]

1. Department of Agricultural Economics, University of Kiel (Germany)
2. Institute of Food and Resource Economics, University of Copenhagen (Denmark)
3. Department of Economics, State University of New York at Binghamton (USA)

* Contact author: arne.henningsen@gmail.com

Keywords: Efficiency, Productivity, Parametric, Semiparametric, Nonparametric

Efficiency and productivity analysis is a major field in applied production economics. It is generally dominated by two methods: the parametric Stochastic Frontier Analysis (SFA) and the nonparametric and deterministic Data Envelopment Analysis (DEA). The SFA can be done in R with the **frontier** package [1] and the DEA might be done with the **FEAR**[1] package [4]. The SFA approach contains a stochastic error term and hence, is suitable even if there is some "noise" in the data. However, this parametric approach requires the specification of an explicit functional form, although the functional form cannot be derived from theory. Selecting a wrong functional form may lead to severely biased estimation results. If the data set includes production units with rather different technologies, even flexible functional forms cannot model their production technologies adequately and hence, the parametric SFA is inappropriate. In contrast, the nonparametric and deterministic DEA approach does not require the specification of a functional form, but it does not include a stochastic component. Hence, the DEA is not suitable in case of "noisy" data.

In many real world applications, the data are noisy *and* production units have rather different technologies (in parametric sense) so that a stochastic and nonparametric approach is required and neither the SFA nor the DEA is appropriate. In cases like this, a semiparametric SFA [2] is appropriate, because it allows for statistical "noise" and does not require the specification of a functional form for production technologies. In a first step, a nonparametric production function is estimated and in a second step the residuals of the first step are used to estimate inefficiencies. Although in many empirical applications this approach seems to be more appropriate than the SFA and DEA, it has not been used much in applied studies, probably because of nonavailability of user-friendly software. However, several soft-ware packages for nonparametric econometrics have become available in recent years. For instance, the powerful and feature-rich **np** package [3] can be used in the first step to estimate the nonparametric production function and the **frontier** package [1] can be used in the second step to estimate the technical efficiencies.

We will demonstrate how the three approaches (SFA, DEA, and semiparametric SFA) can be used for applied efficiency analysis in **R** and we compare the results obtained from all three approaches.

References

[1] Tim Coelli and Arne Henningsen, *frontier: Stochastic frontier analysis*, 2008, R package version 0.9, http://CRAN.R-project.org.

[2] Yanqin Fan, Qi Li, and Alfons Weersink, *Semiparametric estimation of stochastic production frontier models*, Journal of Business and Economic Statistics **14** (1996), no. 4, 460–68.

[3] Tristen Hayfield and Jeffrey S. Racine, *Nonparametric econometrics: The np package*, Journal of Statistical Software **27** (2008), no. 5, 1–32.

[4] Paul W. Wilson, *FEAR: A software package for frontier efficiency analysis with R*, Socio-Economic Planning Sciences **42** (2008), no. 4, 247–254.

---

[1]Please note that the non-academic use of the **FEAR** package is restricted and that this closed-source software is available as binary package for MS-Windows only.

Subject Randomization System

Solomon Henry[1], Douglas Wood[1], and Balasubramanian Narasimhan[1,2,*]

1. Department of Health Research and Policy
2. Department of Statistics

* Contact author: naras@stanford.edu

Keywords: Clinical Trial, Randomization, Web interface

SRS is a system that can be used for subject randomization for clinical trials. Currently it implements biased coin designs of Efron, Wei and the minimization method of Pocock and Simon. The core system is implemented as an R package, and a web interface allows one to define the characteristics of a clinical experiment and register subjects. Built using open-source software, The system can be used for subject randomization in multi-center clinical trials.

References

Efron, 1971 B. Efron, Forcing a sequential experiment to be balanced, Biometrika 58 (1971), pp. 403417.

Wei, 1978b L.J. Wei, The adaptive biased coin designs for sequential experiments, Ann. Statist. 6 (1978) pp 92–99.

Pocock and Simon, 1975 S.J. Pocock and R. Simon, Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trials, Biometrics 31 (1975), pp. 103115.

Sublogo dendrograms: visualizing correlation in biological sequence motifs

Toby Dylan Hocking[1,*]

1. LSTA; Université Paris 6, 175 Rue du Chevaleret, 75013 Paris, France

* Contact author: tdhock@ocf.berkeley.edu

Keywords: Sequence logo, statistical graphics, biological sequence correlation, clustering

DNA and protein sequence motifs are usually visualized using graphical sequence logo plots (Schneider and Stephens, 1990). Though sequence logos excel at communicating the information at each position in the motif, one problem with their use is that they make no attempt to show the joint distribution between positions. I propose the sublogo dendrogram as a new type of statistical plot that uses logos and dendrograms to show this joint distribution. In addition, sublogo dendrograms reveal significant details in subfamilies of sequences that can go unnoticed using a standard logo.

In this work, sublogo dendrograms have been implemented using the WebLogo program (Crooks *et al.*, 2004) in conjunction with R's `grImport` package (Murrell, in press). First, R is used to perform a hierarchical clustering on the aligned input sequences. The tree that results from the clustering is cut, yielding several subfamilies of sequences. Next, WebLogo creates an overall logo image for the entire set of sequences as well as a "sublogo" for each subfamily. The dendrogram resulting from the hierarchical clustering is drawn in the center of the plot, with the overall logo on the left, and the sublogos on the right, next to the relevant leaves of the dendrogram.

Zinc finger protein recognition helix sequences selected to bind triplet GGC

For user convenience, a web server has been set up to make sublogo dendrograms. Furthermore, the R software and code for the web interface is freely available for download, usage and modification.

`http://www.ocf.berkeley.edu/~tdhock/sublogo_dendrogram/form.php`

References

Schneider TD Stephens RM (1990). Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.*, 18, 6097–6100.

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004). WebLogo: A sequence logo generator. *Genome Research*, 14, 1188–1190,
`http://weblogo.berkeley.edu`

Paul Murrell (in press). Importing Vector Graphics: the GrImport package for R. *Journal of Statistical Software*,
`http://www.stat.auckland.ac.nz/~paul/R/grImport/import.pdf`.

CoxFlexBoost: Fitting Structured Survival Models

Benjamin Hofner[1,*] **, Torsten Hothorn**[2] **and Thomas Kneib**[2]

1. Institut für Medizininformatik, Biometrie und Epidemiologie; Friedrich-Alexander-Universität Erlangen-Nürnberg
2. Institut für Statistik; Ludwig-Maximilians-Universität München

* Contact author: `benjamin.hofner@imbe.med.uni-erlangen.de`

Keywords: likelihood-based boosting, hazard regression, model choice, smooth effects, time-varying effects

In many situations, medical applications require flexible survival models that allow to extend the classical Cox-model via the inclusion of time-varying and nonparametric effects. These structured survival models are very flexible but additional difficulties arise when model choice and variable selection are desired. In particular, it has to be decided which covariates should be assigned time-varying effects or whether linear effects are sufficient for a given covariate. Component-wise boosting (e.g., Bühlmann & Hothorn, 2007) provides a means of likelihood-based model fitting that enables simultaneous variable selection and model choice. Extending likelihood-basting boosting algorithms for generalised additive models proposed in Tutz & Binder (2006), we developed a component- wise, likelihood-based boosting algorithm for survival data that permits the inclusion of both parametric and nonparametric time-varying effects as well as nonparametric effects of continuous covariates utilizing P-splines as the main modeling technique (Hofner *et al.*, 2008). The properties and performance of the algorithm were investigated in a simulation study. A software implementation is available in the R package **CoxFlexBoost** (Hofner, 2008).

References

Bühlmann, P. & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22, 477–505.

Hofner, B. (2008). *CoxFlexBoost: Boosting flexible Cox models (with time-varying effects)*. R package version 0.6-0.
`http://R-forge.R-project.org/projects/coxflexboost`.

Hofner, B., Hothorn, T. and Kneib, T. (2008). Variable selection and model choice in structured survival models. *Technical Report No. 43, Institut für Statistik, Ludwig-Maximilans-Universität München*.

Tutz, G. & Binder, H. (2006). Generalized additive modelling with implicit variable selection by likelihood-based boosting.. *Statistical Science*, 22, 477–505.

partykit: A Toolbox for Recursive Partytioning

Torsten Hothorn[1] **and Achim Zeileis**[2]

1. Institut für Statistik, Ludwig-Maximilians-Universität München, Germany
   `Torsten.Hothorn@R-project.org`

2. Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Austria
   `Achim.Zeileis@R-project.org`

Recursive partitioning methods, or simply "trees", are simple yet powerful methods for capturing regression relationships. Since the publication of the automated interaction detection (AID) algorithm in 1964, many extensions, modifications, and new approaches have been suggested in both the statistics and machine learning communities. Most of the standard algorithms are available to the R user, e.g., through packages **rpart** (Therneau *et al.*, 2009), **RWeka** (Hornik *et al.*, 2009), **party** (Hothorn *et al.*, 2009), or **mvpart** (De'ath, 2007).

However, no common infrastructure is available for representing trees fitted by different packages. Consequently, the capabilities for extraction of information—such as predictions, printed summaries, or visualizations—vary between packages and come with somewhat different user interfaces. Furthermore, extensions or modifications often require considerable programming effort, e.g., if the median instead of the mean of a numerical response should be predicted in each leaf of an 'rpart' tree. Similarly, implementations of new tree algorithms might also require new infrastructure if they have features not available in the above-mentioned packages, e.g., multi-way splits or more complex models in the leafs.

To overcome these difficulties, the **partykit** package (Hothorn and Zeileis, 2009) offers a unified representation of tree objects along with `predict()`, `print()`, and `plot()` methods. Trees are represented through a new flexible class 'party' which can, in principle, capture all trees mentioned above but can also accommodate multi-way or functional splits, as well as complex models in (leaf) nodes. The package is currently under development at R-Forge but already provides conversion methods for trees of classes 'rpart', 'J48', and 'pmmlTreeModel' as well as a re-implementation of conditional inference trees (Hothorn *et al.*, 2006).

In our presentation, we will only sketch details of these classes and corresponding methods and focus on applications of the toolkit including extended visualizations for 'rpart' or 'J48' objects, fast predictions on millions of new observations, and a new implementation of the classical CHAID algorithm.

References

De'ath G (2007). *mvpart: Multivariate Partitioning*. R package version 1.2-6, URL `http://CRAN.R-project.org/package=mvpart`.

Hornik K, Zeileis A, Hothorn T, Buchta C (2009). *RWeka: An R Interface to Weka*. R package version 0.3-16, URL `http://CRAN.R-project.org/package=RWeka`.

Hothorn T, Hornik K, Strobl C, Zeileis A (2009). *party: A Laboratory for Recursive Partytioning*. R package version 0.9-995, URL `http://CRAN.R-project.org/package=party`.

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674. doi:10.1198/106186006X133933.

Hothorn T, Zeileis A (2009). *partykit: A Toolkit for Recursive Partytioning*. R package version 0.0-1, URL `http://R-Forge.R-project.org/projects/partykit/`.

Therneau TM, Atkinson B, Ripley BD (2009). *rpart: Recursive Partitioning*. R package version 3.1-42, URL `http://CRAN.R-project.org/package=rpart`.

A suite of R packages for the analysis of DNA copy number microarray experiments

Philippe Hupé[1,2,*]

1. U900 Institut Curie, INSERM, Mines ParisTech
2. UMR144 Institut Curie, CNRS

* Contact author: philippe.hupe@curie.fr

Keywords: DNA copy number, microarray, biostatistics algorithms, oncology

Microarray technology is a powerful tool very helpful in oncology in order to better understand the molecular mechanisms involved in tumoral progression. A ccommon charecteristic of tumours is the presence of chromosome alterations and especially a change of their DNA copy number. There are microarrays which allow the quantification of DNA copy number. The raw data obtained from the microarray technology need appropriate statistical processing so that they can be biologically and clinically meaningful. Thus, we developed statistical methods in order to normalise and extract the biological information from microarrays devoted to the study of DNA copy number in tumours. Our methods are implemented within three R packages which are part of the Bioconductor project:

MANOR - Neuvial et al. (2006) This package implements a normalisation method devoted to the spatial normalisation of DNA copy number data. Briefly, the method consists of a spatial smoothing of the data followed by a segmentation which identifies aberrant spatial areas on the chip.

GLAD - Hupé et al. (2004) This package allows the detection of breakpoints in the DNA copy number molecular profiles (this step is called *segmentation*) and the assignment of a status (either loss, normal, gain or amplification) to each region identified (this step is called *calling*). The calling step provides valuable information for downstream analyses. The development of such an algorithm also avoid the tedious task of a manual expertise which is subject to error, non-reproducible and time-consuming (and even untractable for high-density chips).

ITALICS - Rigail et al. (2008) This package proposed a normalisation method devoted to the analysis of Affymetrix® Genome-Wide Human SNP Array. Besides normalisation, the proposed method has the originality to perform the identification of the DNA copy number alterations using the GLAD algorithm. The algorithm alternatively identifies the DNA copy number alterations and normalises the data. Those two alternative steps are iterated to improve the signal-to-noise ratio of the data at each iteration. The normalisation step takes into account the information of the genome alterations to better estimate the sources of variability to correct during the normalisation step.

The packages we have developped have already been widely used within the scientific community. Moreover, Institut Curie has developed client/server application named CAPweb which integrates the three previous packages (Liva et al., 2006). CAPweb is a user-friendly tool enabling biologists to analyse DNA copy number experiments from raw data to visualisation and biological interpretation. With CAPweb it is possible to manage the data, to normalise the DNA copy number experiments data with MANOR, to detect breakpoints with GLAD, to analyse Affymetrix data with ITALICS, to visualise and analyse the genomic profiles with VAMP (La Rosa et al., 2006).

References

Hupé et al. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics.

La Rosa et al. (2006). VAMP : Visualization and Analysis of CGH array, transcriptome and other Molecular Profiles. Bioinformatics.

Liva et al. (2006). CAPweb: a bioinformatics CGH array Analysis Platform. Nucleic Acids Research.

Neuvial et al. (2006). Spatial normalization of array-CGH data. BMC Bioinformatics.

Rigaill et al. (2008). ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics.

Eulerian tour algorithms for data visualization and the `PairViz` package

Catherine B. Hurley[1,*] , R.W Oldford[2]

1. Department of Mathematics, National University of Ireland, Maynooth, Co. Kildare, Ireland
2. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

* Contact author: catherine.hurley@nuim.ie

Keywords: Graph; Hamiltonian; Eulerian tour; ordering; visualization

`PairViz` is an R package that produces orderings of statistical objects for visualization purposes. We abstract the ordering problem to one of constructing edge-traversals of (possibly weighted) graphs. `PairViz` implements various edge traversal algorithms which are based on Eulerian tours and Hamiltonian decompositions. We describe these algorithms, their `PairViz` implementation and discuss their properties and performance. We illustrate their application to various visualization problems.

References

C.B. Hurley and R.W. Oldford (2008). Visualization using Eulerian tours and Hamiltonian decompositions .R package on CRAN.

C.B. Hurley and R.W. Oldford (2008). Eulerian tour algorithms for data visualization and the `PairViz` package. Submitted.

C.B. Hurley and R.W. Oldford (2008). Pairwise display of high dimensional information via Eulerian tours and Hamiltonian decompositions Submitted.

Financial econometrics based on stochastic differential equations and the `sde` package

Stefano Maria Iacus[1,*]

1. Department of Economics, Business and Statistics * Contact author: stefano.iacus@unimi.it

Keywords: financial econometrics, quasi-likelihood analysis, stochastic differential equations, simulation, mathematical finance

In this talk we will introduce the package `sde` which contains generic functions for simulation and inference on stochastic differential equations. In particular, stochastic differential equations corresponding to diffusion processes driven by the Wiener process are considered.

Most of the theoretical results in modern finance rely on the assumption that the underlying dynamics of asset prices, currencies exchange rates, interest rates, etc are continuous time stochastic processes driven by stochastic differential equations. Continuous time models are also at the basis of option pricing and option pricing often requires Monte Carlo methods. In turn, the Monte Carlo method requires a preliminary good model to simulate whose parameters has to be estimated from the data. On the other side, most applications in financial econometrics make use of pure time series modeling because many statistical procedures are already available in many statistical packages.

The discrepancy between theoretical and applied mathematical finance is motivated by the fact that while the model is continuous, the observations always come in discrete time. Inference for continuous time data from stochastic differential equation dates back to Jacod and Shiryayev (1987) and today is considered as a solved problem. On the contrary, the likelihood function for discretized stochastic differential equations is available only for a very limited class of models and exact likelihood inference is usually not possible. Also, discretization of the estimators obtained from continuous time analysis is always biased and not useful in practice.

Recently, many authors have considered ways to establish approximate and/or quasi-likelihood inference for stochastic differential equations (for a review see Iacus, 2008). The `sde` package implements those methods in the hope to fill the gap between theoretical results and applied financial econometrics. In particular, the package allow to build several kinds of likelihood functions to be used in a standard R context via the `mle` function.

The `sde` package also implements model selection procedures based on AIC statistics for stochastic differential equations, identification of structural changes in the volatility component of the model and hypotheses testing along with other estimation procedures like estimating functions, the method of the moments, etc. Some tools for nonparametric statistics are also available.

Due to the fact that simulation is part of modern financial analysis, the `sde` package includes the function `sde.sim` which implements several simulation schemes for one dimensional stochastic differential equations, including those presented in the fundamental reference of Kloden and Planten (1999), e.g. Euler's and both Milstein's schemes, as well as several new simulation methods appeared in the last ten years, e.g. Ozaki and Shoji-Ozaki local linearization methods, Berkos et al. Exact Sampling, and Kloden-Platen-Soeresen method.

References

Iacus. S.M. (2008). Simulation and Inference for Stochastic Differential Equations with R Example, Springer, New York.

Jacod, J., Shiryayev, A.N. (1987) Limit Theorems for Stochastic Processes, Springer-Verlag, New York.

Kloden, P., Platen, E. (1999). Numerical Solution of Stochastic Differential Equations, Applied Mathematics, **23**, Third corrected printing, Springer, New York.

Modeling recovery rates of municipal waste using generalized linear models and beta regression

M.V. Ibánez [1,*] **, M. Prades** [2] **, A. Simó** [1]

1. Departamento de Matemáticas. Universitat Jaume I. Spain.
2. INGRES. Departamento de Ingeniería Mecánica y Construcción. Universitat Jaume I. Spain.

* Contact author: mibanez@uji.es

Keywords: GLM, Beta Regression, Waste Management

The economic viability of waste management and environmental impact of these issues are of great interest today in most cities. There are a number of economic, social and cultural factors that determine the characteristics of the waste and the value of design parameters used in the calculations of a collection system (2). The aim of this work is to model the recovery rates of municipal waste in Spanish cities over 50,000 inhabitants. Different regression models to manage continuous proportion data are compared. Two kinds of variables can be considered: demographic variables and those related with the waste collection system. The regression models considered have been: Generalized linear models (3) with Binomial, Poisson and Gamma errors after several transformation on the data and Beta regression (1) on the raw data.

References

Ferrari, S. and Cribari-Neto F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31, 799–815.

Gallardo, A. (2000). *Metodología para el diseño de redes de recogida selectiva de RSU utilizando sistemas de información geográfica. Creación de una base de datos aplicable a España* . Valencia: Ed. UPV.

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models.* 2nd. edn. London: Chapman and Hall.

Binary attributes quantification with external information

Alfonso Iodice D'Enza[1]*

1. University of Cassino

* Contact author: iodicede@unina.it

Binary data bases (DB) characterize several fields: transactional data, web-clickstream data, gene-expression data. Binary data mart are often characterized by two main features: high dimensionality and sparsity. This is the case, for example, of transactional data. In tabular form, a transaction data mart is a binary matrix: customers choices/DB records are stored on rows, products/attributes on columns. Cell values are '1' if a customer buys a product, '0' otherwise. It is then fair considering that the data matrix has a large number of columns,as many as the products available in a store (high dimensionality), and that most of the cells are null (sparsity). Both of these aspects make difficult to identify and describe relations among blocks of columns and blocks of rows. A well known exploratory approach to analyze this structures is frequent pattern mining which identifies groups of highly co-occurring attributes. Although this approach is computationally efficient, it produces huge output that hides away interesting relations and that is generally difficult to interpret. An alternative approach is to quantify the binary attributes on the basis of the underlying association structure using Multiple Correspondence Analysis (Greenacre, 2007). Dimensionality of data is drastically reduced: this aspect eases exploratory purposes such as the identification of groups of records or groups of co-occurring attributes. It also provides a visualization display of the association structure that eases user interpretation. In this paper it is proposed a quantification of binary attributes that takes into account external information coded as a categorical variable. In particular, the proposed quantification enphasize both the co-occurrences among attributes *and* the groups of records defined by the external categorical variable. The reference method is a reformulation of Nonsymmetric Correspondence Analysis (Lauro and D'Ambra, 1984). The general criterion beyond the quantification is to maximize the following quantity

$$\frac{1}{n}\sum_{k=1}^{K}\left(\sum_{j=1}^{p}\frac{f_{kj}^2}{f._j} - \frac{f_{k.}^2}{n}\right)(k=1,\ldots K, j=1\ldots J) \tag{1}$$

where f_{kj} is the number of records that present the kth category of the external variable and the jth binary attribute; $f._j$ is the number of records the jth binary attribute; $f_{k.}$ is the number of records that present the kth category of the external variable; n is the total number of records. Remark that the external categories can be refferred to a previous clustering of the records. This method can also be integrated in a two-step procedure together with a clustering step, as in Palumbo and Iodice D'Enza (2009). The paper illustrates an application of the proposed method to a binary data set representing a group of italian students of the University of Macerata. In particular, binary attributes refer to the exams passed by the students, while external information will refer to socio-demographic characteristics of the students.

References

M. J. Greenacre(2007). *Correspondence Analysis in Practice, second edition.* Chapman and Hall/CR.

N.C. Lauro and L. D'Ambra (1984). L'analyse non symmétrique des correspondances. In E. Diday et al., eds, *Data Analysis and Informatics, III.* North-Holland, 1984.

F.Palumbo and A.Iodice D'Enza (2009). Clustering and Dimensionality Reduction to Discover Interesting Patterns in Binary Data. In *Proceedings of GFKL08.* Hamburg, accepted, in press.

Simultaneous Use Probability of Mobile Internet and Other Media by Multivariate Probit Model

Hisashi Ishida [1] , Fumiyo Kondo [1, *]

1. University of Tsukuba
* Contact author: kondo@sk.tsukuba.ac.jp

Keywords: Bayesian multivariate probit model, Mobile Internet, Amusement services, Customer segmentation by needs, Multiple media use

According to a report of 2008 Japanese government, the percentage of the contracts of mobile phones in Japan has reached 85.6%, which shows a popularity of accessing to information anytime, anywhere by Japanese people. Under these circumstances, marketing vehicle to approach consumers via multiple media such as a combination of TV and mobile Internet is getting very popular in Japan.

The purpose of our research is to find out factors to influence on simultaneous use of mobile Internet with 11 kinds of media especially focusing on "amusement services" and "study information services" that were chosen among 21 kinds of mobile information services. Those services have mostly penetration rates of around 70% except "E-mail with pictures" and there are still possibilities of further market expansion in the near future. A simultaneous use probability of mobile Internet and other media is estimated by using Bayesian Multivariate Probit Model that is suitable for the analysis on individual users' behavioral patterns.

The results provide efficient approaches to the consumers of mobile information services and will contribute to further expansion of mobile information service market. The results of our research have shown which explanatory variables of potential customers have influence on which simultaneous use of mobile Internet with other medium in which direction. By using appropriate pair-wise advertisement vehicles, improvements of accessibility to potential customers can be achieved. Moreover, since it became obvious that customers have a variety of information service needs and diversities of simultaneous use of a pair of media among customers were observed, customer segmentation by needs as well as by simultaneous media uses appears to be necessary.

References

Lieven, M. D. , V. Partrick, and B. Katrien(2007). Adoptor segments, adoption determination and mobile marketing. , *Journal of Targeting, Measurement and Analysis for Marketing ,*Vol.16, pp. 78-95.

Martin, B. S., J. Durme, M. Raulas, and M. Merisavo, *Journal of Advertising Research*, pp. 293-300.

Suzuki, T. (2007). *i-mode & Mobile Multimedia Services on Mobile Phones*. NTT DoCoMo.

Rossi, P. E., G. Allenby, and R. McClloch (2005) . *Bayesian Statistics and Marketing*. Wiley.

Using R in a Corporate Environment

John James[1*], Francisco Gochez[1]

1. Mango Solutions, UK
* Contact author: jjames@mango-solutions.com

Keywords: Validation, Assurance, Remoting, Cloud

R is a very powerful data analysis tool. However, its very flexibility often generates the following problem: different users may produce different results on the same data. This is not acceptable in many cases for work in various commercial environments such as the Pharmaceutical and Finance industries –as well as many others - and therefore adoption of R as a mission critical tool is resisted by CIO's in corporate environments.

On the other hand, R is the often the only practicable tool for the end user. This may be because of the advanced nature of the analysis, the size of the data or the short timescale needed for the analysis to be undertaken. At Mango we have worked with numerous large commercial organizations and have implemented a range of solutions to these problems. In this presentation we will discuss some of these approaches and their various pros and cons, and make suggestions about where the R community could usefully contribute to an even wider acceptance and user of R.

In particular, the following points will be covered:
- Validation of code and lockdown. Previous presentations have commented on the ease manipulating an R environment and quality.
- Auditability and reproducibility of results
- Full access to R by end users. Although R is installed in a controlled environment how do users get the access to the command line that is required?
- Supply of the computing resource needed by R: access to large amounts of data, requiring distributed/cloud/grid computing.

The presentation concludes that R is best installed centrally in Corporate Environments with access provided by various bespoke interfaces.

References

Tony Rossini, David James, "*Open Source Statistical Software in Pharma Development: Case Study with R*", User!2007 (Iowa State University, USA), August 2007
http://user2007.org/program/presentations/rossini.pdf.

Mat Soukup, "Using *R: Perspective of a FDA Statistical RevieweR*", User!2007 (Iowa State University, USA), August 2007
http://user2007.org/program/presentations/soukup.pdf.

Karim Chine, "*Biocep, Towards a Federative, Collaborative, User-Centric, Grid-Enabled and Cloud-Ready Computational Open Platform*," escience,pp.321-322, 2008 Fourth IEEE International Conference on eScience, 2008
http://biocep-distrib.r-forge.r-project.org/.

wiiRemote

Landon Jensen[1,*] , Vatsal Shah[1,2,&]

1. Micron Technology, Inc.
2. Purdue University
* Contact author: lsjensen@micron.com
& Contact author: vshah@purdue.edu

Keywords: Virtual, Immersive, Wii Remote, 3D, sensor

Building upon the creative ideas of Johnny Chung Lee using the Wii Remote as a powerful, exciting, and economical input device, we propose a connection of the Wii Remote with R using R (D)COM. Specifically, using the Wiimote libraries we access data sampled from the Wii Remote's 3-axis accelerometer and infrared sensor to create a novel virtual 3D environment that allows for interesting interaction with multidimensional data.

References:

Lee, Johnny Chung (2008). *Johnny Chung Lee - Projects - Wii*,
http://www.cs.cmu.edu/~johnny/projects/wii/

Brindza, Szweda, Striegel (2009). *WiiMote - Edu – Twiki.*
http://netscale.cse.nd.edu/twiki/bin/view/Edu/WiiMote

Baier, Neuwirth (2006). *Tutorial: Embedding R in Applications on Windows*, User!2006 (Vienna, Austria).

hzAnalyzer: Detection, quantification, and visualization of contiguous homozygosity in human populations from high-density genotyping datasets using R and Java

Todd A. Johnson[1,2,*] , Yoshihito Niimura[2] , Tatsuhiko Tsunoda[1]

1. Laboratory for Medical Informatics, Center for Genomic Medicine, RIKEN Yokohama Institute, Yokohama, Kanagawa-ken, JAPAN
2. Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, JAPAN

* Contact author: tjohnson@src.riken.jp

Keywords: genetics, polymorphisms, homozygosity, linkage disequilibrium

Research since the initiation and completion of the Human Genome Project [4] and its follow-up, the International HapMap Project [1, 2], has shown that much of human genetic variation is non-randomly organized into regions of restricted diversity in which a limited range of haplotypes can be observed. Such non-random partitioning of variation has been especially important for the design and performance of genome-wide association studies (GWAS), in which genetic variation between case and control samples are examined for associations with human diseases. In contrast to such regions where diversity is locally decreased across a population, recent reports have shown that individuals exist even in generally outbred human populations who possess extended regions of homozygosity, in which both large regions or even complete chromosomes apparently carry the same genetic variation on both chromosomes[2, 3]. Taken together, the locally restricted patterns that can be seen across populations and the long homozygous segments that can be seen within single individuals likely represent the extremes of a spectrum of relatedness that can be observed between individuals within human populations[6].

To analyze how the extent of contiguous homozygosity in high density single-nucleotide polymorphism (SNP) datasets varies within and between populations at genome-wide, chromosomal, and locally defined levels, we developed hzAnalyzer, a suite of programs using R and Java. This program uses rJava[5] to integrate Java code within our R programs, the combination of which allowed for a synergy with R contributing its ready-made statistical components, ease of scripting, and quick proto-typing and Java contributing enhanced performance in dealing with large datasets and an object-oriented construction that allowed us to represent some of the complexity inherent in population genetic data such as the fact that our sample population consisted of data subsets such as sample populations, families, and individuals. The functions making up hzAnalyzer can be broken down into three categories: 1) Homozygous segment detection and processing, 2) Quantification of variation in the extent of contiguous homozygosity within individuals and populations at different resolutions (i.e. genome, chromosome, local chromosomal regions), and 3) Visualization of raw segment positions and summarized/aggregated data. Our presentation will provide details about the functions that make up hzAnalyzer as well as figures using real human genotyping data from the International HapMap Project.

References

Altshuler D *et al* (2005). A haplotype map of the human genome. *Nature*, 437**,** 1299-1320.

Frazer KA *et al* (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449**,** 851-861.

Gibson J *et al* (2006). Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet*, 15**,** 789-795.

Lander ES *et al* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409**,** 860-921.

Urbanek S (2008). *rJava: Low-level R to Java interface,* http://www.rforge.net/rJava/.

Weir BS *et al* (2006). Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet*, 7**,** 771-780.

Introducing the R to PowerPoint Package

Wayne R. Jones
Wayne.W.Jones@shell.com

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

Keywords: Automatic Report Generation, PowerPoint, rcom.

In this presentation I will introduce the package 'R2PPT' which automatically generates Microsoft PowerPoint presentations directly from R. The package is essentially a suite of wrapper functions for the 'rcom' package developed by Thomas Baier (Thomas@statconn.com) which serves as a COM (Component Object Model) interface to R.

A demonstration of the major functionality of 'R2PPT' will include:

- Adding different slide types.
- Applying a template design.
- Adding R graphics.
- Displaying R data frames.

It is hoped that this package will prove extremely useful to users of R who produce PowerPoint reports on a daily basis.

An R implementation of bootstrap procedures for mixed models

Jose A. Sanchez-Espigares[1] , **Jordi Ocaña**[2]

1. Department of Statistics and Operational Research. Universitat Politecnica de Catalunya
2. Department of Statistics. Universitat de Barcelona

* Contact author: josep.a.sanchez@upc.edu

Keywords: Mixed Models, Bootstrap, inference

The implementation of mixed models estimation in the $lme4$ package provides a general common framework for linear, generalized and nonlinear mixed models with nested and/or (partially) crossed random effects. This package reimplements the estimation procedures for mixed models from the standard $nlme$ package, Pinheiro and Bates (2000), in a more efficient way. Inference about fixed and variance components parameters can be done by means of MCMC techniques (using the mcmcsamp method).

We present an extension of the package to include methods for generating data according an specified model and fitting it to obtain bootstrap samples of the estimators. Several bootstrap methods can be applied to generate the data: specifying the distribution for the variance components (parametric bootstrap), resampling with replacement any transformation of the random effects/residuals from the fitting process (semi-parametric bootstrap) or using extensions of the empirical distribution (wild bootstrap). Generation of data is performed keeping the design matrices for the fixed and random part of the model. First, the random effects are obtained under one of the above strategies in order to calculate the resampling linear predictor and the corresponding mean for each observation. Next, the resampling response variable is generated according to the conditional distribution considered for the response given the random effects. Trying to keep the general framework for the three kind of models, a bootstrap method based on resampling the quantile residuals, Dunn and Smyth (1996), is proposed to obtain the resampling data in the second step.

Due to the fact that estimation of generalized and nonlinear mixed models are computer intensive procedures, efficient strategies are needed to reduce the computational cost of these bootstrap methods. This work presents the features of the routines implemented and evaluates several options in comparison of Bayesian $lme4$ approach and other R packages alternatives.

References

Pinheiro, J. C., and Bates, D. M. (2000). *Mixed-effects models in S and S-plus.* New York: Springer-Verlag.

