Recent Events

Statistical Computing 2003 at Reisensburg

The Reisensburg meeting has become a regular attraction for those interested in Computational Statistics. It is organized by three special interest groups (*Computational Statistics* of the Biometric Society -DR, *Statistical Analysis Systems* of the German Association of Medical Informatics, Biometry and Epidemiology GMDS, and *Classification and Data Analysis in the biosciences* of the Gesellschaft für Klassifikation GfKl) and it takes place near Ulm, Germany, in beautiful Reisensburg castle, situated above the river Danube.

The main topics of this conference are fixed one year in advance by the members of the working groups. The organizers take great care that there is sufficient time for discussion after the talks and at the famous Reisensburg bartizan round tables.

Recent developments of statistical software has been a major topic of previous meetings. Merits and miseries of the various packages were discussed in depth. This has changed. Discussion of the large packages played a minor role this year, and R was featured in many presentations. F. Bretz, T. Hothorn and P. Westfall gave an overview on the `multcomp` package for multiple comparisons. F. Leisch introduced `flexmix`, a framework for fitting discrete mixtures of regression models.

As we all know, a lot has still to be done in R to support advanced visualization and interactivity. A. Zeileis, D. Meyer and K. Hornik demonstrated visualizations using mosaic plots in R . S. Urbanek showed Java based interactive graphics for R and H. Hoffmann demonstrated what can be done in other environments, using visualizations for conditional distributions as an example and tools derived from the Augsburg Dada collection.

The list of speakers and topics is too long to be repeated here in full. Program and abstracts are available from http://www.dkfz.de/biostatistics/Reisensburg2003/.

The 2003 meeting highlighted the analysis of genomic data. Robert Gentleman presented a keynote session on exploring and visualizing genomic data. Subsequent sessions covered methodological aspects, in particular techniques for combining classifiers or variables and methods related to machine learning. On the more applied side, topics included, among others, C. Ittrich and A. Benner addressing the role of microarrays in clinical trials, and U. Mansmann discussing simulation techniques for microarray experiments. E. Brunner (Göttingen) used the opportunity to demonstrate how classical statistical analysis of the design of experiments may be applied in this field to give concise answers instead of vague conjectures.

High dimensional observations, combined with very low sample sizes, are a well known peculiarity of genomic data. Another peculiarity comes from the strong dependence between the observed data. The data refer to gene activities, and these are only an aspect of the metabolic and regulatory dynamics of a cell. Little is known about how to include the knowledge of metabolic pathways and the resulting dependencies in a statistical analysis. Using statistical inference from genomic data to identify metabolic or regulatory structures is largely an open task. F. Markowetz and R. Spang studied the effect of perturbations on reconstructiing network structure; C. Becker and S. Kuhnt addressed robustness in graphical modelling. From the application side, A. v. Heydebreck reported on estimation of oncogeneic tree models and W. Huber talked about identification of protein domain combinations.

The next Reisensburg working conference will take place 2004, June 27.-30. By the time you read this article, the call for papers should have been issued. The main topics will be: applications of machine learning; statistical analyis of graphs/networks; statistical software; bioinformatics; exploration of large data sets.

Till then, working groups in cooperation with the special research unit in Erlangen will organize a workshop on Ensemble Learning, Erlangen 2004, Jan. 23.-24. Stay tuned, and see http://www.imbe.med.uni-erlangen.de/links/EnsembleWS/.

Günther Sawitzki
Universität Heidelberg
gs@statlab.uni-heidelberg.de

Statistical Inference, Computing and Visualization for Graphs

On August 1–2, 2003, a workshop on using graphs in statistical data analysis took place at Stanford University. Quoting the workshop homepage at http://www.research.att.com/~volinsky/Graphs/Workshop.html *"Graphs have become an increasingly popular way of representing data in many different domains, including telecommunications research, genomics and bioinformatics, epidemiology, computer networks and web connectivity, social networks, marketing and statistical graphical models. Analyzing these data effectively depends on contributions from the areas of data representation, algorithms, visualization (both static and interactive), statistical modeling (including graphical models) and inference. Each of these areas has its own language for describing graphs, and its own favorite tools and methods. Our goal for the workshop is to explore*

synergies that may exist between these different areas of research."

