gRaphical Models in R

A new initiative within the R project

Steffen L. Lauritzen

What is this?

In September 2002 a small group of people gathered in Vienna for the brainstorming workshop gR 2002 with the purpose of initiating the development of facilities in R for graphical modelling. This was made in response to the facts that:

- graphical models have now been around for a long time and have shown to have a wide range of potential applications
- software for graphical models is currently only available in a large number of specialised packages, such as BUGS, CoCo, DIGRAM, MIM, TETRAD, and others.

The time has come to integrate such facilities in general software, such as R, with flexible extension and modification of prepackaged modules. The workshop web page can be found on http://www.ci.tuwien.ac.at/Conferences/gR-2002/.

Summary of workshop

Two rather separate clusters of activities could be identified. In one, model selection and identification based on i.i.d. repetitions was the main issue and in the second, the primary issue was modularity in modelling and computation for complex patterns of observations.

Although some effort would be needed to accomodate both of these aspects, the aim of the initiative is to do so and get software from both of these clusters into R.

Further research is needed to make R move from computing within models to computing directly with 'abstract' models as objects, which seems necessary to represent the natural modularity of graphical models.

It was decided to take the following simple steps immediately:

WWW A web-page for the project has been set up at http://www.R-project.org/gR/.

SIG A special interest group for the gR project was formed with the associated mailing list R-sig-gR@lists.r-project.org. See the gR project web-page for subscription information.

DSC 2003 A session at the Distributed Statistical Computing workshop, taking place in Vienna in the period 20-22 March 2003, will be devoted to gRaphical models and the gR project.

Software Some software for graphical models was already integrated or is easily integrable in R and these packages would as quickly as possible be made available through CRAN. This includes

- **deal** for learning Bayesian networks (C. Dethlefsen and S. G. Bøttcher) is already available as an R package;
- the extensive program CoCo for analysis of discrete data (J. H. Badsberg);
- an interface making it possible to access MIM within R (S. Højsgaard);
- GRAPPA, a suite of R functions for probability propagation (P. Green).

gR 2003 A larger workshop is tentatively planned in Aalborg, Denmark, in September 2003.

Graph computations A special interest group had already been formed with the purpose of creating a module under R for computation with graphs. Such a module will be extremely valuable for the gR project.

Organisation Kurt Hornik will act as main contact between gR and the R Core team, and Claus Dethlefsen, Aalborg University, will serve as the maintainer of the CRAN view for R.

Steffen L. Lauritzen
Aalborg University, Denmark
steffen@math.auc.dk

Recent and Upcoming Events

R at the ICPSR summer program

R played a prominent role at the 2002 ICPSR Summer Program. Headquartered at the University of Michigan, the Inter-University Consortium for Political and Social Research (ICPSR) is an international organization of more than 400 colleges and universities. The Consortium sponsors a variety of services and activities, including an extensive social-science data archive and a highly regarded Summer Program in Quantitative Methods of Social Research.

The eight-week 40th edition of the ICPSR Summer Program attracted more than 700 participants—mostly graduate students and faculty in the social sciences—to Ann Arbor, Michigan to attend 30 courses. These courses ranged from the elementary to the advanced, most presented in intensive one-week, all-day classes, and in four-week, two-hours-per-day classes. Additionally, participants attended lectures on a variety of statistical topics. More information, including course outlines, is available at the ICPSR web site, http://www.icpsr.umich.edu/.

Computing in the ICPSR Summer Program has traditionally been eclectic, employing a wide range of statistical software. This year, several relatively advanced courses coordinated their use of R. In support of these courses and to gain more exposure for R among social scientists, I taught a two-week, two-hours-per-evening lecture on Statistical Computing in S, which featured R. More than 100 participants attended these lectures. R was installed in the ICPSR Windows-based computer labs, and a CD/ROM with R for Windows and Windows binaries for all of the packages on CRAN was made available to participants.

Four-week courses that employed R included Bayesian Methods for Social and Behavioral Sciences, taught by Jeff Gill of the University of Florida; Linear, Nonlinear, and Nonparametric Regression, taught by Bob Andersen, then of Oxford University, now of the University of Western Ontario; and Maximum Likelihood Estimation for Generalized Linear Models, taught by Charles Franklin of the University of Wisconsin. The combined enrollment of these classes was more than 100. In addition, Bill Jacoby of the University of South Carolina gave several lectures on statistical graphics and data visualization which featured R.

John Fox
McMaster University, Canada
jfox@mcmaster.ca

R featured at JSM

A good time was had at the Joint Statistical Meetings (JSM) in New York City, August 11–15, 2002. After several years of dismal locations (who can forget Dallas?), this year's JSM began a series of venues possibly even more interesting than the Meetings: in the next two years, the JSM will be in San Francisco and Toronto.