Dunn, K. P., and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5, 236-244.

The `CalciOMatic` package: a new tool for calcium imaging quantitative analysis

Sébastien Joucla[1,*], Andreas Pippow[2], Peter Kloppenburg[2], Christophe Pouzat[1]

1. Cerebral Physiology Laboratory, Université Paris-Descartes, CNRS, UMR 8118, 45 rue des Saints-Pères, 75006 Paris, France

2. Institute of Zoology and Physiology, Center for Molecular Medicine Cologne (CMMC), Cologne Excellence Cluster in Aging Associated Diseases (CECAD), University of Cologne, Weyertal 119, 50931 Cologne, Germany

* Contact author: sebastien.joucla@parisdescartes.fr

Keywords: calcium imaging, ratiometric measurements, parametric modeling, nonlinear regression

Measuring variations of intracellular free calcium concentration through the fluorescence changes of a calcium sensitive dye is an ubiquitous technique in Neuroscience. Despite its popularity confidence intervals on the estimated parameters of calcium dynamics models are never given. To address this issue, we have developed a model for ratiometric measurements obtained with a CCD camera.

We built a 2-stage model whose first element links the fluorescence intensity to the calcium dynamics and whose second element describes fluorescence measurements through a photon counting process. At each time sample and for both wavelengths, the photon count read out by the camera can be described as a realization of a Poisson random variable. In experimental situations encountered in practice, this distribution can be approximated by a Gaussian distribution with variance equal to the mean.

Using Monte-Carlo simulations, we first show that using the classical *ratiometric* transformation to fit calcium signals does not yield reliable confidence intervals on the fitted calcium dynamics parameters. This is due to the heteroscedasticity of the signal. We then introduce a *direct* approach, based on the *square-root* transformation of the original fluorescence signals. This transformation stabilizes the signal variance and leads us back to a standard nonlinear regression setting. Our *direct* approach has many advantages over the *ratiometric* approach:

1. The construction of confidence intervals is reliable, for both the calcium dynamics parameters and the experiment-specific ones (such as the background fluorescence at each wavelength).

2. Using approaches inspired by constrained linear regression, we can take into account the finite precision on calibrated parameters (such as the dye dissociation constant in the cell).

3. It is also possible to estimate the variations of the dye concentration during the experiment.

All these features will be illustrated on simulated data using the Monte-Carlo approach. Moreover, we show on experimental data that using the last two features leads to major improvements in the goodness of fit. These improvements are characterized with classical diagnostic plots of the `nls` function. Finally, the *direct* method allows us to formally decide between several nested models of the calcium decays.

The *direct* method was implemented in `R`, all pieces of codes being gathered in the `CalciOMatic` package, which will be submitted to `CRAN`. This package includes easy-to-use functions to simulate data, fit either simulated or experimental data, and plot results as well as diagnostic plots.

References

Joucla S, Pippow A, Kloppenburg P and Pouzat C. *Getting more out of ratiometric calcium measurements with an explicit data generation model.* Program No. 497.18. 2008 Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience, 2008. Online.

A Generalized Motif Bicluster Algorithm

Sebastian Kaiser[1,*], Friedrich Leisch[1]

1. Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstrasse 33, 80539 München, Germany.

* Contact author: Sebastian.Kaiser@stat.uni-muenchen.de

Keywords: Biclustering, Two Way Clustering, Ordinal Data

In many application domains different clusters in data may be defined by different sets of variables. E.g., in maketing one group of consumers could mainly be concerned about price and technical features of a product, while others care most about design and how "cool" the product is (almost regardless of the price). Standard clustering algorithms use all variables for all clusters and hence may fail to detect such structures in the data. Biclustering is the simultaneous clustering of columns and rows in a data set: each cluster is defined by a different subset of variables, these subsets can of course be overlapping. R package `biclust` (Kaiser & Leisch 2008, Kaiser et al 2008) contains a comprehensive collection of bicluster algorithms, preprocessing methods, and validation and visualization techniques for bicluster results.

The main focus of this presentation will be on recent additions to the package: There are new functions for bicluster validation and comparison. A new generalization of the well-known motif bicluster algorithm has been developed which is particularly suited for biclustering of marketing survey data. While the standard motif algorithm only searches for constant entries in the data matrix, our generalization is better suited for ordinal and metric data. The user can specify "neighborhood patterns" like intervals or density kernels of pre-specified size for metric data. In addition to finding more general patterns than constant groups only this also allows to calculate a posterior probabilities of cluster membership and can be seen as a first step towards fully model-based biclustering. All new methods will be demonstrated using real data from marketing applications.

References

Sebastian Kaiser and Friedrich Leisch (2008). A toolbox for bicluster analysis in R. In Paula Brito, editor, *Compstat 2008–Proceedings in Computational Statistics*, pages 201-208. Physica Verlag, Heidelberg, Germany.

Sebastian Kaiser, Rodrigo Santamaria, Roberto Theron, Luis Quintales and Friedrich Leisch (2008). biclust: BiCluster Algorithms. `http://cran.R-project.org/package=biclust`.

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

Shape analysis in R: GM library in the light of recent methodological developments

Stanislav Katina[1,2,*]

1. Department of Anthropology, University of Vienna, Vienna, Austria
2. Department of Applied Mathematics and Statistics, Comenius University, Bratislava, Slovakia

* Contact author: stanislav.katina@univie.ac.at

Keywords: shape and size analysis, GM library

Introduction. Lately, geometric morphometrics (GM) has focused considerable methodological ingenuity on data from landmarks, curves and surfaces in-between landmark points. Here we discuss not the algebra of these so-called (semi)landmark methods, an algebra that is nearing consensus at the present time, but rather their scientific yield for the kinds of questions (paleo)anthropologists ask. Some of typical investigative designs ask questions about size and shape in a context of species differences and sexual dimorphism, or questions about integration and modularity. Recently few image/statistical softwares have been developed to help to answer some of these questions but researchers have to combine them to get the answers they are searching for. Some of the methods are not covered yet.

2D example. We recruited 20 young women (aged between 19 and 31) who reported to have a regular menstrual cycle and did not take any hormonal contraceptives (Oberzaucher et al., 2008). We took standardized facial photographs neutral expression, eye height, facial ornaments and hair removed from the face, no make-up, 5m distance from the camera, evenly lighted daily for 30 days. In a forced choice task, 50 male and 50 female subjects were presented with two photographs of each participant one taken in the ovulatory and one taken in the luteal phase. The task was to pick out the more attractive, healthy, sexy, and likeable, of the two. We cut skin patches sized 150×150 pixels from the cheek and subjected them to the same forced choice task with slightly modified adjectives. We measured the facial photographs by setting 72 anthropological landmarks and semilandmarks. We analysed the texture of the skin patches calculating co-occurrence parameters, such as homogeneity, energy, entropy, contrast and correlation. The colour information was calculated in an RGB-space in terms of hue, saturation and intensity.

3D example. Our example re-uses part of a Vienna data set of 372 skulls from various collections (include Zoological Dpt. of the Natural History Museum, Vienna, Austria; Dpt. of Anthropology, University of Zurich-Irchel, Switzerland; Royal Museum of Central Africa, Tervuren, Belgium; Dpt. of Anatomy and Human Genetics, Frankfurt/Main, Germany): data from 32 landmark points and 7 ridge curves totalling 161 semilandmarks. The landmark points on both sides of every cranium and 161 semilandmarks on the left side of every cranium were digitalized using a MicroScribe 3DX (Mitteroecker et al., 2004). The right side semilandmarks were calculated by TPS based on the set of all landmarks and left side semilandmarks.

Conclusion. Up-coming GM library includes the methods as Generalized Procrustes Analysis, affine and non-affine component, unwarping, missing value estimation, Multivariate Multiple Linear Regression Model of shape on size, Relative Warp Analysis, shape-space PCA, form-space PCA, size-adjusted PCA, 2-block PLS (two shape blocks, one shape block and one block of external variables), sliding of semilandmarks on open and closed curves and surfaces, analysis of asymmetry, statistical inference (Katina, 2008). The GM library has practical implications for (paleo)anthropologist and also researchers from the other fields.

References

Mitteroecker P., Gunz P., Bernhard M., Schaefer K., Bookstein F.L., (2004). Comparison of cranial ontogenetic trajectories among great apes and humans. *Journal of Human Evolution.* **46** (**6**): 679–697

Oberzaucher E., Blantar I., Schmehl S., Holzleitner I., Katina S., Grammer K. (2008). The myth of hidden ovulation how the face changes during the menstrual cycle. Book of Abstracts. XIX Biennial Conference of the International Society for Human Ethology, Italy, University of Bologna, 13th to 17th July 2008, 174–175, `http://www.ishe08.org/`.

Katina S. (2008). Geometrical aspects of statistical size and shape analysis – Multivariate statistical models and statistical inference. *Habilitation Thesis.* Comenius University. 87 p.

Application of R for classification of main tree species using terrestrial laser scanner data

Hans-Joachim Klemmt[1,*]

1. Technische Universität München
2. Chair of Forest Growth and Yield, Am Hochanger 13, 85354 Freising (Germany)
* Contact author: h-j.klemmt@alrz.tum.de

Keywords: forest, tree species, terrestrial laser scanner

Terrestrial laser scanning is becoming more and more popular for forest mensuration purposes. Meanwhile many groups concerned with this topic across the whole world have developed algorithms and software solutions to derive dimensional aspects of trees more or less automatically (Wezyk, 2007; Maas et al., 2008). For real forest mensurational applications, e. g. to apply terrestrial laser scanning technology for forest inventories, a lack of knowledge exists. So far no performant solution, which can distinguish the tree species automatically from point cloud data, exists.

At the chair of Forest Growth and Yield at Technische Universität München a methodology has been developed to distinguish the tree species by using terrestrial laser scanner data. This system consists of a training component, which trains classification algorithms using forest inventory data of the last inventory period. The classifying algorithms use for example tree bark metrics or color distribution metrics as variables for tree species distinction. After an evaluation of the goodness of the trained classifiers by cross-validation these classifiers are applied to the new forest inventory data. At the end of the process each automatically detected and located tree gets assigned a tree species mark (Witten and Frank, 2005).

For the application of the "lingua franca" R for this problem two reasons are responsible. The first reason is, that the scientific working group on the application of terrestrial laser scanning on forests has made good experiences with the already in R developed routines for automated derivation of dimensional measures of trees (Klemmt, 2008). The second reason is the ever growing number of R packages which provide for the mentioned problem e. g. image processing routines as well as several classification routines.

References

Klemmt (2008): *Using R as an environment for automatic extraction of forest growth parameters from terrestrial laser scanner data* , User!2008 (Dortmund, Germany), July 2008, http://www.statistik.uni-dortmund.de/useR-2008/abstracts/Klemmt.pdf

Maas, H-G., Bienert, A., Scheller, S., Keane, E. (2008): Automatic forest inventory parameter estimation from terrestrial laser scanner data. International Journal of Remote Sensing, Volume 29, Issue 5 March 2008, pp. 1579-1593.

Wezyk, P., Koziol, K., Glista, M., Pierzchalski, M. (2007): Terrestrial laser scanning versus traditional forest inventory. First results from the Polish forests. ISPRS workshop on laserscanning 2007 and SilviLaser 2007, Espoo, Sept. 12-14, 2007, Finland, pp. 424-429.

Witten, I.A., Frank, E. (2005): Data Mining – Practical Machine Learning tools and techniques. Elsevier, 524 p.

rSMILE, an interface to the Bayesian Network package GeNIe/SMILE

Christoph M. Friedrich[1,*] and Roman Klinger[1]

1. Fraunhofer Institute for Algorithms and Scientific Computing (SCAI); Department of Bioinformatics; Schloss Birlinghoven; D-53754 Sankt Augustin; Germany
* Contact author: friedrich@scai.fraunhofer.de

Keywords: Bayesian Networks, Graphical Models, Interface

Bayesian Networks are well known directed probabilistic graphical models. There are many implementations of graphical models in R and a summary of existing packages is given in [1]. The

learning effort to master these implementations is typically high. The combination GeNIe (Graphical Network Interface) [2,3] and SMILE (Structural Modeling, Inference, and Learning Engine) provides an easy way to develop and diagnose Bayesian Networks with categorical variables and allow the inclusion into other applications with the inference engine provided.

Unfortunately, the implementation does not include evaluation possibilities like cross-validation, bootstrapping or ROC analysis. Users have to implement this externally using the SMILE interface. To circumvent this problem, rSMILE, an interface using the Rjava bridge [4] and jSMILE the java implementation of SMILE has been developed. rSMILE allows for training the structure and the conditional probability tables as well as inference in R. Existing models can be loaded and new models saved in the GeNIe format. The networks can be inspected using bar chart depictions and graph-layout algorithms like the spring-embedder method. During structural learning with the Greedy Thick Thinning algorithm background-knowledge can be included by enforcing or forbidding edges of the network.

The interface has been successfully applied in the course of the European integrated project @neurIST where rSMILE has been used to develop risk and treatment outcome models for intracranial aneurysms [5].

References

1. Højsgaard, S. *Graphical models in R* webpage. Last accessed 2009-02-20
   http://www.ci.tuwien.ac.at/gR/
2. Druzdel, MJ. GeNIe *: A development environment for graphical decision-analytic models*. In : Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association (AMIA). **1999**. Pages 1206
3. Druzdzel, MJ. *GeNIe/SMILE* webpage. Last accessed 2009-02-20
   http://genie.sis.pitt.edu
4. Urbanek, S. *The Rjava interface*, webpage. Last accessed 2009-02-20
   http://www.rforge.net/rJava/
5. Friedrich, C. M.; Hofmann-Apitius, M.; Dunlop, R.; Chronakis, I.; Sturkenboom, M. C. J. M.; Risselada, R.; Oliva, B., and Sanz, F. *Initial Results on Knowledge Discovery and Decision Support for Intracranial Aneurysms* Proceedings of HEALTHINF 2008 Conference, INSTICC*, **2008**, 265-272

Visualising a web site with tag clouds generated by R

Sigbert Klinke[1,2,*]

1. Humboldt-Universität zu Berlin, School of Business and Economics, Institute of Statistics and Econometrics, Spandauer Strasse 1, D-10718 Berlin, Germany
2. Johannes Gutenberg University Mainz, Dept. of Law and Economics, Chair of Business and Human Ressource Education, Jakob-Welder-Weg 9, D-55099 Mainz, Germany
* Contact author: sigbert@wiwi.hu-berlin.de

Keywords: `igraph`, network, page rank, visualisation, Wikipedia

The Wikipedia is the first source for a lot of users to gather information about a specific topic. To get an overview about a topic the user needs to follow a number of links to various pages in the Wikipedia. To visualise the link structure between pages, outbound **and** inbound, would help the users to cover a topic more easily.

The Wikipedia itself allows the categorisation of pages. Each page may belong to at least one category which reflects the topic and classes that are directly related to the subject of the page (Wikipedia, 2009). For example, the article about *Student's t-test* belongs to the categories *Statistical tests, Statistical methods* and *Parametric statistics*. It is possible to build hierarchies of categories, for example all three categories are part of the category *Statistics*.

In the German Wikipedia, the category *Statistics* consists of approximately 500 pages and only 14 sub-categories, in the English Wikipedia the category *Statistics* consists of 8 pages and 54 sub-categories. It is obvious that the categories, as hand-made by user, may not provide an easy way to get an overview about a topic.

Search engines, such as Google, use, amongst other things, the link structure between pages to measure the importance and the closeness of pages. Based on all the links between pages (unidirectional: inbound, outbound and bidirectional) in one category, we generate a distance matrix for the pages. Using multidimensional metric scaling we determine the position of the page and its direct neighbours in a two-dimensional space. The page rank (Page and Brin 1998) of each page gives us the importance of each page. The R package `igraph` (Csardi 2008) supports the generation of the network (page positions and page importance).

For each page in the German Wikipedia, in the category *Statistics* a tag cloud with the page names will be generated. The position of the page names are determined by the multidimensional scaling and the font size by the page rank.

Ähnlichkeitsmaß Distanzmaß	**Kategorie:Statistik** **Maßtheorie**	Wahrscheinlichkeitsrechnung Gegenwahrscheinlichkeit
Wahrscheinlichkeitsfunktion Satz von Cramer-Wold	Maßraum Fast-Überall	Wahrscheinlichkeitstheorie

Figure 1: Tagcloud for "Maßtheorie". Note that only links to pages which belong to the category *Statistics* are included in the tag cloud although many more pages link to and from the page "Maßtheorie".

References

Wikipedia (2009). *Wikipedia:Categorization – Wikipedia, The Free Encyclopedia (Online; accessed 26-Feb-09),* http://en.wikipedia.org/wiki/Wikipedia:Categorization

Page, L. and Brin, S. (1998). The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the seventh international conference on World Wide Web*, 7:107-117

Csardi, G. (2008*). Igraph: Routines for simple graphs, network analysis (Online; accessed 26-Feb-09),* http://cran.r-project.org/web/packages/igraph

Factor Analysis for Multiple Testing (FAMT) : an R package for large-scale significance testing under dependence

Maela Kloareg, Chloé Friguet, David Causeur

Agrocampus Ouest, Applied mathematics department, 65 rue de Saint-Brieuc, 35000 Rennes, France
* Contact author: maela.kloareg@agrocampus-ouest.fr

Keywords: Factor analysis, Multiple testing, False discovery rate, Dependence.

The method proposed in this package takes into account the impact of dependence on the multiple testing procedures for high-throughput data. The common information shared by all the variables is modelled by a factor analysis structure, as proposed by Friguet *et al.* (2009). New test statistics for general linear contrasts are deduced, taking advantage of the common factor structure to reduce correlation and consequently the variance of error rates. Thus, the False Discovery Proportion is controlled, which is not the case when classical method are used (see for instance Gordon *et al.*, 2007). Moreover, the FAMT method increases the global power, regarding the Non Discovery Rate. In this presentation, the methodology will be applied on genomic data, and compared to other competitive methods, such as the Optimal Discovery Procedure (Storey, 2007).

References

Friguet C., Kloareg M. and Causeur D. (2009). A factor model approach to multiple testing under dependence. In press

Gordon, A., Glazko, G., Qiu, X. and Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni, and stability of multiple testing. *Ann. Appl. Statist.* 1 (1), 179-190.

Storey, J.D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 69 347-368.

Maximum Likelihood Conjoint Measurement in **R**

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

1. Inserm, U846, Stem Cell and Brain Research Institute, Dept. Integrative Neurosciences, 69675 Bron cedex, France
2. Department of Psychology, Center for Neural Science, New York University, New York, NY 10003, USA

* Contact author: ken.knoblauch@inserm.fr

Keywords: conjoint measurement, scaling, glm

Conjoint measurement is a psychophysical method that allows the assessment of separate contributions of two (or more) attributes (dimensions, factors) to *perceived* differences in stimuli (Luce & Tukey, 1964).We present a parametric model of difference judgments and statistical methods that allow maximum likelihood estimation (MLE) of the relevant parameters as well as testing of hypotheses concerning the parameters. We describe the model and its implementation in R using glm and show how to use it to determine the separate contributions of surface irregularity (bumpiness) and surface gloss to perceived bumpiness and glossiness.

The stimuli were computer-rendered surfaces that varied in physical glossiness and physical bumpiness (Figure 1, left). In one condition, observers viewed every possible pair of the 25 surfaces in Figure 1 (left) and judged which one was bumpier. In a second condition, a different group of observers judged which one was glossier. Ho *et al.* (2008) developed a model of the judgment process by assuming that physical gloss level g_i and bump level b_j of surface S_{ij} contribute to perceived bumpiness and perceived glossiness after scaling by unknown functions $B^g(\cdot), B^b(\cdot)$ and $G^g(\cdot), G^b(\cdot)$, respectively.

Figure 1

Perceived bumpiness B_{ij}^A for the *Additive Model* is modeled as the sum of contributions (cues) from physical bumpiness $B^b(b_j)$ and physical gloss $B^g(g_i)$,

$$B_{ij}^A = B^g(g_i) + B^b(b_j) = B_i^g + B_j^b,$$

with a parallel formulation for perceived glossiness. In comparing the bumpiness of surfaces S_{ij} and S_{kl}, we assume that the observer forms the noise-contaminated decision variable, $\Delta = B_{ij}^A - B_{kl}^A + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$ and judges surface S_{ij} as bumpier precisely when $\Delta > 0$. The parameter σ represents the observer's precision in judgment. We estimate σ and the remaining free parameters $B_2^g, \ldots, B_5^g$ and $B_2^b, \ldots, B_5^b$ using MLE. We fit a similar model to gloss comparisons. The MLE estimates are easily obtained using the function glm with a binomial family. The decision variable serves as the linear predictor which is related to the observer's judgments via a probit link. Results from two observers are shown in Figure 1 right . In the additive model, we can test whether the two surface properties influence one another (if not, then B_i^g (and G_j^b) should equal zero for all i, j). We also tested this simple additive model against more complex, non-additive models. We discuss the bias, variability and robustness of the method as well as methods for testing the underlying assumptions of the model (Luce & Tukey, 1964).

References

Ho, Y.-H., Landy. M. S. & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science*, 19, 196–204. New York: Academic Press.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement. *Journal of Mathematical Psychology*, 1, 1–27.

Tree Algorithms in Data Mining: Comparison of R-rpart and Rweka

Ahmet KOÇYİĞİT*

* Contact author: ahmetkocyigit2008@gmail.com

Keywords: rpart, RWeka, benchmarking, decision , tree

The subject that I would like to present is "Tree Algorithms in Data Mining: Comparison of R-rpart and Rweka". This topic is based on the benchmarking of two different decision tree algorithm packages for R-Project. These "decision tree algorithms" are frequently used methods in Data Mining for Statistics.

This methods are used as a predictive model which uses information about a subject to be able to find possible conclusions for that subject's goal.

I'm currently working on this topic as a thesis project and using R-Project with decision tree algorithm packages rpart and RWeka. I would like to join useR!-2009 as a graduate student from Computer Science Department from Türkiye Bilgi Universitesi.

And Finally my comparison will be based on efficiency, usability, performance.

References

Terry M Therneau and Beth Atkinson. R port by Brian Ripley (2007). rpart: Recursive Partitioning, http://cran.r-project.org/web/packages/rpart/index.html .

Kurt Hornik (2007). RWeka: R/Weka interface. http://cran.r-project.org/web/packages/RWeka/index.html

R Package RobLoxBioC: Infinitesimally robust estimators for preprocessing gene expression data

Matthias Kohl[1]

1. Department of Mathematics, University of Bayreuth. Contact: Matthias.Kohl@uni-bayreuth.de

Keywords: gene expression, preprocessing, infinitesimal robustness, radius-minimax estimator

The preprocessing of gene expression data for several platforms routinely includes the aggregation of multiple raw signal intensities to a single expression value. Examples are the computation of a single expression measure based on the perfect match (PM) and miss match (MM) probes in case of the Affymetrix technology, the summarization of bead level values to bead summary values in case of the Illumina technology, or the aggregation of replicated measurements in case of other technologies including real-time quantitative polymerase chain reaction (RT-qPCR) platforms.

Our new package `RobLoxBioC` provides a way to use infinitesimally robust estimators (cf. Rieder (1994), Kohl (2005)) for this purpose. More precisely, we assume normal location and scale and envelop this (ideal) model with an infinitesimally (i.e., shrinking) contamination neighorhood (Tukey's gross error model) where the exact size/radius of the neighorhood is unknown. The optimally robust radius-minimax (rmx) estimators for this setup, minimizing the relative asymptotic minimax MSE for some given radius interval, can be read off from Rieder et al. (2008) and are implemented in our new package `RobLoxBioC`.

In case of Affymetrix data we implemented an algorithm which is similar to MAS 5.0 (cf. Affymetrix, Inc. (2002)). The main difference is the substitution of the Tukey one-step estimator by an rmx k-step ($k \geq 1$) estimator. The rmx estimators can also be applied to Illumina bead level data as well as to data from other platforms or other omics disciplines (e.g., Proteomics or Metabolomics) incorporating replicated measurements.

We will give some comparisons between the results obtained for our rmx estimators and estimators implemented in Bioconductor (cf. Gentleman et al. (2004)) using datasets from literature.

References

Affymetrix, Inc. (2002). *S*tatistical Algorithms Description Document. Affymetrix, Santa Clara.

R. Gentleman, V. J. Carey, D. M. Bates, B.Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *G*enome Biology, 5:R80.

M. Kohl (2005). *N*umerical Contributions to the Asymptotic Theory of Robustness. PhD-thesis, University of Bayreuth, Bayreuth, `http://www.stamats.de/ThesisMKohl.pdf`.

H. Rieder (1994). *R*obust Asymptotic Statistics. Springer, New York.

H. Rieder, M. Kohl and P. Ruckdeschel (2008). The Costs of not Knowing the Radius. *S*tat. Meth. & Appl., 17(1): 13–40.

Psychometrics in R: Rasch Model and beyond

Christophe Lalanne[1,2,*]**, Martin Duracinsky**[2,3]**, Laurence Vaivre-Douret**[1,4]**, Olivier Chassany**[2]

1. INSERM U669, Univ Paris Sud and Univ Paris Descartes, UMR-SO669, Paris, France
2. Hôpital Saint-Louis, Department of Clinical Research, Paris, France
3. Hôpital Bicêtre, Internal Medicine and Infectious Diseases Department, Paris, France
4. Univ Paris 10, Nanterre, France
* Contact author: ch.lalanne@gmail.com

Keywords: psychometrics, item-response models, Rasch, HRQoL

The analysis of a set of items responses collected on a sample of individuals usually relies on classical test theory and item response theory (IRT), whereby test or composite scores are modelled under either a linear or generalized linear model. The well-known Rasch Model shows how it is possible to introduce probability considerations into subjects' responses, and statistical analysis focuses on several form of validity and reliability of test scores. Furthermore, it can be shown that most of IRT models can be expressed as mixed-effects models (e.g. De Boeck and Wilson, 2004), which facilitates exploration of complex effects like Differential Item Functionning (DIF).

Though several compiled packages have been made available for psychometrics in educational assessment or biomedical investigation, only recently have languages such as SAS and Stata incorporated IRT modelling capabilities (e.g. Rabe-Hesketh and Skrondal, 2008; Hardouin and Mesbah, 2007, but see the Free IRT Project, `http://freeirt.anaqol.org/`, for an overview). The open source R software also provides a well integrated statistical framework for psychometrics, ranging from classical analysis to modern techniques based on IRT and derived methods (but see *Journal of Statistical Software*, volume 20, 2007). For instance, two packages (`eRm` and `ltm`) available on CRAN website allow to estimate IRT models under conditional or marginal likelihood approach.

We show how R core functionnalities may be used in applied biomedical research, in particular for the analysis of Patient Reported Outcomes (PRO). It is now well recognized that PRO and Health related Quality of Life have to be taken into account in the evaluation of therapeutic strategies. Following standard guidelines, a stepwise analytic approach must demonstrate that a given HRQoL questionnaire has all of the desirable characteristics of a valid and reliable measurement instrument. After having analyzed the inter-items correlation matrix and responses distribution, a factorial analysis of polychoric or linear correlation matrix aims at determining a minimal set of underlying latent factors which explains the maximum of scores variance. Scores gathered on other questionnaires allow to study convergent and discriminant validities, together with Multi-Trait scaling analysis. Finally, studying composite scores in relation to biomedical indicators help to highlight sensibility or responsiveness of the questionnaire to patients' relative disease. Likewise, scores on unidimensional domains, like physical or mental states, may vary according to cultural-based factors, whether it be DIF or not, and this can be studied using explanatory IRT models.

In summary, R now offers a flexible and reliable way to carry out Exploratory Factor Analysis, Multi-Trait Scaling, Reliability assessment and IRT modelling. Obviously, this avoids the need to switch between different dedicated software and facilitates a seemly integration of statistical analysis and project management.

References

De Boeck, P. and Wilson, M. (2004). *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach.* Springer.

Hardouin, J.-B. and Mesbah, M. (2007). The SAS macro-program `\%AnaQol` to estimate the parameters of IRT models. *Communications in Statistics: Simulation and Computation, 36(2)*, 437–453.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata*, 2nd ed. Stat Press

Institutional Change on a Network

Jennifer M. Larson[1,*]

1. Ph. D. Candidate, Harvard University

* Contact author: jmlarson@fas.harvard.edu

A lingering puzzle in the study of political and economic development is why institutions form successfully in some places and not in others. Institutions desinged to regulate common pool resources (fisheries, forests, irrigation systems, watersheds, etc.) are particularly puzzling. Even if resources are scarce, each individual has an incentive to overuse the resource; even if all others agree to regulate their own use of the resource, an individual has an incentive to deviate from the agreement and continue to overuse. Despite the clear prediction of collective action problems, some groups are able to create and sustain regulatory institutions. Other groups are not as successful. Certain geographic and demographic variables correlate with successful institutions, and some assumptions imposed on collective action games result in equilibria that entail institutions (1; 2). Nonetheless, we are still left without a compelling explanation for the *emergence* of institutions. I use a network approach to explicitly account for the underlying information structure of a population. Some individuals have more information than others; some sources of information are more valuable than others. Since regulatory institutions are more valuable as the number of people willing to submit to the institutions increases, the spread of willingness to adopt an institution is analogous to the diffusion of communication technology. I use simulation techniques in R to characterize the relationship between network variables (size, degree distribution, shape) and the emergence of institutions.

References

[1] E. Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge University Press, 1990.

[2] _____, *Understanding institutional diversity*, Princeton University Press, 2005.

Exploratory interactive tools for spatial data analysis

Thibault Laurent[1,*] , **Anne Ruiz-Gazen**[1], **Christine Thomas-Agnan**[1]

1. Toulouse School of Economics (GREMAQ)

* Contact author : thibault.laurent@univ-tlse1.fr

Keywords: spatial exploratory data analysis, spatial econometrics and statistics, interactive graphics, brushing and linking, innovations of GeoXp.

At User!2006, we introduced GeoXp, an R package implementing interactive graphics for exploratory spatial data analysis. Besides elementary plots like boxplots, histograms or simple scatterplots, GeoXp also couples maps with Moran scatterplots, variogram clouds, Lorenz curves, etc. In order to make the most of the multidimensionality of the data, GeoXp includes dimension reduction techniques such as principal components analysis and cluster analysis whose results are also linked to the map. We intend to present now the innovations of GeoXp. We describe the interactive analysis of a neighborhood structure given by a spatial weight matrix (created with package spdep) and the detection of outliers analyzing the relationship between pairwise Euclidean and pairwise Mahalanobis distances (calculated with package mvoutliers). We use a data basis concerning public schools of the French Midi-Pyrénées region to illustrate the use of these exploratory techniques based on the coupling between a statistical graph and a map.

References

Filzmoser P., Reimann C., Ruiz-Gazen A., Thomas-Agnan C. (2009). Exploratory tools for spatial multivariate outliers detection, preprint 2009.

Laurent T., Ruiz-Gazen A., Thomas-Agnan C. (2006). GeoXp : an R package for interactive exploratory spatial data analysis, User!2006 (Vienna, Austria), June 2006.

High throughput flow cytometry analysis with Bioconductor

Nolwenn LeMeur[1,2,*], Florian Hahne[1], Deepayan Sarkar[1], Byron Ellis[3], Josef Spidlen[4], Ryan R. Brinkman[4], Robert Gentleman[1]

1. Life Sciences Department, Computational Biology Program, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, M2-B876, PO Box 19024, Seattle, Washington 98109-1024, USA
2. EA SeRAIC INSERM, IRISA - Symbiose, Campus Beaulieu, Université de Rennes I, 35042 Rennes Cedex, France
3. AdBrite Inc., 731 Market St., 5th Floor, San Francisco, California 94103, USA
4. Terry Fox Laboratory, British Columbia Cancer Agency Research Centre,675 West 10th Avenue, Vancouver, BC V5Z 1L3 Canada, Canada
* Contact author: nlemeur@irisa.fr

Keywords: flow cytometry, high throughput, data structure, data analysis

Recent advances in automation technologies have enabled the use of flow cytometry high content screening (FH-HCS), in both basic and clinical research, generating large complex data sets with many covariates. However, data management and data analysis methods have not yet progressed sufficiently far from the initial small-scale studies to support modeling in the presence of multiple covariates.

To those aims, we developed a set of computational tools in the R package flowCore to facilitate the analysis of these complex data. We propose R data structures to handle flow cytometry data through the main steps of importing, storing, assessing and preprocessing data from flow cytometry experiments. For example, this package provides facilities for compensation, transformation and filtering preprocessing steps. A key component of the flowCore package is to have suitable data structures that support the application of similar operations to a collection of samples or a clinical cohort. In addition, our software constitutes a shared and extensible research platform that enables collaboration between bioinformaticians, computer scientists, statisticians and biologists.

The software has been used in the analysis of various data sets and its data structures have proven to be highly efficient in capturing and organizing the analytic work flow. Finally, a number of additional Bioconductor packages successfully build on the infrastructure provided by flowCore, offers new opportunities for flow data analysis.