It was very interesting to see the number of different areas of applied data analysis in which graphs (structures with nodes and edges) are used. There are differences, most notably the sizes of the graphs, ranging from a dozen nodes to several millions, which has an impact on "natural" and efficient computations. However, we also identified commonalities, and having a central infrastructure in R for representing graphs and performing common operations will certainly help to prevent reinventing the wheel several times.

The Bioconductor project has started to provide this infrastructure with the **graph** package and interfaces to standard libraries for graph computations and visualization (**Rgraphviz**, **RBGL**, ...). Development versions of ggobi also have support for graphs that can be tightly linked to R. If you are interested to learn more about the workshop: you can download the slides for any of the presentations from the workshop homepage.

Finally, I want to thank the organizers for the great job they did in organizing the workshop; both the scientific program and the social atmosphere made it a pleasure to participate.

Friedrich Leisch
Technische Universität Wien, Austria
Friedrich.Leisch@R-project.org

JSM 2003

At the 2003 Joint Statistical Meetings in San Francisco, an invited session was organized that is of particular interest to the R community. Jan de Leeuw from University of California, Los Angeles, led off the session with the a talk on "The State of Statistical Software" (http://gifi.stat.ucla.edu/pub/jsm03.pdf). He began with a overview of types of statistical software one might use for activities such as consulting, teaching and research providing some history and thoughts for the future along the way. Luke Tierney from University of Iowa, spoke on "Some New Language Features of R" (http://www.stat.uiowa.edu/~luke/talks/jsm03.pdf) focussing on namespaces, code analysis tools, exception handling and byte compilation. Duncan Temple Lang from Bell Laboratories spoke on "Connecting Scientific Software" (http://cm.bell-labs.com/stat/duncan/Talks/JSM2003). The talk dealt with connecting other software packages to R, with particular attention to R DCOM services. The discussant, Wolfgang Hartmann from SAS, provided an industry perspective (see http://www.cmat.pair.com/wolfgang/jsm03.pdf) comparing the features of different software, commercial and open-source, with specific attention to R.

Balasubramanian Narasimhan
Stanford University, CA, USA
naras@stat.stanford.edu

gR 2003

On 17-20th September 2003, Aalborg University hosted a workshop bringing together people from many communities working with graphical models. The common interest is development of a package for R, supporting the use of graphical models for data analysis. The workshop followed up on the gR initiative described by Steffen Lauritzen in R News 2/3.

The workshop provided a kaleidoscope of applications as well as insight in experiences dealing with practical graphical models. The applications presented were from the areas of epidemiology, geostatistics, genetics, bioinformatics and machine learning.

The wide range of applications and methodology showed that a unifying software package for graphical models must be widely extensible and flexible — utilizing a variety of data formats, model specifications and estimation algorithms. The package should also provide an attractive user interface that aids in working with complex models interactively.

Development of a gR-package is evolving at many levels. Some 'old' stand-alone programs are being ported as R-packages (CoCoR, BRugs), some are being interfaced (mimR, JAGS, BugsR), while others have been developed in R (ggm, deal, GRAPPA).

Experiences from other existing packages can inspire the gR project. For example, the Bayes Net Toolbox for Matlab includes many features that gR will include. Intel is currently re-implementing the Bayes Net Toolbox in C++ (called Probability Network Library, PNL) and plan a December 2003 release, expected to be open source. An R interface to PNL could be a possibility.

During the workshop an outline of a package **grbase** with basic elements was discussed and thought to become a common ground for extensions. Important features were to separate data, model and inference. The **grbase** package will include

- support for a variety of data formats, eg. as a list of cases, a dataframe or a database connection. It should also be possible to work with a model without data.
- a general model language capable of specifying eg. (block-) recursive graphical models and BUGS models.
- a variety of representation forms for graphs, eg. using/extending the **graph** package from bioconductor.