R was featured prominently at the JSM this year. In session 190, "The future of electronic publication: Show me ALL the data," organized by Brian Yandell and chaired by David Scott, a number of members of the R core team discussed R and related tools. Friedrich Leisch described StatDataML, an XML-based markup language for statistical data for the more easy transfer of data between programs, and Sweave, a system for creating statistical reports that combine text and code such that data analysis output can be created and inserted on the fly. Robert Gentleman emphasized that all statistical papers should be accompanied by sufficient data and software so that the results may be reproduced, a concept he called a "compendium." Such a compendium could be created as an R package containing a Sweave document. The key issue will likely be in distribution. Kurt Hornik described his experience in managing the R repository, which now contains over 165 packages. Central to the proper curation of such a software repository is the development of tools for automated testing, especially as the core system is updated. Duncan Temple Lang described a system for distributing verifiable, self-contained, annotated computations with interactive facilities so that readers may examine and explore the content on their own. The session concluded with a discussion by James Landwehr concerning the status of the electronic publication of the ASA's journals.

Session 349, "R Graphics," organized and chaired by Paul Murrell, included discussions of an R interface to OpenGL, the **scatterplot3d** package, and the integration of R graphics within Excel. In addition, Deborah Swayne demonstrated the GGobi data visualization system, a more modern version of xgobi, which allows multiple, linked graphical displays.

By my count, at least 22 of the technical sessions at the JSM included some discussion of the analysis of data from gene expression microarrays. The importance of the Bioconductor project (initiated by Robert Gentleman, and consisting largely of R packages for the analysis of microarray data) for statisticians working in the area was made clear. In particular, the R package **affy** (maintained by Rafael Irizarry), for the analysis of data from Affymetrix chips, was frequently mentioned.

Karl W. Broman
Johns Hopkins University
kbroman@jhsph.edu

DSC 2003

The third international workshop on *Distributed Statistical Computing (DSC 2003)* will take place at the Technische Universität Wien in Vienna, Austria from 2003-03-20 to 2003-03-22. This workshop will deal with future directions in (open source) statistical computing and graphics.

Topics of particular interest include

- Bioinformatics

- Database Connectivity
- Graphical Modeling
- GUIs and Office Integration
- Resample and Combine Methods
- Spatial Statistics
- Visualization

Emphasis will be given to the R (http://www.R-project.org/), Omegahat (http://www.omegahat.org/), and BioConductor (http://www.bioconductor.org/) projects. DSC 2003 builds on the spirit and success of DSC 1999 and 2001, which were seminal to the further development of R and Omegahat.

Deadline for registration is 2003-03-14. There will be a conference fee of EUR 200 for 'early' registrations made before 2003-02-14, and EUR 250 for registrations made afterwards.

On 2003-03-19 there will be several half-day tutorials, with topics currently including

- An Introduction to BioConductor
- Exploring Genomic Data using R and BioConductor
- R Graphics
- Writing R Extensions

Fees for each tutorial are EUR 50 (academic) or EUR 250 (non-academic).

Please contact the organizing committee at dsc-org@ci.tuwien.ac.at for further information.

Friedrich Leisch
Universität Wien
leisch@R-project.org

Computational and Statistical Aspects of Microarray Analysis

This one week intensive school is intended to give a clear view of current statistical and computational problems linked to microarray data along with some solutions. This self-contained course will touch on many aspects of genome biology as it applies to microarray analysis. Topics include preprocessing, estimating gene expression levels, microarray data and hybridization, experimental design, dimension reduction and pattern recognition techniques including boosting, bagging and other recent statistical techniques for microarray data analysis.

The course is primarily intended for PhD students and researchers in the areas of Statistics, Biology and related fields. A small background on data analysis is required. The course is computationally intensive and laboratory sessions are associated with methodology ones.

It will take place at the University of Milan, Italy, from 2003-05-26 to 2003-05-30 with two morning sessions on methodology and computer labs in the afternoon of each day. The course will be given by Anestis Antoniadis (Universite Joseph Fourier, Grenoble, France) and Robert Gentleman (Harvard School of Public Health, Boston, USA). Further information is available at http://www.eco-dip.unimi.it/marray.

Stefano Iacus
University of Milan
stefano.iacus@unimi.it

Editor-in-Chief:
Kurt Hornik
Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien
Wiedner Hauptstraße 8-10/1071
A-1040 Wien, Austria

Editorial Board:
Friedrich Leisch and Thomas Lumley.

Editor Programmer's Niche:
Bill Venables

Editor Help Desk:
Uwe Ligges

Email of editors and editorial board:
firstname.lastname@R-project.org

R News is a publication of the R Foundation for Statistical Computing, communications regarding this publication should be addressed to the editors. All articles are copyrighted by the respective authors. Please send submissions to regular columns to the respective column editor, all other submissions to Kurt Hornik, Friedrich Leisch, or Thomas Lumley (more detailed submission instructions can be found on the R homepage).

R Project Homepage:
http://www.R-project.org/

This newsletter is available online at
http://CRAN.R-project.org/doc/Rnews/