Hierarchical Clustering based on Principal Components

Guillaume Le Ray[1], Quentin Molto [1], François Husson[2]

1. Student from Agrocampus Ouest Majored in Applied Statistics.
2. Professor of Statistics from Agrocampus Ouest.

Keywords: Hierarchical clustering, Factorial Analysis, program, FactoMineR

The HCPC program for Hierarchical Clustering on Principal Components is dedicated to the Clustering especially after a Factorial Analysis. Factorial Analysis and clustering are complementary tools to explore data. Factorial Analysis enables to remove the last components, which means remove the noise and get the clustering more robust. The program structure is the following one: a Factorial Analysis to preserve the signal and delete the noise in the data, a Hierarchical Clustering, a choice of the number of clusters, a partition in clusters, a consolidation by K-means and a description of the clusters.

There is in a wide range of options, uses, results and graphical representations. It includes all the option of the function *agnes* (pachage *cluster*). As it is used in the Factorial Analysis the default distance is the *Euclidean* distance and the aggregation criterion for the Hierarchical clustering is the *Ward* indice. The program suggests a level of cutting calculated on the inertia gains between two partitions. You can choose either to follow this suggestion or to cut the tree on another level by clicking on the tree. The consolidation with *Kmeans* is optional to get more robust clusters. It returns a more optimal partition from the point of view of the inertia criterion. At this process, it adds interpretation tools like the descriptions of the clusters with *catdes* by the variables, the axes and description of the clusters by the individuals (the closest from the cluster barycentres and the most specific of each cluster).

The principal benefit of this function is the multiplicity of graphical representations. It is possible to plot the tree, factor maps with individuals colored by clusters, or both together, in 2D or 3D, where the tree is above the map. This function will be included in the *FactoMineR* package.

References

Jérôme Pagès, Brigitte Escofier(2008). Analyse Factorielles Simples et Multiples 4ème édition. p 31-63

Package for Deciding the Number of Factors in Exploratory Factor Analysis

Yen Lee

National Chengchi University
Contact author: 96752008@nccu.edu.tw

Keywords: Factor Analysis, Factor Numbers, Minimum Average Partial Test, Permutation Analysis, Bootstrapping

Exploratory factor analysis is one of the most widely used methods in psychology. The most crucial procedure in this technique is determining the number of factors to retain. A numbers of rules have been proposed. Among them, Kaiser's lower bound (eigenvalue-greater-than-one, Kaiser, 1960) is most widely used. Horn (1965) proposed parallel analysis to modify Kaiser's rule in taking variation of eigenvalues duo to sampling. Eigenvalues of real data are compared with eigenvalues from simulated normal random variables in ordinary parallel analysis. However, the normality assumption might be improper. Permutation analysis (Buja & Eyuboglu, 1992) which compares data with identical marginal distribution seems more proper for non-normal data. In addition, Lambert, Wildt and Durand (1991) suggested using nonparametric bootstrapping to decide the number of factors. Alternatively, Velicer (1976) proposed that a minimum average partial test which employs a matrix of partial correlations be considered. Procedures above consider different aspects in deciding factor number, so.Gorsuch (1983) had suggested using more than one way to decide the number of factors.

However, popular statistical software packages do not have all the procedures described above. Moreover, there are even no any code of permutation analysis and nonparametric bootstrapping in literatures, so I would like to present a package which implements all procedures. With the impressive graphic ability of R, the package shows the plot of eigenvalues of data, lower bound, parallel analysis, and permutation analysis and the plot of the confidence intervals of eigenvalues. Besides, the package would suggest the number of factors according different rules. Users would only need to read the raw data or correlation matrix into the package, number of factors and plots of eigenvalues through 5 different ways will be obtained easily and comprehensively at the same time.

References

Buja, A. & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate behavioral research*, 27, 509-540.

Gorsuch, R. L. (1983). *Factor analysis*, Philadelphia, PA, Laerence Erlbaum Associates.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrica*, 30, 179-185.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20, 141-151.

Lambert, Z. V., Wildt, A. R. & Durand, R. M. (1991). Approximating confidence intervals for factor loading. Multivariate behavioral research, 26, 421-434.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321-327

Visualizing Cluster Results
Using Package FlexClust and Friends

Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstrasse 33, 80539 München, Germany.
Friedrich.Leischr@R-project.org

Centroid-based partitioning cluster analysis is a popular method for segmenting data into more homogeneous subgroups. Visualization can help tremendously to understand the positions of these subgroups relative to each other in higher dimensional spaces and to assess the quality of partitions. In this talk we present several improvements on existing cluster displays using neighborhood graphs with edge weights based on cluster separation and convex hulls of inner and outer cluster regions. Using symbols or complete high-level plots in the nodes of the graph help to understand the association of background variables and clusters. A new display called shadow-stars can be used to diagnose pairwise cluster separation with respect to the distribution of the original data. Barplots of centroid profiles are improved by shading bars corresponding to statistically significant or user-relevant differences in darker colors. All methods will be demonstrated using real data from market segmentation and microarray data analysis.

IRLB SVD methods for R

Bryan W. Lewis

1. Affiliation of author A and author B
2. Second affiliation of author A
3. Second affiliation of author B
4. Third affiliation of author B

* Contact author: email@adresse.fr Bryan@revolution-computing.com

Keywords: XVD,PCA, LDA

Abstract: The singular value decomposition (SVD) is used by many important statistical methods and applications including principal components analysis and linear discriminant analysis. Numerical implementations of the SVD are computationally intensive. Many applications of the SVD often require only a few singular values and corresponding singular vectors. We introduce Baglama's recent implicitly-restarted Lanczos (IRLB) methods for computing a few singular vectors of a matrix to the R language. These state of the art methods significantly outperform existing R-language SVD implementations in computational and memory efficiency. Moreover, the IRLB algorithm is simple and easily scalable to parallel implementations appropriate to huge data.

References

(2007). *J. Baglama and L. Reichel,*
Augmented Implicitly Restarted Lanczos Bidiagonalization Methods, SIAM J. Sci. Comput., 27 (2005), pp. 19-42.

URL IS NOT AVAILABLE BUT WILL BE PRIOR TO EVENT

Computational Aspects and Windows Related Community Services on CRAN

Uwe Ligges

Fakultt Statistik, Technische Universitt Dortmund, Dortmund, Germany
ligges@statistik.tu-dortmund.de

Keywords: CRAN, packages, community service, parallelization, Windows

With the growth of the package repositories on CRAN, more and more has to be automated:

We will be talking about computational issues like the parallelization of package installation and package checking to be able to check a 1700 packages containing repository in less than 24 hours. It will be shown how much of parallelization is possible and how does this affect the useR! world of contributed packages on CRAN.

Another issue is the update handling on CRAN. If package updates are submitted, 'inverse recursive' checks on packages that depend on those updated packages need to be run in order to see that newly introduced features or changes do not break code in dependent packages. Some package maintainer might have seen first results of these fairly new check services.

Moreover, the community services provided (such as the binary repositories for Mac and Windows binaries) or the winbuilder service (a machine that allows users to upload source packages and that returns check results and a Windows binary package) will be presented – the latter in an online session.

Appliantized Massively Parallel High Performance R

Justin Lindsey

CTO, Netezza

jlindsey@netezza.com

To overcome performance limitations in both time (elapsed time of a computation) and space (total amount of information to operate on) general purpose computing has left the confines of single core computing. This departure has created powerful and significant systems that are multi-core (multithreaded) and multi-processor (parallel or distributed computing). The challenge with these paradigms is they have added significant complexity to the task of controlling and instructing the machine. A path forward using a parallel R appliance will be presented with a detailed discussion of a particular implementation and a simpler model for controlling the machine. The presentation will highlight accomplishments of the present and suggest areas for future work toward a common parallel R standard.

Estimation and Testing of Portfolio Value-at-Risk based on L-Comoment Matrices

Wei-han Liu
Department of Banking and Finance, Tamkang University
Taipei, Taiwan
Contact author: weihanliu2002@yahoo.com

Keywords: order statistics, L-moments, Trimmed L-moments, L-comoments, portfolio value-at-risk, backtesting, saddlepoint approximation

The estimation performance of portfolio value-at-risk (PVAR) hinges on the approximation of the multivariate profit-and-loss distribution (PL). This study applied the multivariate L-moments developed by Serfling and Xiao (2007) and resorts to nonparametric multivariate estimators and descriptive measures. The PVAR estimates are examined via four backtesting methods. In addition to the three backtesting approaches: unconditional coverage, independence, and conditional coverage (Christoffersen 2003), the new approach developed by Wong (2008), based on saddlepoint approximation technique, is included.

References

Christoffersen, Peter F. 2003. Elements of Financial Risk Management. San Diego, CA: Academic Press.

Hosking, J. R. M. 1990. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society, Series B 52:105-124.

Jurczenko, Emmanuel, and Bertrand Maillet, eds. 2006. Multi-moment Asset Allocation and Pricing Models: Wiley.

Lugannani, R., and S. O. Rice. 1980. Saddlepoint approximation for the distribution of the sum of independent random variables. Advanced Applied Probability 12:475-490.

Serfling, Robert, and Peng Xiao. 2007. A contribution to multivariate L-moments: L-comoment matrices. Journal of Multivariate Analysis 98 (1765 - 1781).

Wong, Woon K. 2008. Backtesting trading risk of commercial banks using expected shortfall. Journal of Banking & Finance 32 (7):1404-1415.

The TextometrieR package: textual data analysis for social sciences and humanities[†]

Sylvain Loiseau[1,*], Jean-Philippe Magué[1], Serge Heiden[1]

[1] UMR 5191 ICAR, Université de Lyon
[*] sloiseau@ens-lsh.fr

Keywords: Statistics in the Social and Political Sciences, Corpus linguistics, Textometry, Textual data analysis

We present the **TextometrieR** package which aims at providing tools for texts and corpus analysis. The package originates in the french tradition of textometry, born after the Benzécri's seminal works on multidimensionnal analysis (Lebart *et al.* 1998). This tradition has developped numerous methods for exploring and visualizing textual data. In the context of a growing interest for R in corpus linguistics (Gries 2008), this package add a new milestone after the packages zipfR (Hevert and Baroni, 2006) and languageR (Baayen, 2008).

The **textometrieR** package is developped as part of a research project gathering several representants of this french tradition, including statisticians, linguists, scolars working on the historical, political or literary discourses, etc. This project aims at summing up the work done in textometry in the past decade. It will provide methods for exploring collocation between words (allowing to identify various phenomenon, from lexicalized multi-word units to stylistic/ideological association between words), building concordance, observing lexical distribution (growing of vocabulary, zipfian distribution) performing multifactorial analysis (for instance in texts typology, comparing diachronical or genre difference between texts, etc.).

From an architectural point of view, the **textometrieR** package is the statistical component of a platform which associates R and the IMS Open Corpus Workbench, a powerfull full text indexer and search engine, under a single Java API. The R/Java communication is based on rJava. This plateform provides a uniform environment to query plain text corpora or annotated corpora (*i.e.* containing metadata or linguistic annotations), in order to build quantitative structures such as frequency lists or contingency tables, and to benefit from R power to analyze and visualize those structures.

References

Baayen R. H. (2008). Analyzing Linguistic Data: A practical introduction to statistics. Cambridge: Cambridge University Press.
`http://cran.r-project.org/web/packages/languageR/index.html`

Evert, S. and Baroni, M. (2006). The zipfR library: Words and other rare events in R, useR!2006 (Vienna, Austria).
`http://zipfR.R-Forge.R-project.org/`.

Gries S. Th. (2008). Quantitative Corpus Linguistics With R: A Practical Introduction. New York: Routledge.

Lebart L., Salem A., Berry L. (1998). Exploring textual data. Dordrecht : Kluwer.

Extracting oceanographic data via R:

An application to habitat modelling of marine species

Maite Louzao[1,2,*], Anne Goarant[3], Clara Peron[1], Jean-Baptiste Thiebot[1], David Pinaud[1], Philippe Koubbi[3], Karine Delord[1], Christophe Barbraud[1], Henri Weimerskirch[1], Charles-André Bost[1], Guy Duhamel[4], Patrice Pruvost[4]

1. Centre d'Etudes Biologiques de Chizé, CNRS UPR 1934, 79369 Villiers en Bois, France
2. UFZ Leipzig-Halle, Permoserstrasse 15, 04318 Leipzig, Germany
3. LOV, CNRS , Station Zoologique BP28, 06230 Villefranche sur Mer, France
4. MNHN, DMPA -UMR 5178, 43 rue Cuvier, 75005 Paris, France

* Contact author: louzao@cebc.cnrs.fr

Thanks to the rapid development of remote sensing technologies, the availability of oceanographic data has dramatically increased during the last years. Oceanographic data available via website are very diverse in terms of sources (NOAA, NASA), file type (netcdf, hdf) and resolution (temporal and spatial). Large extraction of data for different time periods and areas can therefore be time-consuming and difficult to handle within the same format. The Environmental Research Division, Southwest Fisheries Science Center and US National Marine Fisheries Service has recently developped a software (Xtractomatic, http://coastwatch.pfel.noaa.gov/xtracto/) that simply make available environmental data (SST, chlorophyll, wind) within the R environment. This R-based tool allows the extraction of oceanographic data along (1) a series of input of time, longitude and latitude (e.g. a track of an animal or ship) specifying an extraction box and (2) a 3-Dimensional cube specified by limits of longitude, latitude and time.

Here, we present how Xtractomatic can be applied to the extraction of oceanographic data which are used for habitat modelling of marine species in the southern Indian Ocean, including top predators. We relied on two types of data: (1) tracking data for seabirds and (2) occurrence patterns of pelagic fishes. Both type of data were placed over a standard grid and the Xtractomatic function was used to extract oceanographic parameters, after adjusting time resolution and spatial scale to the species biology and locations accuracy. Once oceanographic variables were obtained, different regression techniques (Generalized Linear Mixed Models and Generalized Additive Models) were applied within the R environment in order to identify those variables which best explained the oceanographic habitat of the species and predict density or habitat use probability.

Given the high conservation concern of marine top predators and current major environmental changes, the standardization of the whole habitat modelling process (including the download of large amount of environmental data) makes much easier and faster the investigation of the oceanographic processes influencing marine species distribution patterns.

Feasibility of using COSA as a genome-wide SNP screen

Gitta Lubke[1*] **, Jacqueline Meulman**[2]

1. University of Notre Dame, USA
2. Leiden University, The Netherlands
* Contact author: glubke@nd.edu

Keywords: COSA, genome-wide association, SNP

Genome-wide association studies (GWAS) of psychiatric disorders have not yet resulted in a significant breakthrough. The common approach in GWAS is to test single nucleotide polymorphisms (SNPs) univariately, thus necessitating substantial corrections for multiple testing. The complexity of the phenotype is reduced to a binary indicator (i.e., case vs. control). The Clustering Objects on Subsets of Attributes (COSA) algorithm is designed to detect clusters of objects (here: subjects) that are similar to each other, and to simultaneously select cluster-specific subsets of attributes (here: SNPs and phenotype items) that are relevant for the clusters. COSA permits a joint analysis of SNPs and phenotype items, and therefore combines the advantages of a multivariate analysis with the flexibility of targeting phenotype symptom patterns and clusters of subjects within the cases. COSA has not been used for the analysis of genome-wide SNP data. We present results of a simulation study that investigates the feasibility of using COSA as a genome-wide SNP screen.

References

Friedman, JF, Meulman, JJ (2004). Clustering objects on subsets of variables (with discussion). Journal of the Royal Statistical Society, Series B, 66, 815–849.

Automating SQL queries from formulas: loading data on demand

Thomas Lumley

Department of Biostatistics, University of Washington, Seattle. * Contact author: tlumley@u.washington.edu

Keywords: model formula, database, programming, large data, regression

A relatively common situation with large data sets is that the variables needed for any specific computation will fit in memory, but the entire data set is large enough to be inconvenient. For example, the 2006 NHIS public use data set has about 25,000 observations on 546 variables, and will take up about 100Mb, enough to slow down a computer with 1Gb memory. The 2007 BRFSS public use data has about 430,000 observations and 343 variables. The whole data set cannot be loaded into 32-bit R, but there is no difficulty in handling a dozen variables or so.

I will describe an approach to automated loading of data from a relational database by extracting the names of the necessary variables from a model formula or expression. This approach can be used to wrap existing code that is unaware of databases. Only read access to the database is needed, since newly defined variables are stored as definitions and created as the data is loaded.

Sparse Matrices in package Matrix and applications

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

1. ETH Zurich, Switzerland and University of Wisconsin, Madison, USA
2. R Core Development Team * Contact author: maechler@R-project.org

Keywords: sparse matrices, S4 classes and methods, large data

Linear algebra is at the core of many areas of statistical computing and from its inception the S language has supported numerical linear algebra via the `matrix` type and several functions and operators. However, these data types and functions do not provide direct access to all of the facilities for efficient manipulation of dense matrices, as provided by the Lapack subroutines, and they do not provide for manipulation of sparse matrices. The Matrix package provides a set of S4 classes for dense and sparse matrices that extend the basic matrix data type. Methods for a wide variety of functions and operators applied to objects from these classes provide efficient access to BLAS (Basic Linear Algebra Subroutines), Lapack (dense matrix), CHOLMOD including AMD and COLAMD and Csparse (sparse matrix) routines. The talk will focus on the *sparse* matrix classes and methods and mention some specific applications, notably in other CRAN packages.

An introduction into the sparse matrix representations, their corresponding Matrix classes, and typical constructor, inspection and visualization functions.

We will explore the space of Matrix classes, its many hundreds of methods and explain why the useR typically does not have to know most of these details.

Least squares fitting with large sparse design matrices can be accomplished via efficient sparse Cholesky decompositions and allows solving systems of sizes that would not be possible using traditional "dense" matrices.

One novel application in spatial statistics, needs to compute the determinant of $\det|I - \rho W|$ for many values of ρ and a large sparse $n \times n$ matrix W (e.g. $n = 15'000$). This problem can be reduced to compute the symbolic part of the cholesky decomposition of A *once* only, and updating the decomposition to the one of $W - \lambda I$ for varying λ, which is comparatively fast.

References

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

Power Analysis for Multivariate Generalised Linear Models

Mariann Borsos[1] , **István Jánosi**[1,*]

1. Planimeter Kft., Budapest, Hungary

* Contact author: janosi@planimeter.hu

Keywords: Statistics, sample size, power, SAS, multivariate generalised lienar model.

Sample size determination is an important and continuously developing part of statistics. New designs, new approaches are continuously emerging and broadening of computational opportunities are also increase the set of avaialable techniques.

Most of our work is based on the considerations of Daniel F. Heitjan [1] who developed a S-Plus package for power and sample size analysis of a special group of Genelised Linear Models (GLMs), the Multivariate Generalised Linear Models.

Multivariate Generalised Linear Models can be mathematically described by the model

$$\mathbf{Y}_{N \times p} = \mathbf{X}_{N \times q}\mathbf{B}_{q \times p} + \mathbf{E}_{N \times p}$$

where

$\mathbf{Y}$ is the vector of response, $\mathbf{X}$ is the design matrix, $\mathbf{B}$ is the vector of effect coefficients, and $\mathbf{E}$ is the vector of errors assuming that $\mathbf{E} \sim N_{n \times p}(\mathbf{0}, \mathbf{I}_N \otimes \boldsymbol{\Sigma})$. Here $\boldsymbol{\Sigma}$ represents the covariance matrix of the $\mathbf{Y}$ columns (varying over "within" factor levels) and often referred as repeated measures.

Sample size determination for these models is also supported by SAS (among others with a SAS-macro developed by Keyes and Muller [2] in 2005, but we could not find any trace of having such a tool in R. This was our motivation to implement Heitjan's methods in R.

What we exactly did and would like to present to R-community is:

1. Implementation of Heitjan's methods in R.

2. Our experiences on application the procedures (in R) on the examples of Muller, LaVange, Ramey and Ramey [3].

3. Extension of Heitjan's original work with plotting procedures which was completely missing from his published procedures.

References

[1] Heitjan, D.F. (2009). Powerex package for S-Plus,
    http://www.cceb.upenn.edu/pages/heitjan/power/.

[2] Keyes, L.L. and Muller, K.E. (2005). *IML POWER PROGRAM USER'S GUIDE: November, 1992*.

[3] Muller, K.E., LaVange, L.M., Ramey, S.L., and Ramey, C.T. (1992). Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications. *Journal of the American Statistical Association, 87, 1209 - 1226*.

metaMA : an R package implementing meta-analysis approaches for microarrays

Guillemette Marot[1,*] , **Jean-Louis Foulley**[1], **Claus-Dieter Mayer**[2], **Florence Jaffrézic**[1]

1. INRA Génétique Animale et Biologie Intégrative, Jouy en Josas (France)
2. Biomathematics and Statistics Scotland

* Contact author : guillemette.marot@jouy.inra.fr

Keywords: Microarrays, meta-analysis, shrinkage

Microarrays have been widely used to detect differentially expressed genes, for example between normal and tumoral samples. Due to the high cost of these experiments, results often rely on small sample size designs. Since more and more microarray data are available in the public domain, meta-analysis, which consists in combining summary statistics from different studies, is of great interest in this field. Thus, meta-analysis offers the possibility to considerably increase the statistical power and gives more accurate results. The package metaMA implements moderated effect size combinations, as proposed by Marot et al.(2009) as well as inverse normal p-value combinations, with p-values calculated from moderated t-tests.
We compared all these meta-analysis methods in an extensive simulation for various amounts of inter-study variability. We found that

1. moderated effect size combination improved existing gene-by-gene effect size approaches

2. effect size combination is more conservative than the p-value combination method, i.e. it stays much below the nomimal FDR. This is also reflected in the fact that the effect size combination eliminates more false positives than the p-value combination among the genes which are significant in at least one individual study but not in the meta-analysis.

3. p-value combination outperformed the other classical meta-analysis methods in terms of sensitivity and gene ranking (larger areas under the ROC curves).

References

G.Marot, J.-L. Foulley, C.-D. Mayer and F. Jaffrézic (2009). Moderated effect size and p-value combinations for microarray meta-analyses. Submitted.

NCBI2R: An R package to navigate and annotate genes and SNPs

Scott Melville[1,*] , Christian Fuchsberger[1]

1. Institute of Genetic Medicine, European Academy Bozen/Bolzano (EURAC), Viale Druso 1, 39100 Bolzano, Italy, Affiliated Institute of the University Lübeck, Germany.
* Contact author: scott.melville@eurac.edu

Keywords: Genome wide association studies, candidate gene association studies, NCBI, annotation, gene interactions.

NCBI2R is a new R package that annotates lists of SNPs and/or genes, with current information from NCBI. Functions are provided that with one command will annotate the results from genome wide association studies to provide a broader context of their meaning. Other functions enable comparisons between a user's GWA results, and candidate snp/gene lists that are created from keywords, such as specific diseases, phenotypes or gene ontology terms. Commands are simple to follow and designed to work with R objects to integrate into existing workflows. The output produces text fields and weblinks to more information for items such as: gene descriptions, OMIM, pathways, phenotypes, and lists of interacting and neighboring genes. Annotation can then be used in R for further analysis, or the objects can be customized for use in spreadsheet programs or web browsers. The NCBI2R package was designed to allow those performing genome analysis to produce output that could easily be understood by a person not familiar with R.

Good Relations with R

David Meyer[1,3,*] , **Kurt Hornik**[2,3]

1. Department of Information Systems and Operations
2. Department of Statistics and Mathematics
3. WU Vienna (Vienna University of Economics and Business), Austria

* Contact author: David.Meyer@R-Project.org

Keywords: relations, sets, consensus ranking, benchmark experiment

Relations are a very fundamental mathematical concept: well-known examples include the linear order defined on the set of integers, the equivalence relation, notions of preference relations used in economics and political sciences, etc. A k-ary (finite) relation is defined by its *domain*, a k-tuple of sets, and its *graph*, a set of k-tuples. Package **relations** provides data structures along with common basic operations for relations and relation ensembles (collections of relations with the same domain). In doing so, it builds on the infrastructure for (generalized and customizable) sets and tuples provided by package **sets**. Package **relations** also features various relational algebra-like operations, such as projection, selection, and joins. In addition, many relations can be visualized by means of Hasse diagrams (see Figure 1). Finally, it contains algorithms for finding suitable consensus relations for given relation ensembles, including the constructive approaches of Borda, Condorcet and Copeland, as well as optimization-based methods which minimize the aggregate symmetric difference distance between the ensemble members and their consensus. We show how relations can be obtained and manipulated, and how the functionality in the package can be employed to rank the results of benchmarking experiments.

Figure 1: Hasse Diagram of the inclusion relation on the power set of $\{a, b, c\}$.

References

K. Hornik and D. Meyer (2007). Deriving consensus rankings from benchmarking experiments. In R. Decker and H.-J. Lenz, *Advances in Data Analysis.* Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag: Heidelberg, 163–170.

F. Marcotorchino and P. Michaud (1982). Agrégation de similarités en classification automatique. *Revue de Statistique Appliquée*, **30**/2, 21–44.
http://www.numdam.org/item?id=RSA_1982__30_2_21_0

S. Rgnier (1965). Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, **4**, 175-191.

R-package *MultiSelection*: Optimizing Multi-stage Selection Gain and Controling the Variance

Xuefei Mi[1*] , **Friedrich Utz**[1]

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

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

Keywords: Multi-normal, Selection, QTL

We developed an R-package for maximizing the gain of a multi-stage selection procedure under certain restrictions, e.g. a given annual budget or certain risk limits of each stage. This package is mainly applied in fields of plants/animals breeding, where a multi-normal heredity regression model is commonly built. By implementing the R-package *mvtnorm*, which calculates the multi-variate normal distribution, the number of independent variables is increased from three upto one thousand. This makes it possible that hudge amount of the Marker and QTL information can be used.

References

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

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

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

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

Uncovering Interactions in Random Forests

Jacob Michaelson[1,*], Marit Ackermann[1], and Andreas Beyer[1]

1. Cellular Networks and Systems Biology, Biotechnology Center, TU Dresden

* Contact author: jacob.michaelson@biotec.tu-dresden.de

Keywords: Random Forests, interactions, gene regulation

Over the past decade, Random Forests [1] have proven to be a powerful machine learning algorithm in many applications. Its usefulness stems not only from its predictive ability, but also from information it gives about the structure of the model underlying the data. The latter attribute can be particularly appealing in problems where $N << P$ and a majority of the predictor variables do not participate in the underlying true model. Such is the case in many biological problems, where Random Forests can be used to sift through large volumes of data to find meaningful interactions between predictor variables. Measures of variable importance already exist [1, 2] but these generally give information about the "main effects" and do not provide direct insight about relationships between predictors.

In this work we propose a novel approach to detecting variable interactions in Random Forests. Information on predictor co-occurrence in the forest's trees is used as a basis for both a frequentist and a Bayesian approach for uncovering interactions between the variables. To interpret the interdependencies, a graph of the variable interactions is constructed. By depicting this information as a graph, we impose no limit on the order or characteristics of the interactions. We apply the methods to gain new insight on neuronal regulatory pathways in the hippocampus.

References

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5, 2001.

[2] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.

A domain-morphing approach to smoothing over complex regions

David Lawrence Miller[1,*]

1. University of Bath

* Contact author: d.l.miller@bath.ac.uk

Keywords: Smoothing; Generalizedd Additive Models (GAMs); Spatial smoothing; penalized regression splines.

Smoothing over complex 2-D regions is difficult. Several approaches have been proposed in past years including finite element analysis (Ramsay, 2002), within-area distance (Wang & Ranalli, 2007) and recently soap film smoothing (Wood, Bravington & Hedley, 2008.) Here I investigate an alternative method based on the Schwarz-Christoffel transform from complex analysis. This takes the region and "morphs" it to a rectangle or disk in a prescribed way. We may then smooth over the transformed area using penalized regression splines and transform this smooth back to the original domain in order to perform analysis. I explore the utility of this transform on both real and simulated data.

References

Ramsay, T. (2002) Spline smoothing over difficult regions. JRSSB, 64(2), 307-319.

Wang, H., M.G. Ranalli (2007) Low-Rank Smoothing Splines on Complicated Domains. Biometrics, 63(1), 209-217.

Wood, S.N., M.V. Bravington and S.L. Hedley (2008) Soap film smoothing. JRSSB, 70(5), 931-955.

splm: econometric analysis of spatial panel data

Giovanni Millo[1,2,*] , **Gianfranco Piras**[3,4]

1. DiSES, University of Trieste
2. Generali Research and Development
3. GeoDa Center, Arizona State University
4. Universidad Catolica del Norte
* Contact author: giovanni_millo@generali.com

Keywords: Spatial Panels, ML, GM, Tests

We illustrate the new `splm` package, aimed at providing a comprehensive resource for spatial panel econometrics. The package fills a gap in applied practice, as the relevant estimators and tests are well established in the literature but to date they lack user-friendly and widely available software implementations.

Building on the infrastructure for spatially referenced data in package `spdep`, we provide estimators for the standard panel models in the spatial econometrics literature: fixed and random effects with either a spatial lag or spatial correlation in the error term, based on both the concurrent approaches prevailing in the literature, i.e. the Maximum Likelihood framework pioneered by Anselin (1988) and the Generalized Moments framework of Kapoor, Kelejian and Prucha (2007).

Some of the model estimation procedures are generalized to the case of spatially *and* serially correlated error terms. GM estimators for systems of equations are also available.

We also provide the Lagrange Multiplier joint, marginal and conditional specification tests from the work of Baltagi et al. (2003, 2007).

The user interface aims at consistency w.r.t. the spatial (non-panel) estimators in package `spdep` and the panel (non-spatial) estimators in package `plm`.

We briefly discuss code optimization aspects of the computationally heavy Maximum Likelihood routines that have up to now hindered the practical implementation of these estimators. The GM approach, on its part, yields very fast estimators that can be applied to comparatively big datasets.

We conclude with an empirical illustration on a well-known data set from the panel data literature.

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Kluwer (UCSB)

Kapoor, M., Kelejian, H. and Prucha, I. (2007). Panel Data Models with Spatially Correlated Error Components *Journal of Econometrics* 140, 97–130.

Baltagi, B.H., Song, S.H., Jung, B.C. and Koh, W. (2003). Testing for Spatial Correlation, Serial Correlation and Random Effects using Panel Data *Journal of Econometrics* 140, 5–51.

A Framework for Hypothesis Tests in Statistical Models With Linear Predictors

Georges Monette[1*] and John Fox[2]

1. Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada
2. Department of Sociology, McMaster University, Hamilton, Ontario, Canada
* Contact author: georges@yorku.ca

A great deal of confusion, and even prejudice, is associated with the differentiation of "types" of statistical tests in linear and similar statistical models. This paper elucidates the distinction between so-called "type-II" and "type-III" tests in linear models, explaining the nature of the hypotheses tested by each kind of test. We then show how type-II Wald tests can be extended to any statistical model with a linear predictor and asymptotically normal coefficients, providing a practical strategy for conducting such tests without having to refit restricted versions of the model. This general method is implemented in the Anova() function in the car package.

Kaleidoscope Graphs in R

Mario A Morales [1,2*]

1. Hunter College of City University of New York
2. Polytechnic Institute of New York University
* Contact author: mario.morales.44@gmail.com

Keywords: Data Visualization

The primary aim of this talk is to present an innovative technique for building and analyzing kaleidoscope images that explain patterns and relationships in time series, cross section and longitudinal problems.
These graphs are programmed in R and one example is presented with simulated data.

An R package for analyzing truncated data

Carla Moreira[1*], Jacobo de Uña Álvarez[1], Rosa Crujeiras[2]

1. Department of Statistics and OR, University of Vigo
2. Department of Statistics and OR, University of Santiago de Compostela
* Contact author: carlamgmm@gmail.com

Keywords: double truncation, nonparametric maximum likelihood.

The goal of this work is to present an R package, developed to analyze truncated data, with application in a number of fields, including Astronomy, Economics and Survival Analysis. In this package, the two EM algorithms proposed by Efron and Petrosian (1999) and the algorithm by Shen (2008) are included as three possible approaches to approximate the non-parametric maximum likelihood estimator under double truncation.

This software is also prepared to deal with data which are one-sided truncated. Specifically, the package allows to compute the estimator introduced by Lynden-Bell (1971) for left-truncated data, or to specify right truncation, if that is the case. A graphical output includes densities and survival estimates.

References

Efron E. and Petrosian, V. (1999) Nonparametric methods for doubly truncated data. Journal of the American Statistical Association 94, 824-834.

Lynden-Bell, D. (1971) A method of allowing for known observational selection in small samples applied to 3CR quasars. Mon. Not. R. Astr. Soc. 155, 95-118.

Shen P-S. (2008) Nonparametric analysis of doubly truncated data. Annals of the Institute of Statistical Mathematics. DOI 10.1007/s10463-008-0192-2.