- a graphics system, for interactively working with models. For example using **R-Tcl/Tk**, **Rggobi** or the R-interface to Graphviz.
- an analyzing unit that combines data and model with the possibility of using different inference algorithms in the analyzing step.

A minimal version of **grbase** is planned for January 2004.

An invited session concerned with the gR developments is being planned for the Joint Statistical Meeting in Toronto, 8-12 August 2004.

See http://www.math.auc.dk/gr/gr2003/ for more information about the workshop and related links, including links to the aforementioned software.

Acknowledgments The gR-2003 workshop was supported by the Danish National Research Foundation Network in Mathematical Physics and Stochastics - MaPhySto. The Danish activities of the gR project are supported by the Danish Natural Science Research Council.

Claus Dethlefsen
Aalborg University, Denmark
dethlef@math.auc.dk

Book Reviews

John Maindonald and John Braun: Data Analysis and Graphics Using R — An Example-based Approach

Cambridge University Press, Cambridge, United Kingdom, 2003
362 pages, ISBN 0-521-81336-0
http://cbis.anu.edu/DAAG/
http://www.stats.uwo.ca/DAAG/

The aim of the book is to describe the ideas and concepts of many statistical methodologies, that are widely used in applications, by demonstrating the use of R on a number of examples. Most examples in the book use genuine data collected by the authors in their combined several decades of statistical consulting experience. The authors see the book as a companion to other books that include more mathematical treatments of the relevant theory, and they avoid mathematical notation and mathematical description of statistical methods. The book is aimed at both scientists and students interested in practical data analysis. Data and new R functions used in the book are included in the DAAG package available from the authors' web sites and through the Comprehensive R Archive Network (CRAN).

The book begins with a nice summary of the contents of the twelve chapters of the book. Chapter 1, *A Brief Introduction to R*, provides enough information on using R to get the reader started. Chapter 2, *Style of Data Analysis*, demonstrates with many examples the use of R to carry out basic exploratory data analysis involving both graphical and numerical summaries of data. The authors not only describe how to create graphs and plots but also show the reader what to look for in the data summaries and how to interpret the summaries in the context of each particular example. Chapter 3, *Statistical Models*, describes the authors' view on the importance of models as a framework for statistical analysis. Chapter 4, *An Introduction to Formal Inference*, introduces the basic ideas of random sampling and sampling distributions of statistics necessary to understand confidence intervals and hypothesis testing. It also includes chi-square tests for contingency tables and one-way ANOVA.

The next several chapters demonstrate the use of R to analyze data using linear models. Chapter 5, *Regression with a Single Predictor*, Chapter 6, *Multiple Linear Regression*, Chapter 7, *Exploiting the Linear Model Framework*, and Chapter 8, *Logistic Regression and Other Generalized Linear Models*, use increasingly complex models to lead the reader through several examples of practical data analysis.

The next three chapters discuss more specialized topics that arise frequently in practice. Chapter 9, *Multi-level Models, Time Series, and Repeated Measures*, goes through examples that use more complicated error structures than examples found in previous chapters. Chapter 10, *Tree-based Classification and Regression Trees*, provides an introduction to tree-based regression and classification modeling. Chapter 11, *Multivariate Data Exploration and Discrimination*, describes both principle components analysis and discriminant analysis.

The final chapter, Chapter 12, *The R System — Additional Topics*, is a far more detailed introduction to R than that contained in the initial chapters. It is also intended as a reference to the earlier chapters.

The book is a primer on the nuts-and-bolts use of R for the types of statistical analysis that arise commonly in statistical practice, and it also teaches the reader to think statistically when interpreting the results of an analysis. The strength of the book is in the extensive examples of practical data analysis with complete examples of the R code necessary to carry out the analyses. Short R commands appear on nearly every page of the book and longer R code examples appear frequently as footnotes.