Optimization challenges in fitting stochastic models for atomic positions in nanoparticles

Katharine M. Mullen*, Victor Krayzman, Igor Levin

Structure Determination Methods Group, Ceramics Division

National Institute of Standards and Technology

100 Bureau Drive, M/S 8520

Gaithersburg, MD, 20899, USA

Contact author: katharine.mullen@nist.gov

Given a material, what is the arrangement of atoms inside it? Traditional crystallographic methods base the answer to this question on models for the material with atomic positions arranged in an infinitely periodic lattice. These models are helpful for the study of bulk materials, but do not well-approximate atomic positions in nanoparticles and other materials in which the atomic positions are disordered. Furthermore, the number of atoms, atomic positions and other structural features characteristic of an ensemble of nanoparticles may be best described by a stochastic model. In this talk we discuss stochastic models for nanoparticles. We show how the atomic pair distribution function (i.e., distribution of interatomic distances) is connected to structural features of nanoparticles such as lattice parameters. Finally, we present the optimization challenge associated with finding the most likely set of atomic positions in an ensemble of nanoparticles given measured pair distribution functions, and discuss the use of R in addressing this challenge.

References

[1] Simon J. L. Billinge and Igor Levin, *The Problem with Determining Atomic Structure at the Nanoscale*, Science **316** (2007), no. 5824, 561–565.

[2] Robert L McGreevy, *Reverse Monte Carlo Modelling*, Journal of Physics: Condensed Matter **13** (2001), no. 46, R877–R913.

A case study on using generalized additive models to fit credit rating scores

Marlene Müller[1,*]

1. Fraunhofer ITWM, Fraunhofer-Platz 1, D–67663 Kaiserslautern, Germany

* Contact author: marlene.mueller@gmx.de

In credit rating, the finally fitted rating score is not only intended to provide the optimal classification result but also to serve as a modular component of a (typically complex) rating system. This means in particular that a rating score should be given by a linearly weighted sum of so-called rating factors, a procedure which can be easily interpreted and understood by non-statisticians. An important issue is also the possibility to run stress-tests on the final model in order to study the effects of extreme inputs.

All of this leads to the fact that the logit model or logistic regression approach is one of the most popular models for estimating credit rating scores. A possible nonlinear (more precisely nonparametric) dependence of the rating score on the original raw data variables is typically separated within an initial transformation step. From a point of view of optimizing the model fit and thus the potential to identify possible credit defaults more precisely, generalized additive models (GAM) would allow for a simultaneous estimation of the initial transformation together with the final logit fit.

Meanwhile R comprises a number of different packages to fit generalized additive models. In this study we compare GAM estimating approaches with a focus on the specific structure of credit data: small default rates, mixed discrete and continuous explanatory variables, possibly nonlinear dependencies between the regressors.

References

Hastie, T. J.; Tibshirani, R. J. (1990). Generalized Additive Models. Chapman and Hall, London.

Müller, M.; Härdle, W. (2003): Exploring Credit Data. In: Bol, G.; Nakhaeizadeh, G.; Rachev, S.T.; Ridder, T.; Vollmer, K.-H. (eds.): Credit Risk - Measurement, Evaluation and Management, Physica-Verlag.

Wood, S. N. (2006): Generalized Additive Models: An Introduction with R. Chapman and Hall, London.

Web Interface to R for High-Performance Computing

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

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

* Contact author: nakanoj@ism.ac.jp

Keywords: Job Queuing, MPI, multi-thread BLAS, Supercomputer

As the amount of data is continuously increasing in these days mainly because of the development of data acquisition systems using powerful personal computers and networks, there is much need for R to handle large data sets. Supercomputers are appropriate means for such purposes. Supercomputer requires batch and job queuing system for achieving highest performance. The usage of such a system, however, is not easy for novices. For the ease of use of supercompers and personal cluster systems, we have developed a Web interface to R working with several job queuing systems such as OpenPBS, Torque, LoadLeveler and a UNIX command "at". Our Web interface enables us to select queue, numbers of active cores for multi-thread BLAS and MPI slaves. It uses a daemon process to execute user authentication, job submitting and file transfer.

Unifying optimization algorithms in R for smooth, nonlinear problems

John C. Nash[1,*] , Ravi Varadhan[2]

1. University of Ottawa, Telfer School of Management, Ottawa, Canada (retired)
2. The Center on Aging and Health, Johns Hopkins University, Baltimore, USA
* Contact author: nashjc@uottawa.ca

The optim() function in the R 'stats' collection provides a powerful yet clean and easy-to-use mechanism for users to launch optimization tasks. Over the years, however, other packages have been introduced in order to provide for special or supposedly more-efficient capabilities. We present some work to unify these tools and to help the user make sensible choices, but aim to retain the essential tidiness of the interface. We welcome commentary and assistance from the R and optimization communities to further these efforts.

Our objectives are:

1. To unify optimization tools in R for solving smooth, nonlinear, box-constrained optimization problems.

2. To provide "guidance" to users for choosing the appropriate algorithm, automatically setting up the appropriate function call essentially in the same style as optim().

3. To update/extend algorithms in optim()

One motivation for this work is that the early tools (Nelder-Mead, CG, BFGS) were chosen and adapted by one of us (Nash, 1979) for use on very limited systems three decades ago. They are still very usable and useful tools, but we believe that users need to be led to choices better suited to their problems given the evolution of both statistical problems and optimization techniques.

Further, many statistical workers are not familiar with the difficulties that attend optimization, so we also wish to provide better defaults and better guidance on the use of a new optim(). We want a new method, really a wrapper for other methods, that we call "GUIDED" that will assist the user in creating the optim() syntax. In this we take inspiration from the Decision Tree for Optimization Software (Mittelman, 2008). Further, we believe some new developments in optimization such as Powell's (2007) BOBYQA are likely better candidates for the default tool than Nelder-Mead.

The GUIDED selection also promotes a unification of optimization methods within R. The optim() function supports only a small subset of the available tools from different packages. We would like to provide hooks to allow different optimization approaches to be called using the optim() structure. Related questions can be posed about cleanly linking optim() to related tools such as nls(), both in the computational structures and in the documentation.

References

Nash J C (1979) Compact numerical methods for computers: linear algebra and function minimisation, Bristol: Adam Hilger.

Mittelman H (2008) Decision Tree for Optimization Software,
    http://plato.asu.edu/guide.html

Powell M J D (2007) *Developments of NEWUOA for minimization without derivatives*,
    www.damtp.cam.ac.uk/user/na/NA_papers/NA2007_05.pdf. (Note: The BOBYQA software is described, but was not released until January 2009 pending selection of some options to be set as defaults.)

Size Estimation - Statistical Models for Underreporting

Gerhard Neubauer[1,*], Gordana Djuraš[1], Herwig Friedl[2]

1. Institute of Applied Statistics, JOANNEUM RESEARCH, Graz, Austria
2. Institute of Statistics, Technical University, Graz, Austria

* Contact author: gerhard.neubauer@joanneum.at

Keywords: Binomial Model, conditional Poisson models, regression

Underreporting is a problem in data collection, when events are counted and for some reason errors occur. The most prominent example is crime reporting, where crimes associated with shame are likely not to be reported to the police, just as theft of low value goods. The same holds for traffic accidents with minor damage. And also counting infectious diseases like HIV may be subject to underreporting.

As a consequence the mean of the observed counts is smaller than the true mean λ. Using a Binomial model the mean of the observed counts is $\mu = \lambda\pi$, with π the reporting probability, and both parameters to be estimated. Neubauer & Friedl (2006) introduced a regression approach for the Binomial model and - to adopt for overdispersion - also for a Beta-Binomial model. The Binomial and a Beta-Binomial regression model are suited for a wide range of applications. However, if the sample variance is larger than the sample mean the binomial approach fails to give reasonable estimates. For this kind of data Neubauer & Djuraš (2008) proposed a regression model based on the Generalized Poisson distribution. This model allows to handle Poisson under- and overdispersion, as it covers the binomial, Poisson and the negative binomial case. Recently Neubauer & Djuraš (2009) proposed a further extension of the binomial approach leading to a Beta-Poisson regression model.

A second approach to underreporting builds upon conditional Poisson models, i.e. $Y|L \sim Poisson(L)$ or $Y|L \sim Poisson(L\pi)$. Here the limit $\pi \to 1$ does not cause a problem as with the binomial approach, where $Y \to \lambda$. Using different distributional assumptions for L we obtain a variety of models for possibly perfect reporting systems.

Inference in all cases is based upon maximum likelihood estimation. The scope of the R implementation in package `sizEst` is to cover all mentioned models in a framework, where estimation, testing and model selection is enabled. The approach is illustrated with examples from real data.

References

Neubauer, G. and Djuraš, G. (2008). A Generalized Poisson Model for Underreporting. In: Proceedings of the 23rd International Workshop on Statistical Modelling, 7-11 July, 2008, Utrecht, Netherlands.

Neubauer, G. and Djuraš, G. (2009). A Beta-Poisson Model for Underreporting (Tech. Rep. No.1). Graz: Joanneum Research.

Neubauer, G. and Friedl, H. (2006). Modelling sample sizes of frequencies. In: Proceedings of the 21st International Workshop on Statistical Modelling, 3-7 July 2006, Galway, Ireland.

R and spreadsheets - combining different programming paradigms

Erich Neuwirth[1]

1. University of Vienna

* Contact author: erich.neuwirth@univie.ac.at

Keywords: Spreadsheets, software integration, direct manipulation software

R has a rather steep learning curve due to the fact that it is a programming language and not a a menu driven program. Most users working with data are familiar with the spreadsheet paradigm, which among other things gives a view on data and on formulas underlying data simultaneously. Combining spreadsheet programs and R allows to embed the computational strength of R into the widely familiar spreadsheet model.

Our systems (RExcel based on the R/Scilab Server and ROOo based on the RUno extension for OpenOffice) make R accessible from within spreadsheet programs.

Since the paradigms of using a spreadsheet is radically different from using a programming language and also from using a menu driven statistics program, it is very important to design different user interfaces for different user groups of our integrated software system.

Different user groups to be considered are

- Learners and students of statistics with no programming background

- Experienced programmers with only little knowledge of statistics

- Users of "canned" statistical method

- Power users able to adapt existing statistical methods and adapt and design spreadsheet formulas

- Software developers preparing in house solutions for naive end users

We will show what kind of interfaces support which kind of user model, and we will also show how it is possible to integrate existing or newly written R packages directly into the spreadsheet model in any of these user models.

Influence.ME: Tools for recognizing influential data in mixed models

Rense Nieuwenhuis[1,*], Ben Pelzer[1], Manfred te Grotenhuis[1]

1. Radboud University, The Netherlands

* Corresponding author. E-mail: contact@rensenieuwenhuis.nl

Keywords: social sciences, influential cases, diagnostics, mixed effects models, lme4

Mixed effects regression models tend to become common practice in the field of Social Sciences. However, diagnostic tools to evaluate these models lag behind. For instance there is no general applicable tool to check whether all units (or cases) roughly have the same influence on the regression parameters. It is however commonly accepted that tests for influential cases should be performed, especially when estimates are based on a relatively small number of cases. Testing for influence with mixed effects models is especially important in Social Science applications, for two reasons. First, models in the Social Sciences are frequently based on large numbers of individuals while the number of higher level units is often relatively small. Secondly, often the higher level units are remarkably similar, for instance in the case of neighboring countries. `Influence.ME` is a new package for R which provides two innovations for evaluating influential cases: it extends existing procedures for use with mixed effects models, and it allows to not only search for single influential cases, but for combinations of cases that as a combination exert too much influence.

The basic rationale behind measuring influential cases is that when iteratively single units are omitted from the data, models based on these data should not produce substantially different estimates. To standardize the assessment of how influential a single observation is, several measures of influence are common practice. First, DFBETAS is a standardized measure of the absolute difference between the estimate *with* a particular case included and the estimate *without* that particular case. Second, Cook's distance provides an overall measurement of the change in *all* parameter estimates, or a selection thereof.

To apply the same logic to mixed effects models one has to measure the influence of a particular higher level unit on the estimates of a higher level predictor. This means that the mixed effect model has to be adjusted to neutralize the unit's influence on that estimate, while at the same time allowing the unit's lower-level cases to help estimate the effects of the lower-level predictors in the model. This procedure is based on a modification of the intercept and the addition of a dummy variable for the cases that might be influential. `Influence.ME` provides several measures of influential cases, and is specifically designed for use with mixed effects regression models using the afore mentioned modified intercept and dummy approach. Using both 'real' and simulated data from Social Science applications of mixed effects models, five tools to detect influential cases which are available in the package will be discussed:

- Cook's Distance

- DFBETAS

- Percent change of the estimated parameter magnitude

- Changes in statistical significance of parameter estimates

- Changes in the sign of parameter estimates

In contrast with other algorithms for detecting influential cases, influence.ME is capable to uncover groups of cases that are influential. Since this rapidly becomes computationally highly intensive, additional script functions are provided that assist in manually dividing the computation into multiple sessions, or to possibly to share the computations between different computers.

References

Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics. Identifying Influential Data and Source of Collinearity.* Wiley.

Snijders, T.A. & Bosker, R.J. (1999). *Multilevel Analysis, an introduction to basic and advanced multilevel modelling.* Sage.

Van Der Meer, T., Te Grotenhuis, M. & Pelzer, B. *Influential cases in multi-level modeling. A methodological comment on 'National context, religiosity, and volunteering' by Ruiter and De Graaf.* Current status: resubmitted to the American Sociological Review.

A Software Framework for Measuring Efficiency

Veska Noncheva [1,2,*] , **Armando Mendes** [1] , **Emiliana da Silva** [3]

1. CEEAplA, Azores University, 9501-801 Ponta Delgada, Portugal
2. Faculty of Mathematics and Informatics, University of Plovdiv "Paisii Hilendarski", Plovdiv, Bulgaria
3. CEEAplA, Azores University, 9700-851 Angra do Heroísmo, Portugal
* Contact author: veska@uac.pt

Keywords: non-parametric data analysis, Data Envelopment Analysis

The producers always aim at increasing the efficiency of their production process. However, the producers do not always succeed in optimizing their production. In the last years, the interest on Data Envelopment Analysis (DEA) as a powerful tool for measuring efficiency has increased. This is due to the large amount of data-sets available for description of the phenomena under study, and at the same time, to the need of timely and not costly information.

The "Productivity Analysis with R" (PAR) framework establishes a user-friendly data envelopment analysis environment with special emphasis on variable selection and aggregation, and summarization and interpretation of the results. The starting point is the following R packages: DEA [Diaz-Martinez and Fernandez-Menendez, 2008] and FEAR [Wilson, 2007]. The DEA package performs some models of Data Envelopment Analysis presented in [Cooper et al., 2007]. FEAR is a software package for computing nonparametric efficiency estimates and testing hypotheses in frontier models. FEAR implements the bootstrap methods described in [Simar and Wilson, 2000].

PAR is a software framework using a variety of models estimating efficiency and providing results explanation functionality. PAR framework has been developed to distinguish between efficient and inefficient observations of performances and to advise explicitly for producers' possibilities to optimize their production. PER framework offers several R functions for a reasonable interpretation of the data analysis results and text presentation of the information obtained. The output of the efficiency study with PAR software is self explanatory.

We are applying PAR framework to estimate the efficiency of the agricultural system in Azores [Mendes et al., 2009]. All Azorean farms will be clustered into homogeneous groups according to their efficiency measurements to define clusters of "good" practices and cluster of "less good" practices. This makes PAR appropriate to support public policies in agriculture sector in Azores.

References

Cooper, W. W., Seiford, L. M. and Tone, K. (2007). Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. Second edition. Springer. New York.

Diaz-Martinez, Zuleyka and Jose Fernandez-Menendez (2008). DEA: Data Envelopment Analysis. *R package version 0.1-2.*

Mendes A., V. Noncheva, E. Silva (2009). *Decision Support for Enhanced Productivity with R Software: An Azorean Farms Case Study*, accepted at the Thirty Eighth Annual Meeting of WDSI, Hawaii, April 7-11, 2009.

Simar, L., and Wilson, P.W. (2000). A general methodology for bootstrapping in non-parametric frontier models, *Journal of Applied Statistics*, 27, 779-802.

Wilson P. W. (2007). FEAR 1.0: A Software Package for Frontier Efficiency Analysis with R, *Socio-Economic Planning Sciences*, forthcoming.

Spatial odds ratio of disease in epidemiological studies with ordinal responses: a methodology using package VGAM.

Ana Carolina Cintra Nunes Mafra[1,2,*], Ricardo Cordeiro[1,2], Celso Stephan[1,2]

1. State University of Campinas – Unicamp
2. Spatial Analysis of Epidemiological Data Lab - epiGeo
* Contact author: anacarol.nunes@gmail.com

Logistic regression models are frequently used in epidemiological studies based on scales or outcomes that may have ordinal responses to analyze data, but without incorporating spatial distribution of disease, some for not having access to techniques that can make this adjustment. Some studies are still classified in binary form to use more accessible tools. To solve this, Generalized Additive Models to ordinal responses - proportional odds model, continuation-ratio model, adjacent-category logistic model - are extended to obtain the spatial odds ratio and map it through them using package VGAM in R-2.7 for Linux. As an illustration, data from an incidence study of occupational accidents were adjusted with an ordinal response variable 'gravity of the accident' in three categories: serious, moderate and light. The analysis has found areas of increased risk for occupational accidents that varied depending on the level of comparison obtained in different fitted models. Some areas had twice the risk compared to the average of the region studied when comparing serious accident with moderate. In parametric analysis were found risk and protection factors. This brings the ability to analyze data by generalized additive models with attachments for epidemiological studies with ordinal response where the spatial odds ratio has to be analyzed.

Managing data.frames with package 'ff'

Jens Oehlschlägel[1, *], Daniel Adler[2]

1. Truecluster.com, Munich
2. Institute for Statistics and Econometrics, Unicersity of Göttingen
* Contact author: Jens_Oehlschlaegel@truecluster.com

Keywords: Large data, databases, column stores

We explain the new capability of package 'ff 1.1' to store large dataframes on disk in class 'ffdf'. ffdf objects have a virtual and a physical component. The virtual component defines a behavior like a standard dataframe, while the physical component can be organized to optimize the ffdf object for different purposes: minimal creation time, quickest column access or quickest row access. Furthemore ffdf can be defined without rownames, with in-RAM rownames or with on-disk rownames using a new ff class 'fffc' for fixed width characters. On a standard notebook we give an online demo of processing an 80 mio row dataframe – size of a German census :-)

References

Adler, D., Gläser, C., Nenadic, O., Oehlschlägel, J. Zucchini, W. (2008-2009) R package ff 2.1.0 "Memory-efficient storage of large atomic vectors and arrays on disk and fast access functions for R" http://cran.at.r-project.org/web/packages/ff/index.html

Adler, Oehlschlägel, Nenadic, Zucchini (2008) Large atomic data in R package 'ff'. Presentation at UseR!2008, statistics department, University of Dortmund

Oehlschlägel, Adler, Nenadic, Zucchini (2008) A first glimpse into 'R.ff'. Presentation at UseR!2008, statistics department, University of Dortmund

R versus SAS in model based drug development

Michael O'Kelly[1*]

1. Quintiles Ireland Ltd, Fairview, Dublin 3, Ireland
* Contact author: mokelly@quintiles.com

Keywords: R, rv, model based drug development, modeling and simulation

Modelling and simulation were required to aid in making decisions about study designs in a drug development program. The R language and its rv package proved a compact tool in implementing a solution, but had important limitations. Results from R were validated using the SAS package. SAS and R are compared as tools for modeling and simulation in the context of model based drug development.

References

Barrett, Jeffrey, et al. (2008). Pharmacometrics: a multidisciplinary field to facilitate critical thinking in drug development and translational research settings. *Journal of Clinical Pharmacology*, 48, 632-649.

Bornkamp, B et al,, (2007) Innovative Approaches for Designing and Analyzing Adaptive Dose-Ranging Trials, *Journal of Biopharmaceutical Statistics*, 17, 965-995

Kerman, J and Gelman A., (2005), Using Random Variables to Manipulate and Summarize Simulations in R, http://www.stat.columbia.edu/~kerman/Software/rv-package-vignette.pdf

Smith, P (2008). Simulation Loops in Splus vs. SAS. http://www.math.umd.edu/~evs/s798c/Handouts/Lec03Pt5D.pdf

Speigelhalter, D et al., (2004) *Bayesian approaches to clinical trials and health-care evaluation*, Chichester: John Wiley and Sons

Wenping (Wendy) Zhang (2007). Using the RAND Function in SAS® for Data Simulation in Clinical Trials, http://www2.sas.com/proceedings/sugi31/198-31.pdf.

Upper contour method in Joint Regression Analysis using R

Teresa A. Oliveira[1,2,*] , **Amlcar Oliveira**[1,2]

1. Open University, Lisbon

2. Center of Statistics and Applications, University of Lisbon

* Contact author: toliveir@univ-ab.pt

Keywords: Upper contour method, joint regression analysis, genotypes, adjusted linear regressions

The use of the upper contour method in Joint Regression Analysis, on the comparison and selection of genotypes, was introduced by Mexia et al.(1997). The method considers adjusted regressions, one per genotype and a range whose limits are respectively the minimum and maximum adjusted environmental indexes.

Among the adjusted regressions in the considered range we enhance those that correspond to maximum production. The genotypes related to these regressions have a particular interest since when multiple comparisons are made, important conclusions can be drawn about the selection of the best genotypes.

Until now the upper contour method was not treated from a computational point of view. Using the R language we suggest a set of procedures, including graphic visualization, in order to simplify the use of that method.

References

Mexia, J.T.; Amaro, A. P., Gusmo, L. & Baeta, J. (1997). Upper contour of a Joint Regression Analysis. J. Genet. And Breed., 51: 253-255.

Seber, G.A.F. & Lee, J.A. (2003). Linear Regression Analysis, 2nd ed.. John Wiley & Sons-New York.

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

Crawley, M. J. (2007). The R Book. John Wiley & Sons-New York. ISBN-13: 978-0-470-51024-7

IRTtool.com

Jeroen Ooms[1]

1. Dept. of Methodology and Statistics, Utrecht University.
* Contact author: jeroenooms@gmail.com

Keywords: Webapplication Interface Internet IRT

IRTtool.com is a webinterface for the cran package ltm by Dimitris Rizopoulos. The application facilitates 1 parameter (Rasch), 2 parameter and 3 parameter IRT modeling, through means of a web-application. Other options include importing and exporting data, plotting information curves and exporting to PDF.

The application is an attempt to make IRT modeling more easily available for applied researchers. However, it is also meant to show the potential of R as a scripting language for statistical web applications. New developments can quickly be made available to a wide audience, and in the near future webbased data management and analysis could become a serious alternative to commercial statistical software.

References

Dimitris Rizopoulos (2006). *ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses*, Journal of Statistical Software, 5, 1–25.

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

`http://www.irttool.com` // `http://www.jeroenooms.com`

Methods and classes for creating *in silico* evolved genetic sequences of HIV.

Christopher E. Ormsby, Santiago Ávila-Ríos, Gustavo Reyes-Terán*.

Department for Research in Infectious Diseases, National Institute of Respiratory Diseases, Mexico City, Mexico. `christopher.ormsby@cieni.org.mx`.
*Corresponding author (`reyesteran@cieni.org.mx`)

Keywords: Simulation, evolution, genomic, phylogeny.

Research on controlling viral infections through vaccines or other strategies relies on methods for measuring the effects of different selective pressures on viruses, and is based on phylogenetic trees which are inferred from the observed data (1).

Simulated genetic evolution has the advantage of being able to know precisely the complete phylogenetic history and all the selective processes that acted on it. Viruses provide an excellent opportunity to use these simulations, since they have a relatively small and simple genome. The data generated can then be used for benchmarking viral evolution models.

We designed a model which accounts for the mayor human immunodeficiency virus genetic selective pressures identified to date, namely HLA driven escape, APOBEC induced hypermutation, recombination, antiretroviral therapy and strong bottleneck events during transmission. Additionally, certain positions are prone to reversion to wild type, and others can cause compensatory mutations. The random mutation rate can be applied with any of the existing nucleotide substitution models. The mutation rate is biased by finite-state reported associations that are applied stochastically at each node. These biases (selective pressures) are read externally from tabular data.

The generated viruses are stored as an object (of novel class `virolver)`, with only substitutions from the parent registered, which avoids large memory usage and allows more efficient iterations. Different components of the class define the virus as a founder (i.e. was transmitted from a different subject) or a quasispecies in the individual, HLA subject composition and parent node.

Output is achieved by methods that produce a `phylo` object (used by the CRAN package `ape` (2)), and/or create FASTA format sequences either of the leaves or of all the tips.

References

1. *Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag.* **Carlson JM, Brumme ZL, Rousseau CM, Brumme CJ, Matthews P, Kadie C, Mullins JI,Walker BD, Harrigan PR, Goulder PJ, Heckerman D.** 4, 2008, PLoS Comput Biol, Vol. 11, p. e1000225.

2. **Paradis, E.** *Analysis of Phylogenetics and Evolution with R.* New York : Springer, 2006.

EURACE Data Visualisation and Analysis Tool with R

Bülent Özel[1,2,*], Vehbi Sinan Tunalıoğlu[1,2], Mehmet Gençer[1,2], Kaan Erkan[2]

1. İstanbul Bilgi University, Computer Science Department, İstanbul, Turkey
2. National Research Institute of Electronics and Cryptology, Kocaeli, Turkey

* Contact author: bulento@bilgi.edu.tr

Keywords: Multi-agent Simulation, Economic Policy Design, Time Series Analysis, RPy 2, R Graphics.

EURACE is an agent-based software platform for European economic policy design with heterogeneous interacting agents. The project is funded by 6th framework programme of the European Commission and runs until September 2009. The EURACE framework takes a bottom up approach to economic modelling and simulation. The bottom up approach requires frequent experimenting on economic policies which require analyses and visualisation of interaction patterns of millions of agents and tracking of emerging economic variables. This work presents the developed tool which aims to serve analyses and visualisations. In particular, we will demonstrate how R analysis and visualization modules are integrated and adopted as the key constituent of the developing workspace(VisGUI) given in Figure 1.
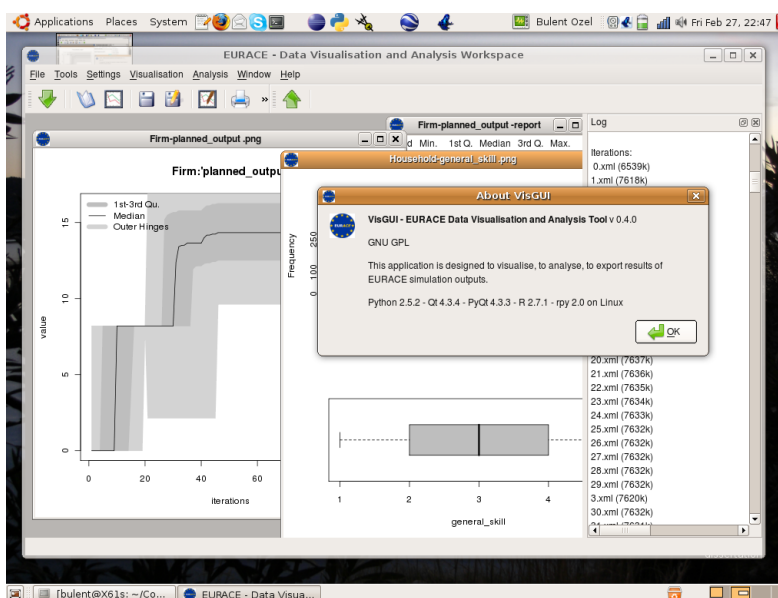

Figure 1: VisGUI Screenshot

VisGUI is an advanced GUI workspace, where policy makers can import, visualize, analyze, edit and export simulation results and reports. It is a platform independent application. It is being implemented using Python2.5 and Qt4. All distributional statistics, and time series analysis and inference analysis are being performed via RPy2. RPy has provided an efficient and practical Python interface to the R Programming Language. In addition, the application allows the user both to enter and execute R scripts in order to generate custom plots or edit existing plots and do perform more specific time series analysis; and to plug in separately developed analysis and visualization modules.

References

EURACE Consortium (2008). The EURACE Project Website,
http://www.eurace.org/.

UEKAE Team (2009). EURACE Project TÜBİTAK/UEKAE Unit Website,
http://eurace.cs.bilgi.edu.tr/.

PEAKPLOT: Visualizing Fragmented Peptide Mass Spectra in Proteomics

Christian Panse[1,*] **Bertran Gerrits**[1] **Ralph Schlapbach**[1]

1. Functional Genomics Center Zurich (FGCZ)

* Contact author: `cp@fgcz.ethz.ch`

Keywords: Proteomics, Mass Spectrometry, Visualization

The goal of proteomics research in general is to identify and quantify proteins on a defined biochemical state. Mass spectrometry is the method of choice for protein identification and post-translational modification assignment in proteomic studies [Roepstorff, P. and Fohlman, J. (1984)]. Due to the advent of accurate and fast sampling mass spectrometers, proteomic experiments often contain thousands of peptide fragmentation spectra. Although it is commonly accepted that no manual validation of individual spectra in such experiments is feasible, annotated spectra of the peptides assignments with their modifications are required for publication and reviewing purposes. Hand-drawn approaches as shown in Figure 1(a) are effective and attractive visualizations. However, their production is very time consuming. The routines provided by [Matrix Science (2009)] are limited by bitmap graphics by the GraphGD library and difficult to customize. Here, we demonstrate a software package called *peakplot*. The software *peakplot* labels the spectra from a peptide sequence assignment by the Mascot search algorithm [Matrix Science (2009)] retrospectively with the appropriate fragment ion labels. The software can can be used either via an easy-to-use web interface or a command line version (for advanced users). We also demonstrate how we embedded our application into existing LIMS (Laboratory Information Management System) infrastructure at our research center. The peakplot application consists of two steps. First, the Mascot Server result file is parsed and the data are processed for the subsequent visualization step. Furthermore, our aim was to provide 'first' visualization functions, which can be easily extended upon individual demands. The difficulty on the 'large-scale, high-throughput' automatic labelling is to avoid overlapping of the labels on each plot. Our heuristic tries to solve this problem by determining the importance of each putative label and drawing only the most important of them. *Peakplot* greatly facilitates the visualization of peptide fragmentation spectra and improves quality assessment of modification sites such as phosphorylation. Availability: `http://fgcz-peakplot.uzh.ch/`
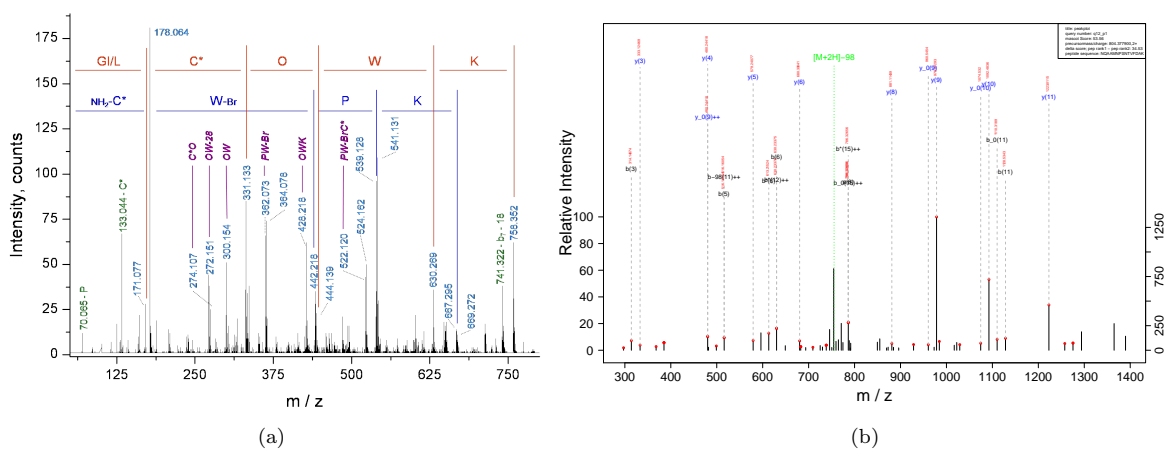

(a)

(b)

Figure 1: The picture displays two labeled mass spectra of two different peptides. (a) hand-drawn by an author, (b) generated from the proposed *peakplot* application.

References

Matrix Science (2009). Mascot Server,
   `http://www.matrixscience.com/server.html`.

Roepstorff, P. and Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom*, 11:601.

Composing HTML documents with hwriter

Gregoire Pau[1]**, Wolfgang Huber**[1]

1. EMBL/European Bioinformatics Institute, Genome Campus, Cambridge, UK

* Contact author: gregoire.pau@ebi.ac.uk

Keywords: HTML, reporting, CSS.

Abstract

HTML documents are structured documents made of diverse elements such as paragraphs, sections, columns, figures and tables organized in a hierarchical layout. The structure of an HTML document is represented by a tree where nodes contain the formatting information and leaves the data to be presented (text, numbers, images).

There are several tools for exporting data from R into HTML documents. The package **R2HTML** is able to render a large diversity of R objects in HTML but does not support combining them efficiently in a structured layout and has a complex syntax. On the other hand, the package **xtable** can render R matrices with simple commands but cannot combine HTML elements and lacks formatting options.

The package `hwriter` allows rendering R objects in HTML and combining resulting elements in a structured layout. It uses a simple syntax, supports extensive formatting options (CSS, Javascript) and takes full advantage of the ellipsis '...' argument and R vector recycling rules. Comprehensive documentation and examples of **hwriter** are generated by running the command `example(hwriter)`, which creates the package web page located at `http://www.ebi.ac.uk/~gpau/hwriter`.

Examples

```
> cap=hwrite(c('Plantae','Monocot','Iris'),
  bgcolor='grey')
> print(cap)
```

```
<table border="1"><tr><td bgcolor="grey">
Plantae</td><td bgcolor="grey">Monocots
</td><td bgcolor="grey">Iris</td></tr>
</table>
```

```
> img=hwriteImage('iris3.jpg')
> hwrite(c(img,cap), 'doc.html', dim=c(2,1))
```



```
> p=openPage('doc.html')
> hwrite('Iris flowers', p, br=T)
> hwriteImage(c('iris1.jpg','iris3.jpg'), p)
> hwrite(iris[1:2, 1:2], p,
  row.bgcolor='#e3dcee')
> closePage(p)
```

Dimensional reduction and clustering of class A G-protein-coupled receptors

Julien Pelé[*] and Marie Chabbert

UMR CNRS 6214 - INSERM U771, Faculté de médecine d'Angers, 49045 ANGERS, FRANCE

* Contact author: julien.pele@etud.univ-angers.fr

Keywords: Dimensional reduction, Clustering, Biological objects

In protein families, amino acid sequences reflect the evolutionary history that led to biological diversity. This available information remains to be unveiled from multiple sequence alignments. Clustering sequences into groups of similar features can provide clues to the evolutionary relationship between them. A solution to cluster objects is to generate a space by dimensional reduction. Objects correspond to points whose mutual distances depend on the metric.

To carry out our project, we used the programming language Perl and the statistical environment of R. We analyzed class A G-protein-coupled receptors from five species. Distance matrices were generated from multiple sequence alignments by different similarity measures. They were analyzed by two statistic techniques: metric multidimensional scaling (cmdscale in R package **stats**) and principal component analysis (dudi.pca in R package **ade4**). Biological objects were grouped by k-means (Kmeans in R package **amap**). Groups were validated with the **clValid** functions from the R package. The clustering analysis was bootstrapped in order to assess the grouping robustness. Best parameters were determined in relation with biological meaningfulness. Our approach indicates that receptors from different genomes share a similar clustering pattern that might relate to major evolutionary determinants of class A G-protein-coupled receptors.

References

Grishin VN, Grishin NV. (2002). Euclidian space and grouping of biological objects. *Bioinformatics,* 18, 1523-34.

Casari G, Sander C, Valencia A. (1995). A method to predict functional residues in proteins. *Nat Struct Biol.*, 2, 171-8.

Dynamic simulation models – is R powerful enough?

Thomas Petzoldt [a]

The R system is increasingly accepted as one of the standard environments for ecological data analysis and modelling. An increasing collection of packages explicitly developed for ecological applications and a growing number of textbooks that use R to teach ecological modelling (Ellner and Guckenheimer, 2006; Bolker, 2008; Soetaert and Herman, 2009; Stevens, prep) are just an indicator for this trend.

The talk will focus on practical experience with implementing and using dynamic models in R as seen from a biological scientist's perspective. A series of concrete case studies was performed analysing particular aspects of ecological interactions. These are presented together with the applied modelling techniques, e.g. sensitivity analysis, parameter fitting, equilibrium analysis and Monte-Carlo simulation.

The example models are organized in an open collection of published models (package **simecolModels**[1]) that cover the range from teaching demos up to the ecosystem level and from spatially aggregated to spatially resolved, for example:

- individual-based simulations of *Daphnia* population dynamics,

- differential equation models of phytoplankton-zooplankton-interactions and of toxin production,

- solid phase / fixed phase simulations in 1D, developed for organismic drift, that may also be useful for physico-chemical systems, i.e. chromatography.

We show how functions of existing packages (e.g. **stats**, **deSolve**, **simecol**, **FME** and **Sweave**) are combined to organize a consistent work flow from data analysis over modelling until publication. The power and the limits of different R implementations are discussed by considering questions like execution speed and implementation effort. It is demonstrated in which circumstances pure R implementations are sufficient and how C-functions can speed up simulations.

We conclude that R is a highly productive system for the dynamic modeller. We propose that it is time to establish a community around dynamic modelling in R and suggest steps in this direction.

References

Bolker, B. (2008). *Ecological Models and Data in R*. Princeton University Press.

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

Soetaert, K. and Herman, P. M. J. (2009). *A Practical Guide to Ecological Modelling Using R as a Simulation Platform*. Springer.

Stevens, M. H. H. (in prep.). *A Primer of Theoretical Population Ecology with R*. Springer.

---

[a]Technische Universität Dresden, Institute of Hydrobiology, 01062 Dresden, Germany, thomas.petzoldt@tu-dresden.de, http://tu-dresden.de/Members/thomas.petzoldt
[1]http://simecol.r-forge.r-project.org/

Statistical assessment of chromosomal aberrations at the cohort level: the CGHSeg package.

Franck Picard [1*] , **Emilie Lebarbier**[2]**, Baba Thiam**[2]**, Stéphane Robin** [2]

1. UMR 5558 CNRS Univ. Lyon 1, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France
2. UMR 518 AgroParisTech/INRA, 16 rue Claude Bernard, F-75231, Paris, France.
* Contact author: picard@biomserv.univ-lyon1.fr

Keywords: CGH, segmentation, clustering, dynamic programming, EM algorithm

Segmentation methods have been successfully applied to the mapping of chromosomal abnormalities when using CGH microarrays. Most current methods deal with one CGH profile only, and do not integrate multiple arrays, whereas the CGH microarray technology becomes widely used to characterize chromosomal defaults at the cohort level. We present CGHSeg, an R package that is devoted to the analysis of CGH profiles at the individual and at the cohort levels. This package performs segmentation in multiple CGH profiles in the framework of linear models, and multivariate segmentation/clustering for the joint characterization of aberration types (status assignment of regions based on the cohort). Overall, linear models offer a unified framework for the joint analysis of multiple CGH profiles, and we will show how they can be used to link the experience acquired in the field of expression arrays (normalization, experimental design) with array CGH data analysis.

Meaningful representation of multivariate analysis output in R : how to solve the trade-off between amount of information and readability?

Timothée Poisot[1,*]

1. Université Montpellier 2, CNRS, Institut des Sciences de l'Évolution, 34095 Montpellier CEDEX 05, France

* Contact author: timothee.poisot@univ-montp2.fr

Keywords: PCA, graphical representation, `match`

"Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away", said the french writer Antoine de Saint-Exupery. How does it apply to graphics? Multivariate analysis have been used for a long time in ecology, because they offer a convenient way to explore the interactions between variables, or the most important factors structuring your data. However, because the purpose of such analyses is to carry the maximum amount of information, their graphical output could be amazingly difficult to read, and the reader is easily overwhelmed by an excess of information. An increasing number of R packages are dedicated to perform such analyses. However, the "out of the box" graphical output is not always easy to customise, and some users may have difficulties to present graphics just the way they want – by including some elements that are not presents by default, or by removing default elements that are useless in their case.

Using datasets from different fields — ecology, physiology, sociology, ... — I present several exemples of visual representation of the same data, and explain how the trade-off between readability and amount of information could be solved, using several ways to approach data representation. The discussion of each exemple is guided by a few questions : How can I do it? What does it tell about my data? Is it informative enough? The use and readability of colors, grey shades, type and sizes of symbols, are discussed. The point of this talk is to adress a fundamental question : How do I convey the maximum quantity of information in an easily readable graphic?

References (of some datasets used)

Poisot & Desdevises (*in revision*). Putative speciation events in *Lamellodiscus* (Monogenea, Diplectanidae) assessed by a morphometric approach. *Parasitology*.

Poisot, Šimková, Hyřsl & Morand (*in revision*). Consequences of rapid water temperature increase on immunocompetence, somatic condition, and parasitism in the chub (*Leuciscus cephalus*), a freshwater cyprinid fish. *Journal of Fish Biology*.

CEM: A Matching Method for Observational Data in the Social Sciences

Stefano Maria Iacus[1,*] **, Gary King**[2]**, Giuseppe Porro**[3]

1. Department of Economics, Business and Statistics, University of Milan, Italy
2. Institute for Quantitative Social Science, Harvard University, USA
3. Department of Economics and Statistics, University of Trieste, Italy

* Contact author: stefano.iacus@unimi.it

We present an R package that implements a new matching method for causal inference in observational data (Iacus, King, Porro, 2009). Observational data are typically plentiful and common in the social sciences; as such, the main issue is reducing bias and only secondarily to keep the variance low. However, most matching methods seem designed for the opposite task, guaranteeing sample size ex ante (such as by choosing matching solutions of one-to-one) but achieving bias reduction (by reducing imbalance between treated and control groups in pre-treatment covariates) only sometimes and with a required extra ex post verification step.

Matching is a simple, intuitive technique of data preprocessing used to control for some or all of the potentially confounding influence of pretreatment control variables by reducing imbalance between the treated and control groups. After preprocessing in this way, any method of analysis that would have been used without matching can be applied to estimate causal effects. The resulting combination reduces model dependence and generally improves inferences with fewer assumptions.

CEM is a matching method with the property that the maximum imbalance between the treated and control groups is controlled by the user ex ante by clear and explicit choices rather than requiring it to be discovered ex post. With CEM, one can control the imbalance on one variable without affecting the maximum imbalance on any remaining variables. It is extremely easy to understand, teach, and use. Unlike many existing methods, it needs no distributional assumptions and so works with any data types. CEM also works on the original space of covariates and hence does not require the adoption of any distance or statistical model to perform the match, and eliminates the need for a separate prior procedure required for other methods that restrict data to common empirical support. It works well with multiple imputation methods for missing data, can be completely automated, and is extremely fast computationally even with very large data sets. CEM also works well for multicategory treatments, determining blocks in experimental designs, and evaluating extreme counterfactuals.

The package `cem` implements such matching method but introduces also a new tool to measure the imbalance in the whole multidimensional distribution of the data. This new index can be used to compare the solutions of different matching algorithms (and so is not specific to CEM). For a given data set the function `cem` returns a vector of weights, one per observation, which can be used later in any statistical model (i.e. `lm`, `glm`, etc) although the package provides also the `att` function for the estimation of the treatment effect (specifically the average treatment effect on the treated)

The package also introduces a diagnostic tool specifically designed for CEM which clearly shows which variable makes the match harder and a graphical tool to represent the distribution of the treatment effect along different strata of the sample rather than just the average treatment effect.

References

Iacus, S.M., King, G., Porro, G. (2009) Matching for Casual Inference Without Balance Checking
http://gking.harvard.edu/files/abs/cem-abs.shtml

Building Information Dashboards with R

Jim Porzak[1,*]

1. The Generations Network, San Francisco, CA
* Contact author: jporzak@tgn.com

Keywords: dashboard, sparkline, bullet graph, graphics, grid

Information dashboards are potentially an important way to communicate the current state of a complex process. They are widely used in business and marketing to track the fundamental metrics, or "key performance indicators," and to highlight exceptional trends and events; either good or bad. Unfortunately, actual implementations do not always live up to their potential (Few, 2006). Often lack of focus is a problem – both visual and data "clutter" distract from effective communication of key points.

The challenges in building an effective dashboard include

- Integrating data from a variety of data sources.
- Detecting trends and exceptional events.
- Building information rich graphical elements.
- Designing a visually attractive, but uncluttered, page.
- Automating timely refresh.
- Easy modification as understanding of requirements evolve.

R is well suited to help with all of these challenges. Data can be easily integrated from various sources: from databases for the core data though spreadsheets for budget numbers. R's core strength is, of course, analysis and graphics. A number of exceptional time series tools are available. `grid` (Murrell, 2006) provides the base upon which to build a well structured and information rich page. While basic sparklines (Tufte, 2004) are easy to code in grid, the `YaleToolkit` package (Emerson & Green, 2007) has some interesting extensions.

The `dashR` package wraps these elements together into an integrated information dashboard toolkit. It also leverages OpenOffice Draw to visually design the dashboard layout and automatically generate nested `grid` viewports. Branding support eases inclusion of logos and the use of specific colors. A number functions generating graphing elements which have been optimized for dashboard use are included. In particular, bullet graphs (Few, 2008) are a clean replacement for meters and gauges often used in dashboards which take the metaphor too literally.

References

Emmerson, J. W. & Green, W. (2007). *YaleToolkit: Data exploration tools from Yale University.* R package version 3.1. http://cran.r-project.org/

Few, S. (2006). *Information Dashboard Design – The Effective Visual Communication of Data.* O'Reilly Media: Sebastopol, 2006.

Few, S. (2008). Bullet Graph Design Specification. http://www.perceptualedge.com/articles/misc/Bullet_Graph_Design_Spec.pdf

Murrell, P. (2006). *R Graphics.* Chapman & Hall/CRD: Boca Raton, 2006.

Tufte, E. (2004). *Sparklines:theory and practice.* http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001OR

STAR: Spike Train Analysis with R

Christophe Pouzat[1,*] **, Antoine Chaffiol**[2] **, Chong Gu**[3]

1. Cerebral Physiology Laboratory, CNRS UMR 8118, Paris-Descartes University, Paris, France
2. Insect Physiology laboratory, INRA UMR 1272, Versailles, France
3. Department of Statistics, Purdue University, West Lafayette, USA

* Contact author: christophe.pouzat@gmail.com

Keywords: Point process; smoothing spline

A central working hypothesis of systems neuroscience is that action potential or spike occurrence times, as opposed to spike waveforms, are the sole information carrier between brain regions. This hypothesis legitimates and leads to the study of spike trains per se. It also encourages the development of models whose goal is to predict the probability of occurrence of a spike at a given time, without necessarily considering the biophysical spike generation mechanisms.

We have adopted the point process / counting process framework to model our spike trains recorded from the first olfactory relay of an insect: the cockroach, *Periplaneta americana*. The key element of this framework is the *conditional intensity* (CI): the instantaneous firing rate of the neuron at time, t, conditioned on potentially every event observed up to t. Despite our growing knowledge of cellular biophysics CI models with a manageable number of parameters are still lacking. We have therefore been lead to nonparametric approaches combining smoothing splines with binomial or Poisson regression models. These efforts have resulted in the STAR (Spike Train Analysis with R; Pouzat, 2009) package which is built "on top" of C. Gu's gss (General Smoothing Spline) package. Both packages are available on CRAN. In addition to nonparametric CI estimation functions, STAR provides numerous goodness of fit tests for CI based spike trains models: the full range of tests developed by Y. Ogata (1988) for earthquakes sequences is implemented together with an original one based on a direct application of Donsker's theorem (Pouzat, Chaffiol, Gu, 2008).

The insight provided by the CI based approach into the function of a "small" neuronal network will be demonstrated by results obtained from spontaneous and odor evoked neuronal activity recordings.

References

Y. Ogata (1988) Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *JASA*, 83, 9-27.

Christophe Pouzat (2009). STAR: Spike Train Analysis with R, `http://cran.at.r-project.org/web/packages/STAR/index.html`.

Christophe Pouzat, Antoine Chaffiol and Chong Gu (2008) Static and dynamic models for spike train analysis: Models, model diagnostics and open-source software. Manuscript available at: `http://sites.google.com/site/spiketrainanalysiswithr/` .

Visualization of Proteomics Data Integrated with KEGG Metabolic Data Using R and Bioconductor

Ermir Qeli[1,*] , Christian Panse[2,*] , Christian Ahrens[1,*]

1. Center for Model Organism Proteomes (C-MOP)
2. Functional Genomics Center Zurich (FGCZ)

* Contact emails: `ermir.qeli@molbio.uzh.ch`,`cp@fgcz.ethz.ch`, `christian.ahrens@molbio.uzh.ch`

Keywords: Proteomics, KEGG, Metabolic Pathways, R, KEGGSOAP, Heatmaps.

Metabolic pathways represent series of chemical reactions occurring within a cell, where each reaction is catalyzed by enzymes. These enzymes are nothing else than proteins. Here we go after the question of how to integrate and visualize quantitative or qualitative proteomics data in the context of metabolic pathways

For illustrating the idea, we focus on the organism *A. thaliana* and we extracted all its possible pathways stored in KEGG (Kanehisa et al, 2000) using KEGGSOAP together with the annotated enzymes present in these pathways.

Furthermore, we extracted from PRIDE tissue specific proteomics data of *A. thaliana* (Bärenfaller et al, 2008) and compared them with the recently acquired pollen proteome (Grobei et al, 2009). From these data, protein spectral counts were calculated, which provide a rough estimation of protein abundance in the respective tissue. For each metabolic pathway, a $m \times n$ matrix with m genes(enzymes) and n tissues with the log values of the respective spectral counts for each gene model for the specific tissue was generated. This matrix is then visualized as a heatmap, where one can compare the enzymatic activity between the different tissues and see which enzymes are active in a specific tissue and inactive elsewhere.

Figure 1 shows only one such heatmap, namely the one related to biosynthesis of phenylalanine, tyrosine and tryptophan. According to literature, the activity in this pathway for pollen should be quite low. The color intensity for the highlighted column for pollen in Figure 1 indicates that this is really the case. Similar pictures are generated for all pathways of *A. thaliana* annotated in KEGG (more than 100). As the scripts written are generic in nature, they can be applied also for other organisms annotated in KEGG where proteomics data is available and are available upon request from the authors.
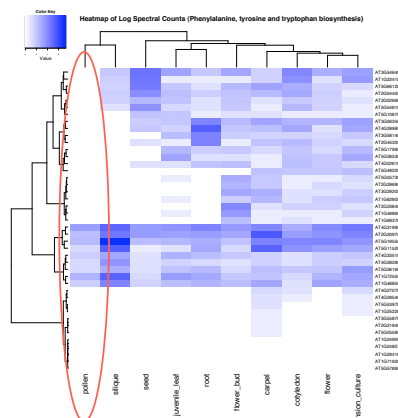

Figure 1: Heatmap showing the enzymatic activity for different *A. Thaliana* tissues in phenylalanine, tyrosine and tryptophan biosynthesis pathway.

References

Kanehisa M, Goto S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes,
   `http://www.genome.jp/kegg/`.

Bärenfaller et al (2008). Genome-Scale Proteomics Reveals Arabidopsis thaliana Gene Models and Proteome Dynamics. *Science*, 320(5878) , 938-41 .

Grobei et al (2009). Deterministic protein inference for shotgun proteomics data of the *Arabidopsis thaliana* mature pollen provides new insights into pollen development and function. *in revision.*

Zhang J, Gentleman R. (2009). Client-side SOAP access KEGG,
   `http://www.bioconductor.org/packages/2.3/bioc/html/KEGGSOAP.html/`.

From relational databases to Linked Data: R for the semantic web

Jose Quesada[1*]

Max Planck Institute, Adaptive Behavior and Cognition (ABC), Berlin Germany,
Quesada@gmail.com

Keywords: Semantic web, RDF, store, RDBMS, modeling

Most of the current structured data in the world lives in relational Data bases (RDB). But statisticians and practitioners sooner or later will find themselves dealing with a different kind of structured representations in the Resource Description Framework (RDF). The logical evolution of the current Web of documents into a Web of Data (and ultimately a Semantic Web) requires mapping of vast quantities of data from RDB to RDF. The conceptual foundations of the relational model and RDF are indeed quite similar, based as they are on set theory and relationships. However, there are important differences between RDBs –or XML- and RDF: in the former the schema often describes a tree, while the latter uses a (more general) graph. I will cover existing options to transform available RDBs into RDF. While there is currently no RDF library, R can interface with different existing frameworks, both for storage (on memory and disk) or processing (reasoning and querying). Then, I will present ways in which R can handle RDF data natively. I will use *package ff* for storage with its C++ core implementing fast memory mapped access to flat files, and *package igraph*, designed to deal with large graphs.

References

W3C Resource Description Framework (RDF)
    http://www.w3.org/RDF/

Adler, D., Gläser, C., Nenadic, O., Oehlschlägel, J., & Zucchini, W. (2008). ff: memory-efficient storage of large atomic vectors and arrays on disk and fast access functions.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. International Journal, Complex Systems, 1695.

Team, R. D. C. (2008). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org.

Automated model generation and selection methods for combinatorially complex biochemical equilibriums

Tom Radivoyevitch[1,2,*]

1. Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106 USA
2. Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, Ohio 44106 USA
* Contact author: txr24@case.edu

Keywords: Enzymes, Rate Laws, Systems Biology, Model Selection

Background: Biochemical equilibrium models can be generated from a full model via hypotheses that some dissociation constants K are infinite and/or that two or more K are equal. For example, in enzyme-substrate-inhibitor (ESI) equilibriums, competitive inhibition models hypothesize that K for ESI is infinite and non-competitive inhibition models hypothesize that K for E_S equals K for EI_S (Fig. 1). In combinatorially complex systems, the number of plausible protein complexes is large relative to the number of reactants, and far more K infinity and equality hypotheses arise than can be specified by hand. Automated model generation and selection methods are needed for these situations.

Results: Biochemical equilibrium models of ATP-induced ribonucleotide reductase R1 hexamerization were generated via K infinity and equality hypotheses from a full model that included three (s, a and h) ATP binding sites on R1. Assuming, based on the crystal structure of yeast R1 dimers [*PNAS* 2006, **103**, 4022-4027], that the s-site is created at the R1 dimer interface, it is reasonable to assume that R1 oligomer s-sites are always fully occupied (i.e. that oligomers cannot form without full s-site occupancy) and that R1 monomer s-sites are always unoccupied (i.e. that the s-site does not exist in R1 monomers). With ATP and R1 denoted by X and R, respectively, the full spur graph system equations are then

$$0 = [R_T] - [R] - \sum_{i=1}^{2} \frac{[R][X]^i}{K_{RX^i}} - 2\left(\sum_{i=2}^{6} \frac{[R]^2[X]^i}{K_{R^2X^i}}\right) - 4\left(\sum_{i=4}^{12} \frac{[R]^4[X]^i}{K_{R^4X^i}}\right) - 6\left(\sum_{i=6}^{18} \frac{[R]^6[X]^i}{K_{R^6X^i}}\right)$$

$$0 = [X_T] - [X] - \left(\sum_{i=1}^{2} i\frac{[R][X]^i}{K_{RX^i}}\right) - \left(\sum_{i=2}^{6} i\frac{[R]^2[X]^i}{K_{R^2X^i}}\right) - \left(\sum_{i=4}^{12} i\frac{[R]^4[X]^i}{K_{R^4X^i}}\right) - \left(\sum_{i=6}^{18} i\frac{[R]^6[X]^i}{K_{R^6X^i}}\right)$$
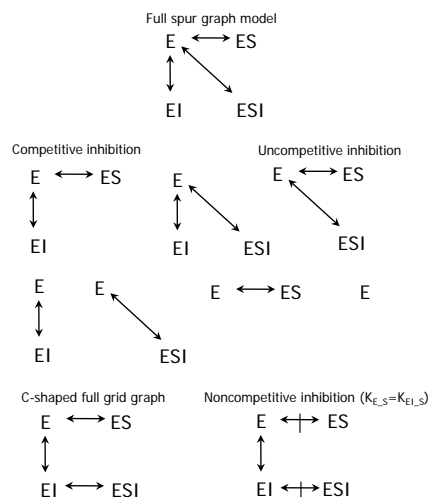
where the T denotes totals and a lack thereof denotes free concentrations. The number of complexes represented is thus $2 + 5 + 9 + 13 = 29$ and this implies $2^{29} = {\sim}500$ million spur models. Not all of these models need to be fitted, however, as one can first fit the 29 single edge models, then the $\binom{29}{2} = 406$ two edge models, then the $\binom{29}{3} = 3654$ three parameter models, etc., stopping once the lowest AIC of the current batch is greater than the lowest AIC of the previous batch. Using this approach to analyze recent dynamic light scattering data [*Biochemistry* 2002, **41**, 462-474], assuming h-sites are filled only after all of the a-sites are filled (and that, in oligomers, these are filled only after all of the s-sites are filled), Figure 2 shows that the best models (those with the lowest Akaike Information Criterion) do not support the existence of an h-site.
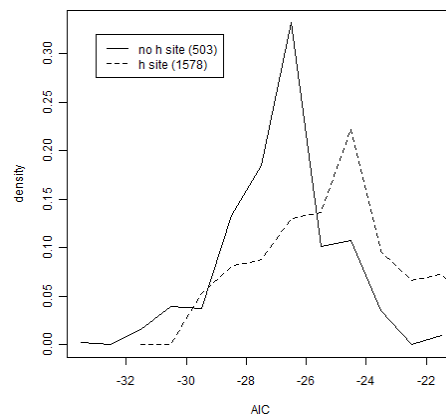
Conclusions: Automated model space generation and analysis methods for combinatorially complex biochemical equilibriums in which the number of models is too large to enumerate by hand, can be realized. Such methods let data speak. They are important because they can lead to inferences that might otherwise be missed.

Figure 1. ESI models/graphs. The full spur graph at the top generates the seven models/graphs below it via hypotheses taken one at a time, two at a time, etc, that dissociation constants are infinite. The C-shaped grid graph is a data-fitting equivalent of the full spur graph. It is important because it generates the non-competitive inhibition model where parallel edges (K_d's) are equal.

Figure 2. Normalized densities of models with SSEs less than twice the minimum SSE (legend indicates model numbers). Though occupied h-site models outnumber unoccupied h-site models 3 to 1, the latter make up 28 of the top 30 models and all of the top 5 models.

References

Radivoyevitch (2008). Equilibrium model selection: dTTP induced R1 dimerization. *BMC Systems Biology*, **2**, 15
Radivoyevitch (2008). *Equilibrium Model Selection*, User!2008 (Dortmund, Germany), July 2008, p 151
Radivoyevitch (2009). *Combinatorially Complex Equilibrium Model Selection*,
    http://epbi-radivot.cwru.edu/ccems/overview.html

Analysis of deep sequencing data to study tumor biology

Wolfgang Raffelsberger[1,*] , Nicodème Paul[1] and Olivier Poch[1] FirstNameA LastNameA[1,2,*] , FirstNameB LastNameB[1,3,4]

1. IGBMC, CNRS UMR7104, Laboratoire de Biologie et Génomique Intégratives, 1 r Laurent Fries, 67404 Illkirch-Strasbourg, France
* Contact author: wolfgang.raffelsberger@igbmc.fr

Keywords: bioinformatics, deep sequencing, SNP statistics

The development of massively parallel sequencing-by-synthesis approaches (such as the Illumina-Solexa and the Roche 454 technologies), also known under the name of deep sequencing, has opened the path for many new applications in biology and medical research. Using such technologies single molecules of DNA (or RNA) can be amplified and sequenced individually at very high throughput. This capacity opens new perspectives in tumor biology since cancer cells acquire during tumor growth novel in a heterogeneous manner mutations, deletions and amplifications in their genome. Furthermore, the deep sequencing approach is very promising, since this provides a uniform platform to compare sequence alterations on the levels of genomic DNA and mRNA.

The mapping of sequences produced from deep sequencing experiments and the statistical analysis of the sequence alterations observed (compared to a reference genome) pose new challenges for users of R. Several packages like Biostrings (Pages et al 2009) and ShortRead (Morgan et al 2009, both on Bioconductor, Gentleman et al 2004) have been developed for running the initial steps of data-analysis, however additional functionalities are needed to study and interpret the characteristics of sequence alterations of inhomogeneous starting material as this is common with cancer biopsies. In this context we are developing a new package dedicated to the reliable identification of sub-populations of sequence alterations form longer sequences (Roche 454 technology). Furthermore, we have developed additional functionalities for the direct comparison of sequence alterations on the levels of genomic DNA and mRNA. This package has allowed us gaining more insight to which degree individual tumors represent actually heterogeneous material on preliminary deep sequencing data.

References

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J.(2004) *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol. 2004;5(10):R80 http://www.bioconductor.org

Morgan M, Lawrence M. and Anders S (2009). *ShortRead: Base classes and methods for high-throughput short-read sequencing data.* R package version 1.0.6

Pages H. (2009), Gentleman R., Aboyoun P. and DebRoy S., *Biostrings: String objects representing biological sequences, and matching algorithms.* R package version 2.10.1

Unbiased variance estimates for multiple imputation in R

James L. Reilly[1,*]

1. Department of Statistics, University of Auckland, Auckland, New Zealand
* Contact author: reilly@stat.auckland.ac.nz

Keywords: multiple imputation, estimating equations, variance estimation

A new R package implementing Robins and Wang's (2000) estimating equation approach to multiple imputation is presented. This produces unbiased variance estimates, even with misspecified models and disagreements between the imputer's and analyst's models.

Extensions to handle data from complex surveys, building on Lumley's (2004) survey package, will be discussed, along with an interface to the Zelig package (Imai, King and Lau; 2008).

References

Imai, K., King, G., and Lau, O. (2008) Toward A Common Framework for Statistical Analysis and Development. *Journal of Computational and Graphical Statistics*, 17(4), 892-913.

Lumley, T. (2004) Analysis of complex survey samples. *Journal of Statistical Software* 9(1): 1-19.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87, 113–124.

Workflows for Data Mining in Integrated multi-modal Data of Intracranial Aneurysms using KNIME

Roelof Risselada[1], Christoph M. Friedrich[2,*], Christian Ebeling[2], Roman Klinger[2], Anna Bauer-Mehren[3], Manuel Pastor[3], Maria Cruz Villa[4], Jose M. Pozo[4], Alejandro F. Frangi[4], and Martin Hofmann-Apitius[2]

1. Erasmus MC; Medical Informatics; PO Box 2040, 3000 CA Rotterdam; the Netherlands
2. Fraunhofer Institute for Algorithms and Scientific Computing (SCAI); Department of Bioinformatics; Schloss Birlinghoven; D-53754 Sankt Augustin; Germany
3. Research Unit on Biomedical Informatics (GRIB), IMIM/UPF; C/Dr. Aiguader, 88; 08003 Barcelona; Spain
4. Research Group of Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Universitat Pompeu Fabra, Pg Circumval·lació, 8, 08003 Barcelona; Spain
* Contact author: friedrich@scai.fraunhofer.de

Keywords: Data Mining, Workflow, multi-modal data

Intracranial aneurysms are bulbous expansions of the intracranial vessels that may rupture and lead to subarachnoid haemorrhage, a bleeding in the space lining the brain. This can result in severe disability or death of the affected person. The prediction of the individual rupture risk of a patient based on information from images, haemodynamic simulations, clinical parameters and genetic markers is one of the aims of the European Integrated Project @neurIST. The project developed an architecture [1] which allows the integration of multi-modal data from clinical information systems. Data mining capabilities have been developed in a Knowledge Discovery application suite called @neuLink [2]. Maintenance, re-use of mining strategies and user presentation of data flows is difficult in monolithic mining scripts. To cope with this problem, @neuLink integrates the Konstanz Information Miner (KNIME) [3] as a workflow engine and provides aggregated data mining results based on R scripts and Weka [4]. The advantage of this solution is a better understanding of the workflow, re-usability of data mining strategies and an increased maintainability.

One example sub-workflow is the application of clustering algorithms to find similar aneurysms. This is based on Zernike moments [5] extracted from images of aneurysms. The resulting similarity matrix is depicted in the screenshot.

Several @neurIST partners developed and implemented rupture risk and treatment outcome models as sub-workflows, which are currently being integrated in consensual models.

Acknowledgements

References

1. Rajasekaran, H.; , Iacono, L. L.; Hasselmeyer, P.; Fingberg, J.; Summers, P.; Benkner, S.; Engelbrecht, G.; Arbona, A.; Chiarini, A.; Friedrich, C. M.; Hofmann-Apitius, M.; Kumpf, K.; Moore, B.; Bijlenga, P.; Iavindrasana, J.; Mueller, H.; Hose, R. D.; Dunlop, R., and Frangi, A. *@neurIST – Towards a System Architecture for Advanced Disease Management through Integration of Heterogeneous Data, Computing, and Complex Processing Services;* Proceedings of the 21st IEEE International Symposium on Computer-based Medical Systems*, IEEE,* **2008**, 361-366.
2. Friedrich, C. M.; Dach, H.; Gattermayer, T.; Engelbrecht, G.; Benkner, S., and Hofmann-Apitius, M. *@neuLink: A Service-oriented Application for Biomedical Knowledge Discovery* Proceedings of the HealthGrid 2008, IOS Press*,* **2008**, 165-172
3. Berthold, M. (2009). *KNIME (Konstanz Information Miner)* webpage; last accessed 2009-02-26 http://www.knime.org/
4. Witten, I. H., and Frank, E. *Data Mining; second edition;* Morgan Kaufmann Publishers*,* **2005**
5. Millán, R. D.; Dempere-Marco, L.; Pozo, J. M.; Cebral, J. R., and Frangi, A. F. *Morphological characterization of intracranial aneurysms using 3-D moment invariants.* IEEE Trans Med Imaging*,* **2007***, 26*, 1270-1282

Dose-response modelling using R

Christian Ritz[1,2,*] , **Jens Carl Streibig**[1,3]

1. Faculty of Life Sciences, University of Copenhagen
2. Statistics, Department of Basic Sciences and Environment
3. Crop Science, Department of Agriculture and Ecology

* Contact author: ritz@life.ku.dk

Keywords: logit/probit models, nonlinear regression, quantal response data, self starter functions

The outcome of dose-response experiments is the effect or stimulus on an organism in response to a dose administered. In this context *dose* can refer to any biological, chemical, radioactive stimulus, or to any other tangible stimuli that can be graduated. Major application areas include agriculture, biology, chemistry, medicine, pharmacology, and toxicology. Specific applications range from experiments in hearing and speech science over pulse oxygen saturation modelling to toxicity testing of chemicals within regulatory frameworks.

This presentation provides an overview of the extension package `drc` (on CRAN), which is designed for the statistical analysis of dose-response data that either could be individual curves or several curves considered jointly (eg. binary mixture models). Package development began in 2004 with a narrow focus on pesticide development and ecotoxicology (Ritz & Streibig, 2005). Mainly in response to user feedback and requests, the package has been improved and extended in various ways for the last 5 years. The package provides a general model fitting function, with much the same basic interface and feel as `lm()`, as well as all standard extractor methods, e.g. `anova`, `coef`, `plot`, `predict`, `residuals`, and `summary`. An arsenal of built-in dose-response functions (relying heavily on the concept of self starter functions) comes with the package. In addition, there are several special functions for after-fitting extraction of particular parameters of biological interest such as effect concentrations or doses (e.g. EC50/ED50 and LC50/LD50 values).

More specifically, the presentation will touch upon topics such as:

- unified parametric modelling framework for several data types

- elaborate infrastructure for built-in functions

- analysis of high throughput dose-response data

- simulation from parametric dose-response models (given a specified design)

- visualisation of the results

Some ideas for future developments will also be presented.

References

Ritz C. and Streibig J. C. (2005). Bioassay analysis using **R**. *Journal of Statistical Software*, **12**, Issue 5, 1–22. http://www.jstatsoft.org/v12/i05.

Combination of protein biomarkers

Xavier Robin[1,2,*], Natacha Turck[1], Alexandre Hainard[1], Laszlo Vutskits[3],
Catherine Fouda[1], Nadia Walter[1], Paola Sanchez-Peña[4], Louis Puybasset[4],
Frédérique Lisacek[2], Jean-Charles Sanchez[1], Markus Müller[2]

1. Biomedical Proteomics Research Group, Department of Bioinformatics and Structural Biology, University of Geneva, Switzerland
2. Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland
3. Department of Anesthesiology, Pharmacology and Intensive Care, University Hospital of Geneva, Switzerland
4. Department of Anesthesiology and Critical Care, Pitié-Salpêtrière Teaching Hospital, Assistance Publique-Hôpitaux de Paris and Université Pierre et Marie Curie-Paris 6, France
* Contact author: Xavier.Robin@unige.ch

Keywords: Biomarkers, Multiplexing, Proteomics, ROC Curve

Recent advances in immunoassay-based protein quantitation methods allow the quantification of a large number of proteins in biological fluids such as serum. When these proteins are differentially expressed between two populations, they are called biomarkers. However, biomarkers often show an insufficient discrimination power. We hypothesized that a combination of biomarkers could increase diagnosis or prognosis efficiency.

In order to predict 6-month outcome of patients after an aSAH (an extracerebral hemorrhage) based on a combination of 6 biomarkers and 3 clinical parameters measured at time of admission, we developed a simple threshold-based panel algorithm where thresholds were determined by exhaustive search. We compared it with 5 other combination methods: SVM (kernlab), Linear Models, Generalized Linear Models, Weighted K-Nearest Neighbors (kknn) and Partial Least Square (pls). 10-fold cross-validation was used to avoid overfitting. In order to get a statistical measure of the differences between the ROC Curve, we used the methods developed by Hanley and McNeil (1983) and DeLong et al. (1988) for comparing ROC Curves, and compared them to bootstrapping methods. Partial Area under the ROC Curve (pAUC) allowed us to focus on 90-100% specificity predictions. We tested this approach on a cohort of 112 patients. All the computations were performed in R.

The best individual biomarker displayed a pAUC of 65% of optimal value (90% specificity for 40% sensitivity). The best clinical measurement had the same pAUC with a specificity of 94% and a sensitivity of 45%. Two combination methods performed slightly better: the threshold-based algorithm with a pAUC of 68% (93% specificity and 55% sensitivity) and SVM with 66% pAUC (90% specificity and 53% sensitivity). That result means the threshold-based test is able to detect 55% of the poor outcome patients while raising only 7% of false positives.

Even though the improvement seems small, detecting 10% more poor-outcome cases without increasing the false alarm rate is of prime importance for physicians and for the management of poor-outcome patients, because no tool specific to prognosis is currently available. This method allowed us to provide a quantitative measure of the differences and to compare the methods between them as well as with individual markers. The threshold-based algorithm was the best predictor of aSAH 6-month outcome. It performed slightly better than individual markers; however cross-validation was applied only to combinations and individual markers performance might be overestimated. We will use the statistical tests described above to validate these results.

References

Hanley J. A. and McNeil B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148, 839–843.

DeLong E. R. et al. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44, 837–845.

McClish (1989). Analyzing a Portion of the ROC Curve. *Medical Decision Making*, 9, 190–195.

Giving syphilis to friends: Using social network methods to study the spread and control of syphilis in Baltimore

Janet E. Rosenbaum[1,2,*]**, Anne Rompalo**[1,2]

1. Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland.
2. Johns Hopkins Sexually Transmitted Disease Center, Baltimore, Maryland.

* Contact author: janet@post.harvard.edu

Keywords: Social Network Analysis, Social Sciences, Public Health, Syphilis.

The United States has declared syphilis elimination a national health priority, but despite low national prevalence, pockets of syphilis persist, including Baltimore which leads the nation in prevalence of primary and secondary syphilis. Currently sexual contact tracing is used to control syphilis: patients with early infectious syphilis are asked to name their sexual partners during their likely window of infectivity so that partners may be notified, tested, and treated. It is known, however, that sexually-transmitted disease (STD) patients generally do not name all of their sexual contacts. Failure to name contacts leaves likely syphilis cases unidentified, untreated, and available to contribute to the further spread of the disease. Past studies of STD patients have found that they often choose sexual partners from within their social, personal, and drug networks. Exploring patients' broader social contexts may be an effective method of discovering unnamed sexual partners and thus additional cases of syphilis.

In this study we test the power of social and personal network analysis in explaining syphilis transmission. These findings can be applied to evaluate a new method of syphilis control by comparing the effectiveness of detecting early infectious syphilis cases by screening social network members of syphilis index cases compared to standard sexual partner notification techniques.

The social network was constructed using administrative data from the two Baltimore STD clinics and the Baltimore City Health Department. In the usual contact-tracing procedure, patients with incident (newly acquired) syphilis are asked at their STD clinic visit to name sexual contacts from their likely window of infectivity for partner notification purposes. For this study, approximately 500 patients with newly acquired syphilis were interviewed privately and asked to name members of their social and drug networks and given a self-administered questionnaire. Named network members were interviewed using the same questionnaire as the index case, and tested for syphilis. Blood and lesion specimens were used for genetic analysis of syphilis strains to confirm connections between reported contacts.

We map the sex and friendship networks of individuals with incident syphilis who have acquired and possibly transmitted syphilis through sexual behavior. We examine and compare the centrality of drug use behavior among the social and sexual networks of individuals with incident syphilis to find the importance of drug use in spreading syphilis. Finally, we compare the number of new syphilis cases identified in social contacts of the index cases versus sexual contacts. The data is analyzed using routines from the Statnet project.

References

Handcock MS, Hunter DR, Butts CT, Goodreau SM, and Morris M (2003). statnet: Software tools for the Statistical Modeling of Network Data.
URL: http://www.statnetproject.org.

The DiceKriging package: kriging-based metamodeling and optimization for computer experiments

Olivier Roustant[1], David Ginsbourger[2], Yves Deville[3]

1. Ecole des Mines de St-Etienne (France). Contact author: roustant@emse.fr
2. Université de Neuchâtel (Switzerland)
3. Statistical Consultant (France)

Keywords: Computer Experiments, Gaussian Processes, Spatial Statistics, Kriging, Global Optimization

The package DiceKriging has been developed for analyses involving computer intensive experiments as met in various industrial contexts (automotive, aeronautics, nuclear, ...) where numerical simulations are required. Kriging stands for the well-known model from spatial statistics, which has recently gained popularity in the computer experiments community, and is sometimes referred to as Gaussian Process Regression (Rasmussen, Williams, 2006). DICE (*Deep Inside Computer Experiments*) is the name of a consortium within the frame of which the works were conducted, joining the industrial partners Armines, Renault, EDF, IRSN, Onera and Total S.A (`www.dice-consortium.fr`). DICE members have shared a growing interest for R, due to its efficiency in including the most recent statistical methods, and its ease of use. The development was both guided by applications and based on published efficient algorithms. DiceKriging was tested on toy & industrial case studies in dimensions 2 to more than 30. It was found to be a valuable complement to other existing packages like *tgp* or *mlegp*. The package contents is the following:

1. Kriging models for deterministic simulators, stochastic simulators with unknown homogenous noise and stochastic simulators with controllable heteroscedastic noise:

   - Maximum Likelihood (ML) estimation of unknown parameters, with possible penalization.
   - Cross Validation (Leave-One-Out, k-fold CV),
   - Prediction: Simple, Ordinary, and Universal Krigings with many kinds of trends,
   - Covariance kernels: Anisotropic Gaussian, Power Exp., and Matérn with $\nu = 3/2$ or $\nu = 5/2$.

2. Kriging-based black-box optimization:

   - Efficient Global Optimization (EGO) algorithm,
   - Several parallelized versions of the EGO algorithm for synchronous distributed computing.

Among the innovations proposed in the package, much effort was devoted to the efficient use of optimization routines, both for the problem of ML estimation and within black-box optimization algorithms.

- To address the known difficulties in ML estimation (multimodality and possible numerical instability) the algorithm proposed by Park and Baek (2001) has been implemented. This allows the use of analytic gradient in the optimisation either with the classical BFGS (optim{stats}) or with the hybrid evolutionary genoud{rgenoud}. The available covariance structures can include a "nugget effect" for stochastic simulators and their list can be extended in the future.

- The EGO algorithm was proposed by (Jones, Schonlau, Welch, 1998). Additionally to EGO, a class of variants for synchronous parallel computing is proposed. In both cases, the maximization of the highly multimodal Expected Improvement is performed using the hybrid evolutionary strategy *genoud*.

- The package also includes Kriging with known parameters, which can be useful for Bayesian Kriging.

The package documentation includes some validation tests (as ML estimation of simulated processes), classical analytical test functions, and several case studies proposed by the members of the DICE Consortium.

References

Jones D.R., Schonlau M. and Welch W.J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global optimization*, 13, 455–492

Park J-S, Baek J. (2001). Efficient computation of maximum likelihood estimators in a spatial linear model with power exponential covariogram. *Computer Geosciences*, 27, 1–7.

Rasmussen C.E., Williams C.K.I. (2006). *Gaussian Processes for Machine Learning*, the MIT Press. `www.GaussianProcess.org/gpml`.

A Tale of Two Theories
Reconciling random matrix theory and shrinkage estimation as methods for covariance matrix estimation

Brian Lee Yung Rowe[1]

1. Bank of America Merrill Lynch
* b_rowe@ml.com

Keywords: random matrix theory, shrinkage estimation

Estimation error in asset returns covariance matrices has plagued the portfolio optimization process ever since Markowitz first proposed the mean-variance approach. To combat this problem, two competing theories have developed to eliminate (or more appropriately, reduce) this estimation error: random matrix theory and shrinkage estimators. While attempting to solve the same problem, the approaches are substantially different. Random matrix theory states that a truly random matrix has a characteristic limit distribution of its eigenvalues. This distribution can be used as a null hypothesis to remove, by scaling to a lower bound, all such eigenvalues associated with this idiosyncratic noise in the sample covariance matrix. In contrast, shrinkage estimation leverages the central limit theorem to shrink covariances toward a biased covariance matrix that lacks estimation error (and hence better represents the unobserved true covariance matrix).

This paper explores these two competing approaches using an R package developed by the author to estimate a covariance matrix of asset returns. This analysis attempts to identify the constraints under which each method best performs and whether there is space to reconcile the two approaches into a single unified framework. In order to compare the performance of the theories, empirical and generated data is used to compute standard portfolio performance metrics. The package itself contains methods for calculating the theoretical Marcenko-Pastur eigenvalue distributions and functions for fitting empirical eigenvalue distributions to the closest Marcenko-Pastur curve. Implementations are also provided for shrinkage using a variety of shrinkage targets, including a single factor model, a constant correlation model, and a multi-factor model.

References

Jim Gatheral (2008). Random Matrix Theory and Covariance Estimation. NYU Courant Institute Algorithmic Trading Conference, October 2008.

Olivier Ledoit and Michael Wolf (2001). Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portofolio Selection. *Economics Working Papers 586*, Department of Economics and Business, Universitat Pompeu Fabra.

Olivier Ledoit and Michael Wolf (2003). Honey, I Shrunk the Sample Covariance Matrix. *UPF Economics and Business Working Paper No. 691*.

Marc Potters, Jean-Philippe Bouchaud, Laurent Laloux (2005). Financial Applications of Random Matrix Theory: Old Laces and New Pieces. *Science & Finance (CFM) working paper archive 500058*, Science & Finance, Capital Fund Management.

Ⓡ–Package robKalman — Ⓡ. Kalman's revenge … or Ⓡobustness for Kalman Filtering Ⓡevisited

Peter Ruckdeschel[1,*] , Bernhard Spangl[2]

1. Fraunhofer ITWM, Abteilung Finanzmathematik, Fraunhofer-Platz 1, D-67663 Kaiserslautern, Germany
2. Universität für Bodenkultur, Gregor-Mendel-Straße 33, A-1180 Wien, Austria
* Contact author: peter.ruckdeschel@itwm.fraunhofer.de

Keywords: Robustness, Kalman filter, S4-classes

Building up on talks on this issue at UseR conferences 2006 and 2008, we report on progress made in the development of package robKalman. Focus of this talk will be

(a) an OOP-layer of S4–classes and –methods on top of the already existing functions which allows for quite flexible "generic" user interfaces

(b) enhanced functionality covering

  – (robust) Kalman smoothing
  – (robust) estimation of (hyper–)parameters
  – IO-robustness, i.e., enhanced tracking features

(c) interfacing functions to other packages providing infrastructure for (multivariate) time series and implementations to state space models and the (classical) Kalman Filter/Smoother.

(d) report on some experience with collaborative package development under r-forge

References

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

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

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

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

R-Forge Administration and Development Team (2008) : *R-Forge User's Manual,* BETA. SVN revision: 47, August, 12 2008.
http://r-forge.r-project.org/R-Forge_Manual.pdf

Shumway, R.H. and Stoffer, D.S. (1982) : *An approach to time series smoothing and forecasting using the EM algorithm.* Journal of Time Series Analysis, 3, 253-264.

Spangl, B. (2008) : *On Robust Spectral Density Estimation.* PhD Thesis at Technical University, Vienna.

Estimating Markov-Switching Regression Models in R: An application to model energy price in Spain

Sara Fontdecaba[1] , **Jose A. Sanchez-Espigares**[1] , **Pilar Muñoz**[1]

1. Department of Statistics and Operational Research. Universitat Politecnica de Catalunya

* Contact author: sara.fontdecaba@upc.edu

Markov switching regression models can be used to study heterogeneous populations that depend on covariates observed over time. The model formulation involves a mixture of regressions models with a Markov chain defining the mixing distribution. In each instant the time series is assumed to be under a determined regime. This unobserved process, that governs the evolution of the series, defines a state variable related to the Markov chain process and is characterized by a matrix consisting on the probabilities of transition between states. Applications of Markovian Switching models can be found in several fields including, for instance, ecology, engineering and econometrics.

Following Hamilton(1989), we have implemented a set of R functions in order to explain time series according to a switching regression model. Estimation of parameters defining the model and imputation of the unobserved state process is performed under the Maximum Likelihood criterion. The implemented routines deal with Ordinary Least Squares regression to relate the response variable to the explanatory variables, although this model can be of different and more complicated types (i.e. Auto-Regressive models or Transfer Functions). The equations from the mixture of models can include some regressors with switching effect (different coefficient for each state) and others with common coefficients for all states. The last ones indicate a constant relationship not depending on the current regime.

Due to the large number of parameters to estimate, standard non-linear optimization procedures can be unstable. To avoid this problem, an EM approach has been included that guarantees a more robust approximation to the global optimum.

An illustration of the use of this routines is presented to model the evolution of the energy price in Spain between 2002 and 2008, according to the demand level, raw material prices (oil, coal and gas) and finance indicators (Ibex35 and exchange rate EUR/USD). A Markovian Switching model with two states has been considered with all regressors with switching effect. The R functions provide estimation of parameters and probabilities of being in each state along the series.

References

Hamilton, J.D. (1989). A new Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57:2,357–384.

Using Qt for GUI tasks in R

Deepayan Sarkar[1,*]

1. Fred Hutchinson Cancer Research Center * Contact author: deepayan.sarkar@r-project.org

Keywords: GUI, Qt, user interfaces, visualization and graphics

Although R is primarily a command-line tool, it provides facilities for interfacing with various GUI toolkits. Qt is a powerful, robust, and cross-platform GUI toolkit licensed under the GPL. Full use of Qt from R requires bindings to Qt, which are not yet available. However, it is relatively simple to write custom tools that use Qt for specific purposes. In this talk, I will give some examples of such tools, including a R graphics device implemented using Qt, an alternative R help browser with full-text search capabilities, a simple data import wizard, and a basic but functional high-level graphics system not unlike lattice, that is completely independent of the R graphics engine.

References

Qt software, `http://www.qtsoftware.com/`.

PLS in Chemometrics with R

Majid Sarmad[1,2,*]

1. Fedowsi University of Mashhad
2. Faculty of Maths. Sciences
* Contact author: sarmad@um.ac.ir

Keywords: NIR, Partial Least Squares

Partial least squares can be used when the number of observations are more than the number of variables. Particularly, in chemometrics, when a huge number of NIR (near infrared) wavelengths are passed through a substance, they are studying on a model between the output of wavelengths and one or more characteristics in the substance. Modeling cannot be made using a large number of initial experiments as it may cost a lot.

pls and some other packages in R do PLS modeling; including PCR, Cross Validation and PLS. In a real example, using R, it will be tried to model PH in Kiwi fruit by 601 NIR wavelengths.

R package gcExplorer: graphical and inferential exploration of cluster solutions

Theresa Scharl[1,2,*] **, Friedrich Leisch**[3]

1. Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria
2. Department of Biotechnology, University of Natural Resources and Applied Life Sciences, Vienna, Muthgasse 18, A-1190 Vienna, Austria
3. Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany
* Contact author: theresa.scharl@ci.tuwien.ac.at

Cluster analysis is commonly applied to microarray data in order to find groups of co–expressed genes where cluster algorithms with the ability to visualize the resulting cluster objects (e.g., a dendrogram for hierarchical clustering) are usually preferred. The display of cluster solutions particularly for a large number of clusters is very important in exploratory data analysis. It gives practitioners an idea of the relationships between segments of a partition and allows to interpret the cluster results. Neighborhood graphs (Leisch, 2006) can be used for visual assessment of the cluster structure of centroid–based cluster solutions. In a neighborhood graph each node represents a cluster and two nodes are connected if there exist data points that have the two corresponding centroids as closest and second closest centroid.

In this work we present new visualization methods based on the neighborhood graph. For node representation different plot symbols visualizing single clusters are used allowing a quick overview of the data. On the one hand the corresponding data points themselves can be visualized using for example line diagrams for gene expression over time. On the other hand node symbols like pie charts can be used to visualize further properties of the clusters like association to functional groups under study. Finally the neighborhood graph can be used for the validation of a cluster solution, e.g., by testing the relationship between a clustering and a priori information about gene functions. All visualization methods and test procedures used are implemented in R package **gcExplorer** (Scharl and Leisch, 2009) which is now available on CRAN. The grid–based node symbols are implemented in R package **symbols** (`http://r-forge.r-project.org/projects/symbols/`).

References

Friedrich Leisch (2006). A Toolbox for K-Centroids Cluster Analysis. *Computational Statistics and Data Analysis*, 51(2): 526–544.

Theresa Scharl and Friedrich Leisch (2009). gcExplorer: Interactive Exploration of Gene Clusters. *Bioinformatics*, doi: 10.1093/bioinformatics/btp099.

Inference, aggregation and graphics for top-k rank lists

Michael G. Schimek[1,2,*] , **Eva Budinska**[3],
Shili Lin[4], **Alena Mysickova**[5,2]

1. Medical University of Graz, Austria
2. Danube University Krems, Austria
3. Masaryk University Brno, Czech Republic
4. Ohio State University, USA
5. Humboldt University Berlin, Germany

* Contact author: michael.schimek@medunigraz.at

Keywords: Graphics Tool, Moderate Deviation, Ordered List, Rank Aggregation, `TopkLists` Package

Lists of common distinct objects in rank order are typical for various fields of application. The rank of an object belonging to the set of interest in a certain list indicates its respective position among all other objects. The rank position might be due to a measure of strength of evidence, to a consumer preference, or to an assessment either based on expert knowledge or a technical device. Let us assume that the rank assignment in each list is independent of the assignment in the other lists.

Let us have ℓ such lists τ_j $(j = 1, 2, \ldots, \ell)$ assigning rank positions to the same set of objects. The ranking is from 1 to N, without ties. Our goal is to identify a subset of objects that is characterized by high conformity across the lists. This implies that there is similarity between the rankings which can be evaluated by a distance measure d (a permutation metric) such as Kendall's τ or Spearman's footrule. In practice, because of truncated rank lists and incomplete rankings of objects in some of the lists caused by missing assignments, we need to penalize these measures accordingly. Moreover, in most applications, especially for large or huge numbers N of objects, it is not likely that consensus prevails, thus only the top-ranked elements are relevant. For the remainder objects their ordering is more or less at random. This is not only true for surveys of consumer preferences but also appropriate for search tasks in the Web and data integration in the field of biotechnology. In many instances we observe a general decrease, not necessarily monotone, of the probability for consensus rankings with increasing distance from the top rank position. Typically there is reasonable conformity in the rankings for the first, say k, elements of the lists, motivating the notion of *top-k rank lists*.

List aggregation by means of brute force is limited to the situation where both N and ℓ are unrealistically small, and k is known (e.g. 'ground truth' in Web search engines). Here our aim is to solve this computational problem for a realistic setting, firstly, via an algorithm for the selection of the $\hat{k}$'s for all $(\ell^2 - \ell)/2$ possible pairs of lists τ_j, secondly, via a graphical tool monitoring the aggregation process of the thus obtained top-k rank list information, and thirdly, via an algorithm for the calculation of a set of objects characterized by rankings of high conformity across the lists up to some global index $\bar{k}$. For the first task we take advantage of a moderate deviation-based inference procedure for random degeneration in paired rank lists (Hall and Schimek, 2009). The graphical tool is a newly developed type of heat map simultaneously displaying three-dimensional information representing the dynamics of the aggregation process based on the input from the inference procedure. For the last task an Order Explicit Algorithm (OEA) is combined with cross-entropy Monte Carlo (CEMC), as outlined in Lin and Ding (2009). For the same input lists the aggregation result does not only depend on the index $\bar{k}$ but also on the chosen distance measure and adopted concept for the handling of partial lists (incomplete sets of objects), apart from necessary tuning parameters. Therefore a graphical monitoring tool is much desirable.

Although the discussed methodology is quite general in terms of application, we take a special interest in the meta analysis of microarray experiments. Hence we apply the above algorithms, which we are implementing in the `R` package `TopkLists`, to both artificial and real gene expression data. `TopkLists` is based on the most recent algorithmic developments, allows for N in the magnitude of thousands, and will serve as universal tool for the objective identification of informative objects (e.g. genes) conforming across rank lists.

References

Hall P. and Schimek M.G. (2009). Moderate deviation-based inference for random degeneration in paired rank lists. Preprint.

Lin S. and Ding J. (2009). Integration of ranked lists via Cross Entropy Monte Carlo with applications to mRNA and microRNA studies. *Biometrics*, to appear.

MGARCH: An R Package for Fitting Multivariate GARCH Models

Harald Schmidbauer[1,*] , Vehbi Sinan Tunalıoğlu[1,] , Angi Rösch[2,***]**

1. Istanbul Bilgi University, Istanbul, Turkey
2. FOM University of Applied Sciences, Study Centres Munich/Germany & Taian/China
* harald@hs-stat.com
** vst@vsthost.com
*** angi.r@t-online.de

Multivariate GARCH processes have been found useful in the analysis of volatility spillovers between several heteroskedastic time series. This phenomenon can frequently be observed in the behaviour of series of returns on stocks.

Our package can fit several MGARCH specifications to data, such as the MGARCH-BEKK and the DCC models. The package's functionality includes easy model diagnostics and model simulation, as well as real-world examples from finance. A comprehensive manual is also available.

References

Bauwens, L., Laurent, S. & Rombouts, J.V.K. (2003). *Multivariate GARCH models: a survey*. CORE Discussion Paper 2003/31, Université Catholique de Louvain.
`http://www.core.ucl.ac.be/econometrics/Bauwens/Papers`.

Tsay, Ruey S. (2002). *Analysis of Financial Time Series.* Wiley.

State-of-the-art in Parallel Computing with R

Markus Schmidberger[1,*]

1. IBE, Ludwig-Maximilians-Universität Munich, Germany

* Contact author: schmidb@ibe.med.uni-muenchen.de

Keywords: Parallel Computing, Computer Cluster, Multi-core Systems, Grid Computing, Benchmark

R is a mature open-source programming language for statistical computing and graphics. Many areas of statistical research are experiencing rapid growth in the size of data sets. Methodological advances drive increased use of simulations. A common approach is to use parallel computing.

This presentation presents an overview of techniques for parallel computing with R on computer clusters, on multi-core systems, and in grid computing. It reviews sixteen different packages, comparing them on their state of development, the parallel technology used, as well as on usability, acceptance, and performance.

Two packages (snow, Rmpi) stand out as particularly useful for general use on computer clusters. Packages for grid computing are still in development, with only one package currently available to the end user. For multi-core systems four different packages exist, but a number of issues pose challenges to early adopters. The presentation concludes with ideas for further developments in high performance computing with R.

References

M. Schmidberger, M. Morgan, D. Eddelbuettel, H. Yu, L. Tierney, U. Mansmann (2009). State-of-the-art in Parallel Computing with R; Journal of Statistical Software; submitted. Preprint: `http://epub.ub.uni-muenchen.de/8991/`

Implementation of Software for Distributions in R

David Scott[1,*] , Diethelm Würtz[2], Christine Yang Dong[1]

1. University of Auckland
2. ETH Zürich

* Contact author: d.scott@auckland.ac.nz

Keywords: distributions; unit testing; variance gamma distribution

Distributions are fundamental to statistics and the implementation of them in software is likewise of extreme importance. Many distributions are implemented in R packages and an overview is given in a CRAN Task View, see Dutang (2009). In this talk we give suggestions for a standardized approach to the implementation of software for distributions. We discuss the functions which should be available in addition to the usual d-p-q-r functions; a possible standard naming system; suggested return structures for functions such as fitting routines; and appropriate testing procedures, focussing on unit tests. The package VarianceGamma which is currently being prepared exemplifies these ideas.

References

Dutang, Christophe (2009). CRAN Task View: Probability Distributions,
    http://cran.r-project.org/web/views/Distributions.html.

The Determination of an Environmental Service for a Contingent Valuation Study – Using R to Compute Estimates

Gary Sharp[1,*], David Friskin[,1], Stephen Hosking[,2], Catherine Logie[1] , Mark Nasila[1], Henri van der Westhuizen[,2]

1. Department of Statistics, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa
2. Department of Economics and Economic History, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa
* Contact author: gary.sharp@nmmu.ac.za

The tension when contrasting the trade-offs between environmental conservation and development growth is an acknowledged fact. Environmental conservationists are pro-active in defending estuarine quality arguing in favour of maintaining or improving environmental conditions. Unfortunately many conservationists ignore the economic implications in their arguments (King & Brown, 2009). This is not the case for development driven advocates, who tend to highlight the financial advantages of development projects.

This paper considers a contingent valuation study of the Bushman's estuary on the Southern coast of South Africa (van der Westhuizen, 2007). We use the R software to bootstrap density estimates of the median, the trimmed mean and the mean predicted willingness-to-pay for a log-linear estimated model. This valuation provides conservationists with a method for attaching an economic value for a recreational service.

References

King, J. (2009). *Building blocks and flow sessions: steps towards integrated flow management*, International conference on Implementing Environmental Water Allocations, (Port Elizabeth, South Africa) February 2009, pp1-2.

Van der Westhuizen, H. (2007), *Valuing preferences for freshwater inflows into the Bira, Bushmans, Kasouga, Keiskamma, Kleinemonde East, Nahoon and Tyolomnqa estuaries*. Unpublished masters dissertation, NMMU, Port Elizabeth.

ALM: An R Package for Simulating Associative Learning Models

Ching-Fan Sheu[1*] , **Teng-Chang Cheng**[2]

1. Institute of Education, National Cheng Kung University, Taiwan
2. Division of Academic Affairs, National Cheng Kung University, Taiwan

* Contact author: csheu@mail.ncku.edua.tw

Keywords: Associative learning, Pearce's configural model, Elemental model of Harris.

Animal and human learning is often studied by experimental tasks in which subjects are required to indicate a predicted outcome based on the presence or absence of stimulus events. Although current learning theories take either an elemental approach or a configural approach to stimulus representations, formal models of associative learning are essentially connectionist models and their output are evaluated, qualitatively, against patterns of experimental results from studies of similarity, discrimination, categorization, and so on.

This paper presents an R package for simulating predictions of Harris's elemental model and Pearce's configural model for associative learning. Researchers can readily generate graphical representations of model predictions by specifying experimental tasks as data frames and providing appropriate parameter values to R functions. The ability to visualize model predictions will facilitate testing these two associative learning models across a variety of experimental situations.

References

Harris, J.A. (2006). Elemental representations of stimuli in associative learning. *Psychological Review*, 113, 584-605.

Pearce. J.M. (2004). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, 101, 587-607.

Schultheis, H., Thorwart, A., & Lachnit, H. (2008). HMS: A Matlab simulator of the Harris model of associative learning. *Behavior Research Methods*, 40, 442-449.

Provenance Tracking in CXXR

Chris A. Silles[1,*] , **Andrew R. Runnalls**[1,**]

1. University of Kent, Canterbury, Kent, CT2 7NF, UK

* C.A.Silles@kent.ac.uk ** A.R.Runnalls@kent.ac.uk

Provenance is a record of lineage of a data object, and describes such things as what source data the data object was derived from, and the sequence of commands which was applied to generate the data object. Information systems are now ubiquitous in application domains where identifying the provenance of data is a critical ability, such as ensuring reproducibility of scientific research. This has led to the establishment of the field of *Provenance-Aware Computing*.

One of the pioneering papers within the provenance-aware computing literature is *Auditing of Data Analyses*, published by Becker and Chambers in 1988. In this article, the authors describe an *S AUDIT* facility which featured in *New S*. When *New S* was released in 1988, it signified a milestone in provenance-awareness. When a user issued a command within a *New S* session it maintained a record of the command, as well as which objects were read from or written to during the course of its execution. *R* currently has no support for an auditing facility such as *S AUDIT*.

Recently, the development of the methods employed within the field of provenance-aware computing for collecting and querying provenance has reflected the growing demand for more detailed information about the origins of data. Questions being asked of provenance information are typically complex, and it would be impossible to answer them using only a facility such as *S AUDIT*. Further advances have been made in the area of interoperability. The *Open Provenance Model* describes a method for the representation of provenance information so that, among other things, it may be exchanged between systems.

This paper describes how we have so far introduced facilities for provenance tracking into *CXXR*. *CXXR* is a project to refactorise the R interpreter into C++ while retaining as far as possible full functionality. The goal of *CXXR* is to allow for easier creation of experimental variants of the R interpreter.

In this paper we will describe and demonstrate the features we have introduced, such as the following facilities for inspecting the provenance of a given object,

- The sequence of operations performed on it (i.e. how it came to be);

- Which objects were used in its creation (i.e. its *ancestors*);

- Which objects used it for their creation (i.e. its *descendents*).

We will also discuss issues surrounding provenance collection in the context of a statistical environment, such as

- At what granularity provenance should be collected, and users be able to query it;

- How interoperability with other provenance-aware systems can be achieved.

Finally the paper will reflect on issues we have encountered while conducting this research, and outline its future directions.

References

Chambers, J.M, Becker, R.A (1988). "Auditing of Data Analyses", *SIAM J. Sci. Stat. Comput.* 9, 4, 747–760

Moreau, L., et al. (2008). "The Open Provenance Model", *Technical Report, University of Southampton*, `http://twiki.ipaw.info/bin/view/Challenge/OPM`

Moreau, L., Groth, P., Miles, S., Vazquez, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., Scheiber, A., Tan, V., Varga, L. (2008). "The Provenance of Electronic Data", *Communications of the ACM* 51, 4, 52–58

Runnalls, A.R. (2009). CXXR: Refactorising R into C++, `http://www.cs.kent.ac.uk/projects/cxxr`

Parallel Computing with Iterators

David Smith[1,*]

1. REvolution Computing, Inc.
* Contact author: david@revolution-computing.com

Keywords: iterator, parallel computing, distributed computing, ParallelR

While the R language provides means for programmer to iterate through the elements of a vector or list, it currently has no support for creating iterator objects as defined in languages like Java or Python. Using an iterator as a "cursor" to a list of values that can be defined dynamically solves a number of problems in R programming. In this talk, we'll begin by introducing a new function for creating iterator objects in R, describe its operators and methods, and provide practical examples of its use.

Although iterators are useful in their own right, one of the most useful applications we have found is in parallel and distributed programming. One of the biggest obstacles to the practical implementation of parallel programming is that for many existing systems a "new way" of thinking about R programming is required. We introduce instead a new function foreach() which we have found to be a natural and elegant construct for programming loops in R which can then easily be run in parallel. foreach() combines the simplicity of a for() loop to repeat arbitrary segments of code, with the power of the lapply() function to iterate over an object (with iterators). foreach() works efficiently to process loops sequentially, but with the addition of the ParallelR library it is trivial to make those same loops run in parallel. This has the result of dramatically reducing the both the time to process loops in R on multicore workstations or clusters, and the time it takes to program and debug them in the first place.

Mathematical modelling of the environment - are there enough data?

Karline Soetaert[a]

Dynamic mathematical models are commonly applied to analyse ecological and biogeochemical data. They allow a.o. to estimate immeasurable quantities, such as reaction rates and fluxes, with the ultimate goal to acquire some predictive capabilities.

Whereas the most complex models are solved by numerical integration of differential equations to obtain rates, the data requirements for applying such models are large. In many real-life applications, essential observations are lacking to obtain a complete picture and the model needs to be simplified in order to still make some sense out of the data. The ultimate simplification is to express the natural system in terms of linear (mass balance) equations, without the need to specify rate equations or to find kinetic parameters.

Which modeling technique is chosen has great implications on the way the model is solved, and in order to deal efficiently with incomplete data sets, a flexible (mathematical) repertoire for model application is required. There are many good reasons to perform ecological modeling in a software package strong in statistics and graphical output (i.c. R), not in the least because of the extensive pre- and post-processing required by these models. Since the introduction of R-package **odesolve** (Setzer, 2001) that offered a numerical integration routine, R has been promoted as a platform to perform dynamic model simulations (Petzoldt, 2003).

To broaden the scope of models that R can deal with, several other packages were recently created, performing certain mathematical tasks or providing utilities to facilitate the modeling process or the confrontation of models with data.

An overview of several modeling types and how to solve them in R, will be given. Although the emphasis will be on environmental models, the techniques presented have a much wider scope.

References

Petzoldt, T. (2003). R as a simulation platform in ecological modelling. *R News*, 3(3):8–16.

Setzer, R. W. (2001). *The odesolve Package: Solvers for Ordinary Differential Equations*. R package version 0.1-1.

[a]Centre for Estuarine and Marine Ecology (CEME), Netherlands Institute of Ecology (NIOO), 4401 NT Yerseke, Netherlands. E-mail: k.soetaert@nioo.knaw.nl

Linking the Offender's age to the Criminal Event: A Statistical Study on Sex-related Homicides

Helmut Tausendteufel[1], Stephan Stahlschmidt[1,2,*], Wolfgang Härdle[2]

1. Fachhochschule für Verwaltung und Rechtspflege Berlin, Faculty 3, Alt-Friedrichsfelde 60, D-10315 Berlin, Germany
2. Humboldt-Universität zu Berlin, School of Business and Economics, Institute of Statistics and Econometrics, Spandauer Straße 1, D-10718 Berlin, Germany

* Contact author: stahlschmidt@wiwi.hu-berlin.de

Keywords: Bayesian Network, Classification, Criminal Event Perspective

Offender profiling has gained much popularity over the last years. But although important progress has been made, little is known about the implications of the offender's age on the crime. The project at hand investigates this issue by means of an exploratory statistical study focusing on sex-related homicides. The project is thereby based on a dataset of 350 sex-related homicides in Germany since 1991, in which the offender was found guilty and was convicted.

The forensic theoretical background is provided by the Criminal Event Perspective stressing the interaction of the offender's behaviour, the victim's behaviour and the underlying situation. These three components together determine the sequences of the crime and therefore any observable variable.

In order to support police profilers with a tool applicable in their investigation, a Bayesian network will be presented mirroring the causalities found in the dataset. To this end Qualitative Comparative Analysis, visualization techniques and different exploratory techniques are applied in the R environment.

References

Bøttcher, S. G. and Dethlefsen, C. (2003). deal: A Package for Learning Bayesian Networks. *Journal of Statistical Software*, 8(20).

Dusa, A. (2008). QCA: Qualitative Comparative Analysis. R package version 0.5-2.

Scutari, M. (2009). bnlearn: Bayesian network structure learning. R package version 1.1.

Swayne, D. F., Buja, A. and Temple Lang,D. (2003). Exploratory Visual Analysis of Graphs in GGobi. *Proceedings of DSC 2003* (Vienna, Austria), March 2003, 20-22.

A Spatial Multinomial Case-Control Modeling Package

Celso Stephan[1,2,*]**, Luciana B Nucci**[1,2]**, Ana Carolina C N Mafra**[1,2 ,]**, Ricardo Cordeiro**[1,2]

1.   UNICAMP – State University of Campinas, SP/Brazil
2.   epiGeo – Epidemiological Spatial Data Analisys Lab
*   Contact author: celso@stephan.mus.br

Case control studies are one of the most important methodological contributions for efficiently population risk estimations in epidemiology[1]. Spatial aspects of such studies have been supported by new geoprocessing tools, dissemination of digital maps and popularization of GPS devices. Concepts on spatial risk[2] and multinomial models theory[3] were used for development of this R package, which is part of a theoretical project for interpretation on this kind of study.

Functions were developed in R (S4) language for modeling spatial case-control data and maps designing based on calculated risk and its significance limits of provided shape files using Kernel smoothers. These functions will compose a new R Package to be available on CRAN.

This work shows how these functions might be used, taking by example a case-control study realized in 2006-7 period. Work related accidents cases were classified in three levels (light, moderate and heavy/fatal), compared each other and against controls (no accident) obtained from a random sample in a 300,000 inhabitants city in Brazil. Maps of the spatial risk magnitude were made and its significant regions were calculated and showed on the maps.

The development of this package will bring the possibility of spatial case-control analysis, under multinomial distributions, where cases were classified in levels and automatic designing risk maps for each level in comparison to others.

References

Cordeiro R (2005). The rare disease Myth. *Revista Brasileira de Epidemiologia, 8:111-6.*

Bithell J (1990). An application of density estimation to geographical epidemiology. *Statistics in Medicine,* 9:691-701.

Agresti A (2002). Categorical data analysis, 2nd ed, New York: John Wiley & Sons.

Customizing the rpart library for multivariate gaussian outcomes: the longRPart library

Sam Stewart[1*], Mohamed Abdolell[2], Michael LeBlanc[3]

1. Faculties of Medicine and Computer Science, Dalhousie University
2. Department of Diagnostic Radiology, Dalhousie University
3. Fred Hutchinson Cancer Research Center, Seatle, Washington
* sam.stewart@dal.ca

Keywords: repeated measures outcomes, classification trees, rpart

Abdolell et al. (2001)[1] implemented a binary partitioning algorithm for the case of continuous repeated measures outcomes, using Mahalanobis distance as a deviance measure to evaluate goodness-of-split. The algorithm was implemented only for a single split at the root node of the tree with the single purpose of dichotomising prognostic variables. The binary partitioning algorithm was implemented in SAS using the PROC MIXED procedure, along with a permutation test to evaluate the p-value of the associated binary split and a bootstrap method to calculate a confidence interval.

This project extends the binary partitioning algorithm of Abdolell et. al to a binary recursive partitioning algorithm[2] which is implemented in R. We utilize the *nlme* library to extend the *rpart* library[3], producing the *longRPart* library for binary recursive partitioning in the case of MVN outcomes, and extends the algorithm to split on unordered categorical variables. A tree plotting function is developed for annotated plots that are applied to terminal nodes of the tree to display the longitudinal profiles of the outcome variable.

A detailed discussion will be presented of how the *rpart* library was extended to accomodate the longitudinal outcome with its associated deviance measure, and how to apply these same principles to the case of other non-standard outcomes using custom R functions.

References

Abdolell et al (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Statistics in Medicine* 21(22), 3395-03409.

Breiman, Friedman, Olshen, and Stone. (1984) Classification and Regression Trees. Wadsworth.

Terry M Therneau and Beth Atkinson. R port by Brian Ripley. (2008). rpart: Recursive Partitioning. R package version 3.1-42.
http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm

Threshold cointegration in R

Matthieu Stigler[1]

1. National Institute for Public Finance and Policy

* Contact author: Matthieu.Stigler@gmail.com

The concept of cointegration suggests that even if two or more variables are non-stationary, there can exist a linear combination of them that is stationary (Engle and Granger 1987). It implies the existence of a stable long-run relationship between the variables. In this framework, however, every deviation from the long-run equilibrium results in an error correction mechanism.

The concept of threshold cointegration (Balke and Fomby 1997) extends the linear cointegration case by allowing the adjustment to occur only after the deviation exceeds some critical threshold. This allows the model to take into account possible effects of transaction costs or stickiness of prices. Further, it permits us to capture asymmetries in the adjustment process, where positive or negative deviations are not corrected to the same extent.

The presentation will review the concept of threshold cointegration, discuss recent developments in the field and show how to use the package "tsDyn" (di Narzo, Aznarte and Stigler 2009) in R to estimate and interpret threshold cointegration models.

References

[1] Nathan S Balke and Thomas B Fomby, *Threshold cointegration*, International Economic Review **38** (1997), no. 3, 627–45.

[2] Fabio di Narzo, Jose Aznarte, and Matthieu Stigler, *Development version of package tsdyn*, `http://code.google.com/p/tsdyn/wiki/ThresholdCointegration`, January 2009.

[3] R. F. Engle and C.W.J. Granger, *Co-integration and error correction: representation, estimation and testing.*, Econometrica **55** (1987), no. 2, 251–276.

Whole Data Approaches to Large-Scale Multiple Hypothesis Testing

John Storey

Department of molecular biology
University of Princeton

jstorey@princeton.edu

A number of modern scientific problems involve simultaneously testing thousands or millions of hypotheses. Methodologies for these problems have typically been developed in the context of p-values or test-statistics, which are one-dimensional summaries of each hypothesis test's data. I will present some of our recent developments where the approaches are instead formulated at the level of the whole data set. This allows information shared across hypothesis tests to be directly captured and utilized. Specifically, solutions to multiple testing dependence and optimal frequentist multiple testing will be presented, both of which require careful computational considerations.

Party on! – A new, conditional variable importance measure for random forests available in `party`

Carolin Strobl[1,*] , **Achim Zeileis**[2]

1. Department of Statistics, Ludwig-Maximilians-Universität München
2. Department of Statistics and Mathematics, Wirtschaftsuniversität Wien

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

Keywords: Permutation importance, variable selection, spurious correlation.

Random forests have become very popular in many scientific fields because they can cope with "small n large p" problems involving complex interactions. Random forest variable importance measures have been suggested as screening tools, e.g., for gene expression studies. However, these variable importance measures have been shown to be biased in favor of predictor variables of certain types and towards correlated predictor variables.

While the former issue could be addressed straightforwardly in `party` by means of unbiased split selection and resampling schemes (Strobl et al., 2007), in the case of correlated predictors the original permutation importance is highly misleading, creating a new source of bias in interpretations drawn from random forests. Therefore, Strobl et al. (2008) recently suggested a solution for this problem in the form of a new, conditional permutation importance measure. Starting from version 0.9-994, this new measure is available in the `party` package.

In the talk, the rationale and application of this new measure is outlined and illustrated by means of a toy example. Moreover, some hands-on advice is given for sensibly using and interpreting random forests in R.

References

C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25.

C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307.

T. Hothorn, K. Hornik, C. Strobl and A. Zeileis (2009). `party`: A Laboratory for Recursive Partytioning. `http://CARN.R-project.org/package=party`.

C. Strobl, T. Hothorn and A. Zeileis (2009). Party on! A new, conditional variable importance measure for random forests available in the `party` package. *Technical report (submitted).* `http://epub.ub.uni-muenchen.de/9387/1/techreport.pdf`.

Multiple imputation with diagnostics: opening windows into the black box

Yu-Sung Su[1,*] , Gelman Andrew[1], Jennifer Hill[2]

1. Columbia University
2. New York University

* Contact author: yusung@stat.columbia.edu

Our new R package, "mi," has several features that allow the user to get inside the imputation process and evaluate the reasonableness of the resulting model and imputations. These features include: flexible choice of predictors, models, and transformations for chained imputation models; binned residual plots for checking the fit of the conditional distributions used for imputation; and plots for comparing the distributions of observed and imputed data in one and two dimensions. In addition, we use Bayesian models and weakly informative prior distributions to construct more stable estimates of imputation models. Our goal is to have a demonstration package that (a) avoids many of the practical problems that arise with existing multivariate imputation programs, and (b) demonstrates state-of-the-art diagnostics that can be applied more generally and can be incorporated into the software of others.

Sound analysis and synthesis with the package Seewave

Jérôme Sueur[1,*], Thierry Aubin[2], Caroline Simonis[3]

1. Muséum national d'Histoire naturelle, Département Systématique et Évolution, CNRS UMR 7205, CP 50, 45 rue Buffon, 75005 Paris, France

2. Neurobiologie de l'Apprentissage, de la Mémoire et de la Communication, CNRS UMR 8620, Bât. 446, Université Paris-Sud, 91405 Orsay Cedex, France

3. Muséum national d'Histoire naturelle, Département Écologie et Gestion de la Biodiversité, MNHN USM 301 & CNRS UMR 7179, CP 55, 57 rue Buffon, 75005 Paris, France

* Contact author: sueur@mnhn.fr

Keywords: time series, sound, Fourier, amplitude, frequency

`Seewave` is a package for sound analysis and synthesis. It has been first submitted to CRAN in 2006, a new version having been released about all semester. The package has been initially written to help analysing sound produced by animals. However, it is now used in different contexts such as telemetry, optical signals, high frequency vibrations, meteor signal shape and others.

Sound input can be achieved using different object classes : (*i*) usual classes (numeric `vector`, numeric `matrix`) if the sampling frequency is provided, (*ii*) time series classes (`ts`, `mts`), and (*iii*) sound-specific classes (`Wave` of the package `tuneR` and `Sample` of the package `Sound`).

`Seewave` currently includes more than 70 functions. Sounds are edited as oscillogram or amplitude envelope in single or multi-framed windows. An option can be set to move along the signal using a time slider. Signals can be modified with cutting, inserting, pasting, muting, fading, and repeating functions.

In the time/amplitude domain, signal and silence durations can be automatically measured. Amplitude filters can help reducing a background noise. Amplitude modulations can be automatically removed or quickly modified interactively and echoes can be generated through Doppler effect.

In the frequency domain, 15 statistical descriptive parameters (dominant peak, quality factor, entropy, spectral flatness, etc) are extracted from a frequency spectrum by calling a single function. The fundamental frequency of harmonic series is detected by the autocorrelation or cepstral method, while the instantaneous frequency is obtained by the zero-crossing method or Hilbert transform. `Seewave` provides a short-term Fourier transform to return mean spectra, 2D and 3D spectrograms. Fourier window size, overlap and zero-padding options allow the user to improve graphical representation, and to reduce the uncertainty principle.

To test for the temporal and frequency similarity of two sounds, cross-correlations, surface computation and coherence can be computed.

New sound is designed with sinusoidal amplitude modulations and linear and/or sinusoidal frequency modulations. Simple additions together with requency filters and linear frequency shifts ensure the modification or generation of complex sound.

References

Sueur J, Aubin T, Simonis C (2008). Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics*, 18: 213–226.

Sueur J (2009). Seewave, http://sueur.jerome.perso.neuf.fr/seewave.html.

Exploring the multivariate structure of missing values using the R package VIM

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

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

* Contact author: templ@tuwien.ac.at

Keywords: Missing Values, Visualization, Missing Value Mechanism, R.

In our presentation we will describe and demonstrate the usefulness of exploring missing values in data by using visualization tools in order to get a first impression of the data but also as pre-processing step before imputation.

Before choosing an imputation procedure to impute missing values in data one should be aware of the missing value mechanism. We propose the use of visualization tools for the detection of the missing value mechanism(s). Univariate plots such as histograms and spineplots, multivariate plots such as multiple scatterplots and parallel coordinate plots and maps are adapted in order to visualize missing values (see also Theus et al., 1997), but new plots ("matrixplot" and others) are provided as well (see, e.g., Templ and Filzmoser, 2008). In addition to that, interactivity is provided when using these plots, i.e. the users have the possibility to highlight or re-arrange the plots by clicking on the plots.

VIM can be used for data from essentially any field. If spatial coordinates are available, it is possible to load a background map and present information about missing values on that map.

Our developed R-package VIM (Templ and Alfons, 2009) includes a graphical user interface for interactive, easy to grasp graphics in order to make the tools available also for non-experts in R.

We will illustrate our proposed visualization methods through a short demo using real-life data sets.

References

Templ, M. and Filzmoser, P. (2008). Visualisation of Missing Values Using the R-package VIM. Research Report CS-2008-1, Department of Statistics and Probability Theory, Vienna University of Technology. http://www.statistik.tuwien.ac.at/forschung/CS/CS-2008-1complete.pdf.

Templ, M. and Alfons, A. (2008). VIM: Visualization and Imputation of Missing Values. R package version 1.2.4. http://cran.r-project.org.

M. Theus, H. Hofmann, B. Siegl, A. Unwin.(1997). MANET: Extensions to Interactive Statistical Graphics for Missing Values. In New Techniques and Technologies for Statistics II, IOS-Press Amsterdam, 247-259. http://home.vrweb.de/~martin.theus/NTTS.pdf

Statistical Graphics! Who needs Visual Analytics?

Martin Theus

martin@theusRus.de

Data visualization seems to be everybody's darling; not based on a formal discipline but used by almost every scientist who wants to display qualitative or quantitative information graphically. Looking at the data visualization literature we find a broad range of aspects that can be touched upon.

As statisticians, we should be more targeted and basically can distinguish between two uses of graphics in statistics: exploration and diagnostics. Only at first sight these two applications of graphics are disjunctive, and further exploration is often the result of the investigation of a model via diagnostic plots.

This talk looks into why R became so popular for creating statistical graphs of all kinds despite the not state-of-the-art graphics core. It further focuses on what needs to be done, in order to use R for exploratory data analysis which is centered around data visualization and supports traditional statistical reasoning in many ways.

Distributed Text Mining with tm

Stefan Theussl[1,*]**, Ingo Feinerer**[2]**, Kurt Hornik**[1]

1. Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, A-1090 Wien, Austria
2. Institute of Information Systems, DBAI Group, Technische Universität Wien, A-1040 Wien, Austria

* Contact author: Stefan.Theussl@wu-wien.ac.at

Keywords: High performance computing, Text mining, MapReduce, Distributed file system

Text mining is a widely used technique utilizing statistical and machine learning methods to extract patterns or knowledge from large unstructured text data sets. Recently R has gained explicit text mining support via the **tm** [2, 3] package. This infrastructure provides sophisticated methods for document handling, transformations, filters, and data export (e.g., term-document matrices).

However, the availability of very large and always growing text corpora poses new challenges for efficient handling of these data sets mainly due to architectural performance limits of single processor environments and memory restrictions. On the other hand we observe an increasing availability of multicore architectures even in commodity computers and high performance computing environments, i.e., distributed and highly integrated computing clusters.

In this context, we propose to make use of a technique called MapReduce [1] which is widely used in high performance computing because of its functional programming nature. Existing building blocks in **tm** allow for adding new layers to support this kind of parallelism and distributed allocation. In particular we identify compute-intensive parts of **tm**, break these parts up into suitable entities for parallel processing and finally encapsulate the emerging parallelism in a functional programming style.

A key factor in large scale text mining is the efficient management of data. Therefore, we show how distributed storage can be utilized to facilitate parallel processing of large and very large data sets. This approach offers us a reliable, flexible, and scalable high performance computing solution for distributed text mining.

References

[1] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In *OSDI'04, 6th Symposium on Operating Systems Design and Implementation*, pages 137–150, 2004.

[2] Ingo Feinerer. *tm: Text Mining Package*, 2009. R package version 0.3-3.

[3] Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in R. *Journal of Statistical Software*, 25(5):1–54, March 2008.

Tools on R for Dose-response curves analysis

Chantal Thorin[*], Yassine Mallem, Jacques Noireaud, Jean-Claude Desfontis

*UPSP 5304 : Physiopathologie Animale et de Pharmacologie Fonctionnelle.
Ecole Nationale Vétérinaire,Route de Gachet BP 40707, 44307 Nantes Cedex, France*

* Contact : thorin@vet-nantes.fr

Keywords: R, Non linear mixed effects models, Dose response curves, Pharmacological parameters

In experimental pharmacology, studies on drug-receptor interactions commonly use dose-response curves (DRC) established under repeated measurements designs. The best approach to analyse a dose-response relationship is to use non linear mixed effects models (nlme) (Davidian and Giltinian, 1995), but specific softwares dedicated to analyse pharmacological data have not yet developed nlme procedures. The aim of this work was to provide accurate and easy-to-use tools on R to assist pharmacologists with using nlme modelling to fit DRC.

Five functions using available R packages have been built. The *Est.Pop* function, using nlme function in nlme package (Pinheiro and Bates, 2000), gives an estimation of the different parameters included in the predicted function and a qqplot of the residuals. The *IC.par* function provides a confidence interval for each parameter of the predicted function for the confidence level asked by users. The *Graph.curves* function displays a graph showing the individual fitted curves and the population fitted curve which illustrate the individual effect on physiological response. Nevertheless, nlme procedures are very susceptible to outliers points in the data sets and the convergence of the iterative calculus is not always achieved. In those situations and when the residuals seem not to be normally distributed the *Est.Boot* function is more accurate to give an estimation of the predicted function parameters by a non parametric bootstrap method using the bootstrap package (Huet et al, 2004). Depending on the bioassay and the relative asymmetry of the curves, four predictive functions (Hill equation, Richards, Gompertz, Hill modified functions) can be tested (Giraldo et al, 2002) with those tools; the *Comp.Mod* function is dedicated to compare established models and to detect the best one.

The nlme modelling analysis of a set of dose-response curves fom β- adrenoceptors-mediated blood vessels relaxation studies (Mallem et al, 2005) will be presented and discussed.

References

Davidian M, Giltinan DM (1995). *Nonlinear Models for Repeated Measurement Data.* Chapman & Hall-CRC : New York.

Huet S, Bouvier A, Poursat M-A, Jolivet E (2004). *Statistical Tools for Nonlinear Regression: a practical guide with S-PLUS and R examples- Second Edition.* Springer-Verlag: New York.

Giraldo J, Nuria M, Vivas B, Badia A (2002). Assessing the (a)symmetry of concentration–effect curves: empirical versus mechanistic models, *Pharmacology & Therapeutics* , **95**, 21-45.

Mallem MY, Holopherne D, Reculeau O, Le Coz O, Desfontis J-C, Gogny M (2005). β-Adrenoceptor-mediated vascular relaxation in spontaneously hypertensive rats. *Autonomic Neuroscience* **118**, 61-67.

Pinheiro JC, Bates DM (2000). *Mixed-Effects Models in S and S-Plus.* Springer-Verlag :New York.

maxLik: A Package for Maximum Likelihood Estimation in R

Ott Toomet[1,2] **and Arne Henningsen**[3,4,*]

1. Department of Economics, University of Tartu (Estonia)
2. Department of Economics, Aarhus School of Business, University of Aarhus (Denmark)
3. Department of Agricultural Economics, University of Kiel (Germany)
4. Institute of Food and Resource Economics, University of Copenhagen (Denmark)

* Contact author: arne.henningsen@gmail.com

The **maxLik** package provides convenient tools for maximum likelihood (ML) estimations in the statistical software environment R. This package is available from CRAN (`http://cran.r-project.org/package=maxLik`), R-Forge (`http://r-forge.r-project.org/projects/maxlik/`), and its homepage (`http://www.maxLik.org/`).

The most important tool for a user of the **maxLik** package is probably the `maxLik` function. It is a wrapper function that delegates the maximum likelihood estimation to the selected optimisation routine. Five optimisation methods are currently available (names of the corresponding functions in parenthesis): Newton-Raphson (`maxNR`), Berndt-Hall-Hall-Hausman (`maxBHHH`), Broyden-Fletcher-Goldfarb-Shanno (`maxBFGS`), Nelder-Mead (`maxNM`), and simulated-annealing (`maxSANN`). While the actual optimisation in `maxBFGS`, `maxNM`, and `maxSANN` is done by `optim`, the Newton-Raphson algorithm is implemented in the function `maxNR` itself. The actual optimisation in `maxBHHH` is done by `maxNR`.

The first argument of `maxLik` (`loglik`) is mandatory and specifies the log-likelihood function. Its first argument must be the vector of the parameters to be estimated and it must return either a single log-likelihood value or a numeric vector where each component is the log-likelihood value corresponding to an individual observations. The second and third argument (`grad` and `hess`) are optional and can be used to specify functions that return the gradients and the Hessian of the objective function, respectively. If these functions are not provided by the user, numerical gradients and Hessians are calculated if necessary. The fourth argument (`start`) is mandatory and must be used to specify a vector of starting values. Finally, the fifth argument (`method`) is optional and can be used to select the maximisation routine. It defaults to `"NR"`, but it can also be `"BHHH"`, `"BFGS"`, `"NM"`, or `"SANN"`. The `maxLik` wrapper capabilities are designed in a transparent way, so that the user can easily swap the methods without changing the arguments. The arguments not used by a particular optimisation method, such as `hess` for the Berndt-Hall-Hall-Hausman method, are ignored.

The **maxLik** package is implemented using S3 classes. The `maxLik` wrapper returns a list of class `"maxLik"`. Corresponding methods can handle the likelihood-specific properties of the estimate including the fact that inverse of the negative Hessian is the variance-covariance matrix of the estimated parameters. The most important methods for objects of class `"maxLik"` are: `summary` for returning (and printing) summary results, `coef` for extracting the estimated parameters, `vcov` for calculating the variance covariance matrix of the estimated parameters, `logLik` for extracting the log likelihood value, and `AIC` for calculating the Akaike information criterion.

Currently, the **maxLik** package is used for maximum likelihood estimations in three packages that are available on CRAN: **mlogit**, **sampleSelection**, and **truncreg**. On the useR! conference, we would like to demonstrate how to use the **maxLik** package for maximum likelihood estimations in R. Furthermore, we would like to highlight its advantages and features to encourage more users and package writers to use the **maxLik** package.

iPlots Extreme - next-generation interactive graphics for analysis of large data in R

Simon Urbanek[1,2,*]

1. AT&T Labs - Research
2. R Core development team

* Contact author: simon.urbanek@r-project.org

Keywords: interactive graphics, data analysis, large data

Interactive graphics have proven to be very helpful in data analysis not only for explorative data analysis but also in the analysis of models and model results. However, R provides no native facilities for interactive graphics. iPlots have been developed to bridge this gap and provide highly interactive and customizable framework for interactive graphics. As the size of datasets and capabilities of modern computers increase to allow even large and larger datasets to be processed in R, the same need is required from interactive graphics software. In this paper we present an entirely new generation of interactive graphics: iPlots Extreme which leverages the potential of modern computers to allow us to visualize and analyze large data. New approaches are necessary not only in the highly-optimized implementation but also in approaches to methods for visualization of such large datasets. iPlots Extreme offer a wide range of plots as well as customization. They support both continuous and categorical data with many interactive features while maintaining a flat learning curve and an intuitive interface. We will discuss the design as well as illustrate the use of iPlots Extreme on large-scale practical examples.

Microsoft Office Dynamic Documents as R Applications

Josh van Eikeren[1,*] and Paul van Eikeren[1]

1. Blue Reference, Inc., Bend Oregon USA
* Contact author: Josh.van.Eikeren@BlueReference.com

Keywords: Dynamic Documents, reproducible research, Microsoft, Word, Excel

Business and technical luminaries argue that future competitive success will be based on the application of data, analysis, predictive modeling, and fact-based decision making. "Competing on Analytics" makes use of a class of software, like R, which enables a range of techniques from statistics to data mining to analyze historical data and make predictions about future events. However, despite these sophisticated developments, common documents remain central to business processes. Documents are portable, persistent and provide contextual views of information organized for the purpose of relating disparate pieces into actionable knowledge.

Modern business practices (e.g., Business Process Management) are driving the convergence of documents and predictive analytics. Business data is typically focused on the "what" of the business and maintained in highly structured databases. Business documents tend to focus on the "why," are unstructured and contextual, and are usually maintained in separate document repositories. And, predictive analytic applications contain the "how" of the business, are highly structured, but typically lack an appropriate storage mechanism. The reality is that business is done at the intersection of "what", "why" and "how"—where facts and context meet actionable analysis.

Traditionally, the domains of data, documents and predictive analytics application have been isolated from one another. Dynamic documents provide a means to combine the strengths of documents with the power and flexibility of predictive analytics applied to data. In fact, dynamic documents can act like situational software applications by combining the best attributes of applications and static documents. To address the needs for dynamic documents, we have extended the capabilities of Microsoft Office to enable embedding a structured entity we call a Parts-Container for holding and managing data (e.g., data frames), software objects (e.g., ASCII and binary files), code blocks (e.g., R scripts) and inline expressions (e.g., R commands). Since the elements of the Parts Container are linked to a computational engine like R, the documents become applications.

In this presentation, we will outline several case studies that illustrate the construction and application of dynamic documents in conjunction with the R statistical computation environment towards

- test-driven new method development;
- reproducible research; and
- study management in pharmaceutical development.

References

Josh van Eikeren (2009). Inference for R,
    http://InferenceForR.com.

Paul van Eikeren (2009). Inference Parts-Container Explained,
    http://inferenceforr.com/documentation/Documents/ProductOverviews/Inference%20Parts%20Container%20Overview.pdf

Bayesian Approach to the Specification of Design Space in Quality by Design

Paul van Eikeren[1,*] and Corey Dow-Hygelund[1]

1. Blue Reference, Inc., Bend Oregon USA
* Contact author: Paul.van.Eikeren@BlueReference.com

Keywords: Bayesian, Response Surface Models, Design Space, Quality by Design

The United States Food and Drug Administration (FDA) has recognized that product development is now the weak link in the "critical path" from scientific discovery to commercial drug products. In response, the FDA has instituted sweeping changes on the way pharmaceutical developers and manufacturers conduct their business. Corresponding global regulatory authorities have followed suit. The FDA's Quality by Design (QbD), a risk-based approach, is focused on process understanding including identifying sources of variability, overall reliability and process robustness. From the FDA's viewpoint, the principles of QbD in pharmaceutical development require establishing a clear linkage between the safety and efficacy of the drug product in the patient with its quality as defined by the attributes of the drug product and then linking it all the way back to the process for preparing the drug product.

Central to the QbD approach is the establishment of a Design Space comprised of the "multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality." Despite the central role of Design Space, limited prescriptive information is available regarding to how to construct such a Design Space and to demonstrate that operation within it "…provides assurance of quality."

Development and calibration of a Design Space typically involves construction of multiple predictive response surface models corresponding to drug product attributes. The Design Space is constraint by requirements of meeting multiple response criteria. Such multiple response surface optimizations are typically approached using overlapping mean response or by a desirability function. However, these approaches fail to account for the uncertainty in model parameters and the correlation structure of the data. As shown by Peterson (2004, 2008), a Bayesian approach employing posterior predictive distributions addresses both of these limitations.

This presentation will be directed at several case studies illustrating the use of R in conjunction with an assortment of R packages towards the construction of design spaces for representative pharmaceutical products as part of Quality by Design activities. Furthermore, these case studies will illustrate the additional benefits of using the Bayesian approach, including the following:

- providing estimates of the uncertainty in model parameters;
- enabling the use of prior information leading to more efficient adaptive design and experimentation;
- enabling an approach towards robust parameter design;
- providing a figure of merit (probability) for meeting product specification criteria in terms consistent and easy to understand by technical workers operating in a regulated environment;
- enabling a means to establish the reliability of the Design Space; and
- providing a basis for selection of alternate process settings within the design space while ensuring "assurance of quality."

References

Paul van Eikeren (2009). Inference for Quality by Design, http://InferenceForQbD.com.

John. J. Peterson (2004). A posterior predictive approach to multiple response surface optimization. *Journal of Quality Technology*, 36, 139-153.

John J. Peterson (2008). A Bayesian Approach to the ICH Q8 Definition of Design Space. *Journal of Biopharmaceutical Statistics,* 18, 959-975.

An integrated and statistical approach using R to assess the economic importance of coastal fisheries in France.

Van Iseghem Sylvie[1,*] Demanèche Sébastien[2], Daurès Fabienne[1], Leblond Emilie[2]

1. IFREMER, Département d'Economie Maritime, Centre de Brest - BP 70, 29 280 PLOUZANE
2. IFREMER, Département STH, Centre de Brest - BP 70, 29 280 PLOUZANE
* Contact author: svaniseg@ifremer.fr

Keywords: economic indicators, sampling plan optimisation, precision levels, regression model, coastal fisheries

Coastal fisheries are highly represented in France and in all European Union EU Member States; vessels less than 12 meters represent almost 75% of the total European fleet. Nevertheless, due to the lack of information available on them, their importance are often underestimated. In 1995, the Code of Conduct for Responsible Fisheries was adopted by the Food and Agriculture Organization. The Code emphasizes that the development of fisheries management plans requires appropriate reliable data on all aspects of a fishery. In particular, the Code stressed that "in order to insure the sustainable management of fisheries and to enable social and economic objectives to be achieved, sufficient knowledge of social, economic and institutional factors should be developed through data gathering, analysis and research" (article 7.4.5). Based on these considerations and in order to provide the scientific basis for the implementation of the Common Fisheries Policy, the Fisheries Council of the European Union decided in 2000 to establish a Community program for the collection of data needed to evaluate the situation of all the fisheries sector.

This paper presents the statistical approach ongoing in France to collect economic data in order to satisfy EU requirements and to characterize the economic status of French coastal fleets.

The methodology includes both an optimized sampling plan and a model used to re-assess the contribution of small-scale fisheries to national production. The optimized sampling plan provides a sample of about 15% of the total French fleet collected from a direct survey of fishermen. The sampling scheme is optimized to represent the economic indicators variability by category of vessels and geographic distribution and to insure that the levels of precision required are satisfied. The modelling combines both official landings and data collected from directs surveys.

The role of small scale fisheries in the French professional fishing sector is re-evaluated and its key role is demonstrated.

References

- Ardilly P. 1994. Les Techniques de Sondage, Paris, Technip, 393pp.
- Berthou, P., O. Guyader, E. Leblond, S. Demanèche, F. Daures, C. Merrien, P. Lespagnol, 2008. From fleet census to sampling schemes: an original collection of data on fishing activity for the assessment of the French fisheries, ICES ASC 2008/K12.
- Cochran, W.G. (1977), Sampling Techniques, third edition. John Wiley & Sons, Inc., New York, 428pp.
- Daurès, F., S. Demanèche, et al. (2003). Methodology for the assessment of aggregated economic indicators in the fishing sector: estimation of a revenue function. XVth EAFE Annual Conference, Brest (France).
- Guyader, O., P. Berthou, C. Koustikopoulos, F. Alban, S. Demanèche, M. Gaspar, R. Eschbaum, E. Fahy, O. Tully, L. Reynal, A. Albert. 2007. Small-Scale Coastal Fisheries in Europe. Final report of the contract No FISH/2005/10, 447pp. http://ec.europa.eu/fisheries/publications/studies_reports_en.htm.
- Leblond, E., F. Daurès, P. Berthou, Ch. Dintheer, C. Merrien, A. Tétard, J. Vigneau, P. Lespagnol, 2008. The Fisheries Information System of Ifremer - a multidisciplinary monitoring network and an integrated approach for the assessment of French fisheries, including small-scale fisheries, ICES ASC 2008/K11.
- Leblond, E., F. Daurès, et al. (2007). La Synthèse des Flottilles de pêche 2005 - Flotte Mer du Nord - Manche –Atlantique. IFREMER, SIH: 58 p. (http://www.ifremer.fr/sih).
- Tillé Y. 2001. Théorie des sondages, Paris, Dunod, 284pp.
- Van Iseghem, S., S. Demanèche, et al. (2004). Optimization of a Sampling Plan for Economic Data Collection: Application to the Atlantic French Fleet. XVIth EAFE Annual Conference, Rome (Italy).

1

Automatic Numerical Differentiation of Noisy, Time-Ordered Data in R

Ravi Varadhan[1,*] , **Ganesh Subramaniam**[2]

1. Center on Aging and Health, Johns Hopkins University
2. AT&T Labs - Research

* Contact author: rvaradhan@jhmi.edu

Keywords: Numerical derivative, Smoothing, FDA, Time Series Data Mining'

In nonparametric regression, it is often of interest to estimate some functionals of a regression function, such as its derivatives. For example, in biomechanics, the estimation of derivatives of displacement data (e.g. velocity and acceleration of body segments) is a common task (Pezzak 1977). In the study of growth curves, the first (velocity) and second (spurt) derivatives of the height as a function of age are important parameters for study (Ramsay & Silverman 1997).

In this work, we study the estimation of derivatives of regression functions using nonparametric regression approaches. Suppose we observe

$$y_j = g(x_j) + \epsilon_j, j = 1, ..., n,$$

where the ϵ_js are uncorrelated with $E(\epsilon_j) = 0$ and $E(\epsilon_j^2) = \sigma^2 > 0$, and the design points $x_j{}_{j=1}^n$ are either equally-spaced or randomly distributed. The main interest lies in the estimation of $g'(x_j)$ and $g''(x_j)$.

This investigation on derivative estimation was motivated by a time series data mining problem in telecommunication network data, where given a large number of time series the objective was to indetify time series that have potentially 'interesting behavior' (Subramaniam and Varadhan 2007 and 2008). Ramsay and Silverman (1997) demonstrated several examples where features that are not visible in the original data are detected in the first and second derivatives. A critical parameter in nonprametric regression is the bandwidth for smoothing noisy data. With the large number of curves, optimal bandwidth selection on a curve-by-curve basis is not practical, but some reasonable approximation of optimal bandwidth using, for example, generalized cross-validation or plug-in-bandwidth, might be acceptable. However, the main problem is that the optimal bandwidth for estimating regression function is generally smaller than the optimal bandwidth for estimating the derivatives (Hardle 1985). Thus automatic smoothing of optimal derivative estimation in a data mining setting presents a unique challenge.

Various nonparametric regression approaches are available in R: kernel regressions, smoothing splines, penalized splines and local polynomial. Our objective here is two-fold: (1) to compare the different smoothing techniques for their accuracy in estimating the first and second derivatives of the regression function using the automatic bandwidth selection techniques that are applied to the noisy data itself, and (2) to propose and evaluate some approaches for approximating optimal bandwidth for derivative estimation. We conduct a comprehensive simulation study involving: (a) various nonparametric regression methods (smoothing splines, penalized splines, kernel smoothing and local polynomial), (b) different regression functions, (c) two types of designs - equally spaced or randomly distributed time points, (d) different signal-to-noise rations, and (e) homoscedastic and heteroscedastic errors. We evaluate the performance in terms of mean integrated squared error, integrated squared bias, and integrated variance. In addition to the simulation results from function and derivative estimation, we will also use simulated telecommunications network data to demonstrate how these techniques can be used in anomaly detection.

References

J. O. Ramsay and B. W. Silverman (1997). Functional Data Analysis. *Cambridge University Press.*

Subramaniam, G. and Varadhan, R (2007). Feature Extraction using FDA, JSM 2007 (Salt Lake City, UT,USA), Aug. 2007

Subramaniam, G. and Varadhan, R (2008). Borrowing Strength In Time Series Data Mining, JSM 2008 (Denver, CO,USA), Aug. 2008

Pezzak, JC, Norman RW, and Winter DA (1977). An assessment of derivative determining techniques used for motion analysis, *J of Biomechanics*, 10: 377–382.

What useR! deals with : a text mining over the ages

Bertrand Vautier[1*]

1. Student in applied statistics, agrocampus-ouest
* Contact author: bertrand.vautier@free.fr

Keywords: text mining, useR, abstracts, lexical field

Since 2004 useR! Conferences gather contributors, users, programmers and everyone in touch with the R project. Now is the time to sum up the subjects presented at useR!

The aim of this work is to give an overview of the subjects presented at useR! To show the evolution of attendance and of interests over years at useR! This work has been done using the abstract books of 2004, 2006 and 2008 and the power of text mining.

Text mining applied to these abstracts showed interesting results : in addition to the expected growth of attendance, the list of participants nationality got bigger over years and, in the group of the most represented countries, each country was quite specialized on a specific subject. This method allowed to draw a list of lexical fields, which were the main subjects developed at useR!, and it showed a particular result : some subjects were not only separated from the mass of words but also opposed to specific fields.

In this work, text mining showed a great aptitude to summarize more than 400 texts dealing with various subjects. It revealed itself as a powerful analysis tool whose scope goes beyond texts, it analyses also the evolution and policy of R-project and highlights the work of thousands of people.

References

useR!2004 Abstract book (2004).
http://www.r-project.org/conferences/useR-2004/abstracts/Abstracts.pdf

useR!2006 Abstract book (2006).
http://www.r-project.org/useR-2006/Abstracts/Abstracts.pdf

useR!2008 Abstract book (2008).
http://www.statistik.uni-dortmund.de/useR-2008/abstracts/_Abstracts.pdf

satin: a R package for extracting and visualizing satellite data for oceanographic applications

Héctor Villalobos[1,*], Eduardo González-Rodríguez [2]

1. Centro Interdisciplinario de Ciencias Marinas (CICIMAR-I.P.N.)
Av. Instituto Politécnico Nacional s/n, Col. Playa Palo de Sta. Rita. La Paz, B.C.S. 23096. México
2. Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE) Unidad La Paz
Miraflores No. 334 e/ Mulegé y La Paz. Fracc. Bellavista. La Paz, B.C.S. 23050. México. email: egonzale@cicese.mx

* Contact author: hvillalo@ipn.mx

While physical oceanographers program their own functions (mostly in MATLAB©) to handle remote sensing data like sea surface temperature, chlorophyll concentration and sea wind speed, and although other commercial software also exists for this purpose, the R package _sat_ellite _i_mage _n_avigator (_satin_) aims to provide an easy-to-use –yet flexible– set of functions for the non-expert to extract and display level 3 satellite data for use in oceanographic applications. Currently, data from AVHRR, Aqua MODIS, SeaWiFS and QuikSCAT sensors are supported by _satin_, these are provided by NASA as Hierarchical Data Format files (HDF4) as either uncompressed or compressed format. Our extraction functions depend on the package _hdf5_, thus data files must be previously converted to HDF5 format, which can be achieved with the tools provided by the HDF Group (`http://www.hdfgroup.org/h4toh5`). By providing the file name to be read and the geographic limits for the area of interest (from -90 to 90 degrees of latitude and from -180 to 180 degrees of longitude), _satin_ extraction functions return an object of class "list" with longitude and latitude vectors and the corresponding oceanographic parameter matrix rescaled to appropriate units according to attributes in the HDF file. The extracted data are then available for further analysis (e.g. obtaining isotherms) or can be pass to _satin_ display functions to create georeferenced maps with a suitable colour palette and scale bar by specifying a minimum of input arguments. The display functions are flexible enough to allow the use of customized colours and maps for the more experienced users. The use of the package _satin_ is illustrated with examples from the northwestern Mexico area.

Stairstep-like dendrogram cut:
a permutation test approach

Dario Bruzzese[1] , **Domenico Vistocco**[2*]

1. Dip.to di Scienze Mediche Preventive, University of Naples "Federico II"
2. Dip.to di Scienze Economiche, University of Cassino

* Contact author: vistocco@unicas.it

Keywords: Hierarchical Clustering, Permutation Test, Significant Clusters

The output of hierarchical clustering methods is typically displayed as a dendrogram describing a family of partitions indexed by an ultrametric distance. Actually, after the tree structure of the dendrogram has been set up, the most tricky problem is that of cutting the tree with a suitable threshold in order to take out a sub-optimal classification. Several (more or less) objective criteria may be used to achieve this goal, e.g. the deepest step, but most often the partition relies on a subjective choice leaded by interpretation issues. Additionally, whatever the chosen criterion is, only one solution can be obtained for each desired granularity, i.e. the one where clusters are joined at consecutive heights starting from the adopted threshold.

We propose an algorithm, exploiting the methodological framework of permutation test, allowing to find out automatically a sub-optimal partition where clusters do not necessarily obey to the afore-mentioned principle.

Starting from the root node of the dendrogram, a *partial threshold* is moved down the tree until a link joining two clusters is encountered. A permutation test is thus performed in order to verify whether the two clusters must be accounted as a unique group (the null hypothesis) or not (the alternative one). If the null cannot be rejected, the corresponding branch will become a cluster of the final partition and none of its sub-branches will be longer processed. Otherwise each of them will be further visited in the course of the procedure. In fact, in both cases, the *partial threshold* will continue its path and the next branch of the dendrogram will be processed. The algorithm stops when there are no more branches that stand the test (i.e. the null cannot be rejected any more).

The permutation test on which the whole procedure is based can be summarized in this way. Under the *Null*, if all the units belonging to each of the two clusters are mixed up together and then randomly split up, with the only constraint of the group cardinality, the distance among the shuffled clusters should not be very different from the original one. Repeating the shuffling m times, a montecarlo p-value can be computed as the number of permuted distances at least as extreme as the original one.

The algorithm allows us to explore partitions which are not directly achievable using a standard cut-level approach. The obtained partition will be evaluated using several criteria proposed in literature.

References

Good, P. (1994). *Permutation tests : a practical guide to resampling methods for testing hypotheses.* Springer, New York.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. (2005). Cluster Analysis Basics and Extensions. *unpublished.*

Rand, W.M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, December 1971, 66, 336, 846–850.

Interactive R server for teaching statistics

Basile Simon-Vermot[1], Richard Baltensperger[1], Jacques Zuber[1], Pascale Voirin[1]*

1. University of Applied Sciences of Western Switzerland, EIA-FR, Bvd de Pérolles 80, CP 32, 1705 Fribourg
2. University of Applied Sciences of Western Switzerland, HEIG-VD, route de Chéseaux 1, 1401 Yverdon
* pascale.voirin@hefr.ch

Keywords: Teaching, ANOVA, interactive graphics, Rserve, TCL/TK

We present elements developed for an introductory course of statistics such as the concept of ANOVA (Analysis of Variance) and its application to linear regression. It is intended to significantly lower the students' entry barrier to statistical methods.

Students can access to this facility via internet [1] and interactively play with graphics to observe dynamically the impact of changes on the methods used (see Figure 1 below).

Figure 1: Example of an interactive graphic

R is used as a server in background to perform all statistical calculations. We take advantage of the Rserve library and TCL/TK environment for that purpose.

This work is part of developments to facilitate statistics learning [3].

References

[1] Voirin P. (2009). *Cours de statistique: ANOVA*
     http://r-use.hefr.ch/.

[2] Larreamendy- Joerns J., Leinhardt G., Corredor J.(2005)*, The American Statistician, 59(3): 240-251.*

[3] Voirin P., Abou Khaled O., Senn T. (2006). *R as integrated engine in blended learning environments*,
     User!2006 (Vienna, Austria), June 2006.

Electrical Load Forecasting in R

Corinne Walz[1,*] , **Franziska Ziemer**[1,*], **Daniele Amberti**[2,*]

1. Julius-Maximilians-Universität Würzburg, Germany
2. O.R.S., Italy

* Contact author: Corinne.Walz@stud-mail.uni-wuerzburg.de, Franziska.Ziemer@stud-mail.uni-wuerzburg.de, amberti@inwind.it

Keywords: load profiling, electricity, timeseries, forecasting

Due to the liberalization of the European energy markets, electrical load forecasting became very important. To achieve accurate medium term forecasting on an hourly basis, forecast models that integrate previous consumptions as well as exogenous variables (like temperature) are needed.

Currently used approach is a 'Two stage modelling in electrical load forecasting, with application to customer management by power distribution utilities' (Amberti et al, 2006) that uses an autoregressive model with external regressors and a day types approach. This work focuses on model selection especially in terms of forecasting performance as a requirement for usability in a production environment. Solution's implementation is done through time series modelling and forecasting libraries, clustering and model selection R's features as well as original contributions in R.

References

D.Amberti and A.Pievatolo (2006). Two-stage modelling in electrical load forecasting, with application to customer management by power distribution utilities. *Proceedings of the ENBIS Sixth Annual Conference.*

Using R for Sensory Analysis,
including a discussion of the S4 class system

Richard Weeks1*, Oliver Kimberlin1

1.   Mango Solutions, UK
*   Contact author: rweeks@mango-solutions.com

   Mango were requested to design and create an R library to facilitate the
graphical and statistical analysis of Sensory data, primarily from the
CompuSense Sensory software. This library imported different types of Sensory
source data and stored each within a common S4 class implementation. The
library also provides bespoke graphical and statistical methods.

This paper will present the functionality of the resulting library. This paper
will also use this library as a way of presenting the S4 class system, proving
an overview of the use of S4 classes and methods and a discussion around this
functionality.

The Reproducible Computing package

Patrick Wessa[1,*] , Ed van Stee[1]

1. K.U.Leuven Association, Lessius Dept. of Business Studies, Belgium
* Contact author: patrick@wessa.net

Keywords: Reproducible Computing, Research, Compendium

The problem of irreproducible research received a great deal of attention within the academic community [1], [2], [3], [4], [5], [6], [7]. Several solutions have been proposed ([5], [7], [8]) - most prominently there is an R package [9] called "Sweave" ([8]) which allows us to create a so-called Compendium: an integrated collection of text, code, and data (that allows the presented science to be reproduced). All the necessary documents that are needed to create the article (with embedded R code) and the data are contained in an archive file (preferably in tar.gz or zip format). A few examples of such Compendia can be downloaded from [10].

This paper discusses a new approach towards reproducible computing by re-defining the concept of a Compendium as a document where each computation is referenced by a unique URL that points to an object which contains all the information that is necessary to recompute it [11], [12]. The key difference with the original definition of a Compendium is the fact that in our proposal there is a complete separation between text and computing. In other words, each computation (and associated meta data) is stored in a central, web-based repository that can be referenced from any text. Over the last two years, an easy-to-use Compendium Platform was developed (based on this new definition) and implemented in statistics education [11], [13].

The novelty about this paper is the introduction of a newly developed Reproducible Computing package which communicates with the Compendium Platform and allows the R user to do the following:

- store/retrieve image files
- archive/reproduce code snippets
- search archived objects in the repository
- data mining about archived objects
- convert code snippets into R modules (= web applications)
- protect data frames while computations are still reproducible,
- etc...

The bottom line about this package is that it allows the R user to quickly produce reproducible and reusable computations for the purpose of research and publishing.

References

1. J. de Leeuw, "Reproducible Research: the Bottom Line," in Department of Statistics Papers, 2001031101, Department of Statistics, UCLA., 2001, URL http://repositories.cdlib.org/uclastat/papers/2001031101
2. R. D. Peng, F. Dominici, and S. L. Zeger, American Journal of Epidemiology (2006).
3. M. Schwab, N. Karrenbach, and J. Claerbout, Computing in Science & Engineering 2, 61–67 (2000).
4. P. J. Green, The Statistician pp. 423–438 (2003).
5. R. Gentleman, "Applying Reproducible Research in Scientific Discovery," BioSilico, 2005, URL http://gentleman.fhcrc.org/Fld-talks/RGRepRes.pdf
6. R. Koenker, and A. Zeileis, "Reproducible Econometric Research (A Critical Review of the State of the Art)," in Research Report Series, 60, Department of Statistics and Mathematics Wirtschaftsuniversität Wien, 2007.
7. D. L. Donoho, and X. Huo, International Journal of Wavelets, Multiresolution and Information Processing (2004).
8. F. Leisch, "Sweave and beyond: Computations on text documents," in Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria, 2003, ISSN 1609-395X.
9. R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2008), URL http://www.R-project.org, ISBN 3-900051-07-0.
10. Bioconductor, "Publications - bioconductor.org," in http://www.bioconductor.org/pub, 2008.
11. Wessa, P.: Learning Statistics based on the Compendium and Reproducible Computing, Proceedings of the World Congress on Engineering and Computer Science (International Conference on Education and Information Technology), UC Berkeley, San Francisco, USA, 2008
12. Wessa, P.: A framework for statistical software development, maintenance, and publishing within an open-access business model, Computational Statistics, Springer, 2008
13. Wessa, P.: How Reproducible Research Leads to Non-Rote Learning Within a Socially Constructivist E-Learning Environment, Proceedings of the 7th European Conference on e-Learning, 2008

Prediction and Fuzzy Logic at ThomasCook to automate price settings of last minute offers

Jan Wijffels[1,2]

1. ThomasCook Belgium
2. BNOSAC – Belgium Network of Open Source Analytical Consultants – www.bnosac.be
* Contact author: jwijffels@bnosac.be

Keywords: Fuzzy Logic, Predictive modelling, Price effects, rpy2

ThomasCook Belgium provides sun and beach holidays to more than 70 Short Haul destinations around the Mediterranean and some Long Haul destinations in the Americas. During the summer approximately 1.25Mo promotional prices for holiday packages including hotel and flight are put on the market. Flight frequency for some destinations goes up to 5 times a day. The prices of these holiday packages can change on a dayly basis as well in the upwards as the downwards direction.

Bookings on these packages depend on a whole range of factors. Namely: prices, holiday information, flight information, competitor risk, weather risk and cannibalization risk (risk of losing passengers to yourself).

We present a practical user case where we automated the price settings of these promotions. This includes the setting up of large predictive models to evaluate the impact of these influential factors as well as the setup of an expert system based on fuzzy logic which automates the price setting.

We will also cover our experiences using the PL/R PostgreSQL interface and our usage of RPy2 to build a simple GUI to help our Yield department in the interpretation of the automation process.

References

David Meyer, Kurt Hornik (2009). *Generalized and Customizable Sets in R*, Research Report Series / Department of Statistics and Mathematics, Nr. 83, January 2009

Conway Joseph (2007). *PL/R - R Procedural Language for PostgreSQL*, http://www.joeconway.com/plr/

Laurent Gautier (2009). *RPy2, a Python package to connect to R*, http://rpy.sourceforge.net/rpy2/doc/html/index.html

Handling Streaming Data in R Using bigmemory

Rory Winston[1]

1. The Research Kitchen Ltd.

* Contact author: rory@theresearchkitchen.com

Keywords: Real-time, shared memory, bigmemory

Typically, R's strength lies in offline data analysis. However, there is a growing interest in supporting real-time data acquisition and processing. This has potential applications in many fields, from engineering to medicine to finance. In this presentation, we examine the state of the art for real-time data handling in R, and the potential issues when dealing with streaming data. Using the new version of the bigmemory package, we show some concrete examples of handling streaming data across multiple sessions using shared memory. To illustrate, we will see a real-life example from the financial domain.

References

Michael Kane and John Emerson (2009). bigmemory,
    `http://cran.r-project.org/web/packages/bigmemory/`.

Rory Winston (2008). Real-Time Market Data Interfaces in R, Proceedings, User!2008 (Dortmund, Germany), July 2008, 31–37.

Portfolio Analysis and Optimization with R/Rmetrics

Diethelm Würtz[1,2]**, Yohan Chalabi**[1,2]**, Andrew Ellis**[2,3]**, Martin Hanf**[3]

1. ITP ETH, Zurich
2. Rmetrics Association, Zurich
3. Finance Online, Zurich

* Contact author: wuertz@phys.ethz.ch

Modern portfolio theory describes how rational investors will use diversification of investments to optimize their portfolios, and how risky assets should be priced. Financial asset returns are modeled by random variables, and a portfolio is composed as a weighted combination of these assets. A well-accepted mathematical description of portfolio theory was introduced by Markowitz in 1952 as a formal risk/return framework to support investment decision-making. However, there are important limitations to his original formulation, which form the starting point of our investigations.

The underlying assumption of modern portfolio theory states that the measure of investment risk is described by the sample variance of asset returns and that all securities can be adequately represented by a multivariate elliptically contoured distribution. These facts do not always represent the realities of the investment markets, where we are confronted with non-stationary behavior and unusual market behavior, due to structural breaks, bubbles, and even market crashes. Risk is becoming more and more related to bad outcomes and losses, which are considered to weigh more heavily than gains. This view has been put forward by researchers in finance, economics and psychology, which has in turn lead to the introduction of more sophisticated risk measures, such as value-at-risk or shortfall risk. Recent advances in portfolio and financial theory, coupled with today's increased computing power, have overcome some of these limitations. In this talk, we present the implementation of algorithms to support decision-making in portfolio risk analysis and optimization in the framework of the R/Rmetrics software environment. We present the software and selected examples for portfolio design beyond the approach of Markowitz. This includes exploratory data analysis of financial assets, risk measures, robust covariance estimates for portfolios, shortfall risk portfolios, performance analysis and rolling benchmark tests, and we also address the question of how portfolios can be made stress-resistant against unexpected market behavior.

R in Hydrological Modelling: Why we should try it ?

Mauricio Zambrano-Bigiarini[*]

Dep. of Civil and Evironmental Engineering, Università degli Studi di Trento, Trento, Italy, I-38100
* Contact author: mauricio.zambrano@ing.unitn.it

Keywords: Hydrological modelling, application fields

Abstract

R and some of its packages are presented as powerful tools in pre-processing and analysing input data of hydrological models and post-processing its results. Hydrological modelling practitioners spent large amount of time in pre and post-processing data and results with traditional tools. This talk describes how R has been used in almost all the stages of a typical hydrological modelling process on a basin of around 85000 km2, and for 30 years, saving time that can be better spent in doing analysis. Operations made cover the analysis of thousands of raw files with time series of precipitation, temperature and streamflow which are read and organized. Gauging stations to be used in the modelling process are selected according the amount of days with information, missing time series data are filled in using spatial interpolation; time series on the gauging stations are summarized through daily, monthly and annual plots. Input files in dbase format are automatically created in a batch process; results of the hydrological model are read, filtered and compared with observed values through plots and numerical goodness of fit indexes. At the end, the R environment has proved being an effective and promising tool in hydrological modelling.

Escaping RGBland: Selecting Colors for Statistical Graphics

Achim Zeileis[1,*], Kurt Hornik[1], Paul Murrell[2]

1. Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Austria
2. Department of Statistics, The University of Auckland, New Zealand

* Contact author: `Achim.Zeileis@R-project.org`

Statistical graphics are often augmented by the use of color coding information contained in some variable. When this involves the shading of areas (and not only points or lines)—e.g., as in bar plots, pie charts, mosaic displays or heatmaps—it is important that the colors are perceptually based and do not introduce optical illusions or systematic bias. Based on the implementation of the perceptually-based Hue-Chroma-Luminance (HCL) color space in the **colorspace** package, originally written by Ihaka (2003), we have extended the package by new convenient functions for more suitable color palettes in version 1.0-0 (Ihaka *et al.*, 2008). We show how these palettes can be used for coding categorical data (qualitative palettes) and numerical variables (sequential and diverging palettes) in various types of displays (see Zeileis *et al.*, 2009). We also illustrate that it is easier to construct palettes suitable for color-blind viewers (which can be easily assessed using the **dichromat** package, Lumley, 2007).

References

Ihaka R (2003). "Colour for Presentation Graphics." In K Hornik, F Leisch, A Zeileis (eds.), "Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria," ISSN 1609-395X, URL `http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/`.

Ihaka R, Murrell P, Hornik K, Zeileis A (2008). ***colorspace**: Color Space Manipulation.* R package version 1.0-0, URL `http://CRAN.R-project.org/package=colorspace`.

Lumley T (2007). ***dichromat**: Color Schemes for Dichromats.* R package version 1.2-2, URL `http://CRAN.R-project.org/package=dichromat`.

Zeileis A, Hornik K, Murrell P (2009). "Escaping RGBland: Selecting Colors for Statistical Graphics." *Computational Statistics & Data Analysis.* Forthcoming. Preprint available from `http://statmath.wu-wien.ac.at/~zeileis/papers/Zeileis+Hornik+Murrell-2008.pdf`.

Linking R and JUICE

Software tool for analysis and visualization of vegetation data

David Zelený & Lubomír Tichý

Institute of Botany and Zoology, Masaryk University
Kotlářská 2, CZ-61137 Brno, Czech Republic
e-mail: zeleny@sci.muni.cz, tichy@sci.muni.cz

JUICE 7.0

is a stand-alone Windows application for editing, classification and analysis of large phytosociological tables and other ecological data. It has been developed by Lubomír Tichý (Masaryk University Brno) since 1998, written in Visual Basic and optimized for Windows OS (but operating also under Linux via Wine emulator). JUICE can also automatically or semi-automatically operate other commercial or freeware softwares, such as PC-ORD, CANOCO, SYNTAX or Mulva. For more information and free download, visit http://www.sci.muni.cz/botany/juice/ or http://tinyurl.com/cwtrf9.

Linking JUICE and R offers quick and effective availability of selected R analytical and visualization functions: JUICE offers user friendly environment of vegetation data handling and import, R works as an engine for analysis and visualization.

Recently implemented functions are multivariate analysis of vegetation data (DCA, PCA, NMDS) and modeling of species response curves (GLM, GAM, Gaussian and HOF models).

R-project

Following R packages are used by JUICE:
• *vegan* (Oksanen, Kindt, Legendre, O'Hara et al.) - number of tools for analysis of ecological communities, diversity and multivariate analysis;
• *gravy* (Oksanen) - modeling of species response curves along environmental gradients;
• *rgl* (Adler & Murdoch) - 3D interactive ordination diagrams;
• *geometry* (Grasman & Gramacy) -calculation of convex hulls;
• *tcltk* and *tkrgl* (Murdoch & Chen) -user interface for interaction with figures;
• *ordijuice* and *srcjuice* (Zelený) - connecting JUICE and R and offering some additional analytical and visualization tools.

Technical solutions

Link between JUICE and R is mediated by sharing the files via R/bin directory and using libraries *ordijuice* (for unconstrained ordinations) and *srcjuice* (for species response curves). JUICE offers wizard for selection of appropriate analysis parameters and type of graphical output. Data are saved into R/bin file, and R is initialized to run in a batch mode by sending the script with selected parameters into Windows command line. R draws the figure onto graphical device (X11 or rgl) and opens Tcl/Tk panel; after either the device or Tcl/Tk bar is closed, R is terminated. This design allows several R session to run simultaneously using the same data, but different parameters. Other option is R running in silent mode: the result of analysis and figures are saved back to R/bin directory, where it's awaited by JUICE, and R is terminated.

Updates of ordijuice and srcjuice libraries are available from R repository in http://www.sci.muni.cz/botany/zeleny/R. If in use, once per day is checked for new updates and user is noticed about the need to update.

Use of R in Genome-wide Association Studies

Jing Hua Zhao[1*] , Qihua Tan[2]

1. MRC Epidemiology Unit, Institute of Metabolic Science, Cambridge, UK
2. Odense University Hospital, Denmark

Keywords: Bioinformatics, Biostatistics, Genome-wide Association Studies, Complex Traits

Recent GeneChip and sequencing technologies have made it possible to use ~1 million or more single-nucleotide polymorphisms (SNPs) in large-scale genetic epidemiological studies. They are the most common genetic variants in human genome and their association with complex traits in relation to the environment is the subject of genome-wide association studies (GWASs), through which important variants have been successfully identified for complex traits ranging from anthropometric measurements, etiology and progression of common diseases, drug response to diversity and evolution of human populations. However, there is still a considerable scope for advancing these initiatives.

In this presentation, we provide an overview of the background and issues in design and analysis for such studies, as well as their connection with international collaborative projects and consortium work. We give real examples to illustrate how the R statistical and programming environment has been used, and discuss the extent to which this could be further developed. We believe that GWAS makes a strong case of being a motivation and inspiration for development of analytical and computational tools and that it also facilitates a vigorous inter-disciplinary collaboration between researchers in substantive areas such as biology, genetics, mathematical statistics and computing.

References

Altschuler D, MJ Daly, ES Lander (2008). Genetic mapping in human disease. *Science*, 322, 881–888.

Donnelly P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature*, 456, 728–731.

Elston RC, M Ann Spence. (2006). Advances in statistical human genetics over the last 25 years. *Statistics in Medicine*, 25, 3049–3080.

Pearson TA, TA Manolio. (2008). How to interpret a genome-wide association studies. *JAMA*, 299, 1335–1344.

Zhao JH, Q Tan (2006). Integrated analysis of genetic data with R. *Human Genomics*, 2, 258-265.

Index