SCIENTIFIC AND STATISTICAL COMPUTING WITH R

A. Mani

Member, Calcutta Mathematical Society
a.mani@member.ams.org
Homepage: http://amani.topcities.com

BARCAMP 2009, IIT KGP KOLKATA CENTER

ABSTRACT

The R environment is an integrated suite of FOSS facilities for data manipulation, knowledge discovery from databases, data analysis, scientific computing and graphical display. Technically R is an interpreted functional programming language with object-oriented programming features like inheritance. I intend to introduce some basic aspects of R, outline more advanced features and speak about providing scientific and statistical computing services over it.

R caters to statistics and other methods of knowledge discovery like PCA, fuzzy techniques, data clustering, neural networks, DNA computing, belief networks, general mathematical modelling, econometrics and of course to the various sciences including bioinformatics.

Contents

What is R?

R is

- A Free and Open Source Software under the GNU-GPL
- A fully planned and coherent system unlike other data analysis software
- A vehicle for developing new methods of interactive data analysis
- Partly based on the statistical programming language **S**, and **Scheme**
- Very usable in cluster and parallel computing environments
- Endowed with its own LaTeX-like documentation format, which is used to supply comprehensive documentation.
- Easily extendable with new packages.
- In R, subroutines have the ability to modify/ construct other subroutines and evaluate the result as an integral part of the language

TECHNICAL FEATURES

- is an interpreted functional programming language
- has all object-oriented programming features
- Allows branching and looping as well as modular programming using functions.
- Allows linking with procedures written in the C, C++, or FORTRAN languages for efficiency.
- Is compatible with computationally intensive libraries like **Lapack** and **Blas**
- Under most compilation environments, compiled code dynamically loaded into **R** cannot have breakpoints set within it until it is loaded.
- Rd files (**R** documentation files) can be processed with any language for special functionalities.
- **R** scripts are almost perfectly portable.
- Stand-alone executables can be created for a particular subset.

GETTING STARTED

- Installation: Planning
- Compiling from Source vs The Package Manager
- IDE: RCommander + Kate, Emacs, Rkward
- Commands: > install.packages() > R CMD INSTALL lmtest.tar.gz
- File Formats: foo.R, out.txt, foo.Rda
- Types can be acquired or forced upon, but are not declared

SIMPLE CODE EXAMPLE

```
> dat ← read.table("clustr.csv", header=TRUE, sep=",", na.strings="NA")
> names(dat)
[1] "X" "BLUE" "RED" "YELL" "WHITE" "GREY" "PINK" "BROWN"
> dat$X
[1] ACIDIC AGRESSIVE ANXIOUS ASIATIC ATTRACTIVE
> x ← dat$BLUE    > y ← dat$RED    > ab ← lm(y x, data=dat)
> names(ab)
"coefficients" "residuals" "effects" "rank" "fitted.values" "assign" "qr"
"df.residual" "xlevels" "call" "terms" "model"
>summary(ab) [summary is a Generic Function]
>ab$coef
```

CONSTRUCTS

- > zwi ← function(x) ts(dat[,x], start=1, deltat=1/12))
- > lapply(2:10, zwi)
- > tapply(x, y, mean)
- > library(lattice)
- > splom(dat[,-1], type="l")
- > yscf ← tapply(xf, xp, function(g) sum(ifelse(diff(g)>0,diff(g)*100000,0)))

Syntax Notes

- Function calls (expressions),Infix and prefix operators, Index constructions, Compound expressions, Flow control elements, Function definitions
- Matrix product: %*%, Tensor product: %x%, Set Membership: %in%
- seq(-1, 1, by=.5)*1:3 ; [1] -1.0 -1.0 0.0 0.5 2.0
- NA, NAN, INF
- > attach(dataframe): like NameSpaces
- browser(): for Breakpoints in code

INTEGRATION WITH DATABASES

- R by itself cannot handle data objects $>$ 500MB with ease
- The core R engine supports persistent, but not concurrent access to data
- The .RData format is TeX like.
- Requirements: Fast access to smaller parts of large datasets
- Secure concurrent access and updates. Client-Server Model
- More: `http://cran.r-project.org/doc/manuals/R-data.html`

DATABASE SYSTEMS

- SQLite, MySQL, *PostgreSQL, Oracle: via ODBC or specific interfaces
- BerkeleyDB (hash tables)
- RSQLite, RODBC, RMySQL, Rdbi and RdbiPgSQL

```
> library(RMySQL)
> con ← dbConnect(dbDriver("MySQL"), dbname = "MySQLTest")
> dbListTables(con)
> data ← dbGetQuery(con, "SELECT * FROM Foo")
> dbDisconnect(con)
```

Object Oriented Programming

Class	A self-contained set of attributes and methods of an object. S3, S4 Classes, R.oo, Proto Extensions
Instance	The actual object created at runtime
Method	An Object's Abilities
Inheritance	'Subclasses' are more specialized versions of a class, which inherit attributes and behaviors from their parent classes.
Message passing	The process by which an object sends data to another object or asks the other object to invoke a method is interfacing
Encapsulation	is meant to conceal functional details of a class from objects that send messages to it

OOPs References

- 3 Levels of OOPs
- 'show'- S4 class. 'plot'-S3 class
- S3 Scheme: > help(NextMethod)
- S4 Scheme: > help(Methods), setGeneric, setMethod, setClass
- R.oo, Proto (Packages)
- Chambers,J.M.: *Software for Data Analysis* Springer'2008

DEBUGGING

- To stop, use undebug(<function>)
- > debug(mean.default)
- > mean(1:10)
- debugging in: mean.default(1:10) debug: <SNIP>.....
- Other Functions: trace, untrace, traceback
- To See Code: > lm, methods('predict')
- For Profiling: system.time, Write Code within Rprof() and Rprof(NULL) and use R CMD Rprof Rprof.out

R WEB INTERFACES

- Rpad
- Rweb, ROnline, Rcgi, CGIwithR - uses the CGI interface for I/O: `http://www.omegahat.org/CGIwithR`
- RApache: R inside Apache $\geq$2. `http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/RApacheProject`
- Rserve implements a TCP/IP server, which allows other programs to use facilities of R
- R-php, Rwui- the latter creates a web UI for R scripts

CREATING PACKAGES

>source("redct.R")
>package.skeleton("redct",c("funrou1","funrout2", "dat"))
>#package.skeleton(name = "", list,environment = .GlobalEnv,path = ".", force = FALSE, namespace = FALSE, code_files = character())
Will create

- redct, redct/man, redct/man/README, redct/man/funrou1.Rd
- redct/man/funrou2.Rd, redct/src, redct/src/README, redct/R
- redct/R/funrou1.R, redct/R/funrou2.R, redct/data, redct/data/dat.rda, redct/DESCRIPTION, redct/README

DESCRIPTION FILE

- Package: redct
- Title: Computes reducts in multi-source approximation contexts
- Version: 1.0
- Date: 2008-06-05
- Author: A. Mani
- Maintainer: A. Mani (a.mani.cms@gmail.com)
- Description: Implements two algorithms for the computation of reducts in multi-source approximation contexts.
- License: GNU GPL Version-3

EDITING DOCUMENTATION FILES

You will need to fill in all of the following fields:
* \name{}, \alias{}, \title{ }, \description{ }
*\usage{redct(x)}: with all arguments and defaults
*\arguments{ \item{} { Describe \ redct(x) here }
*\details{ Long Description }, \value{ Desribe the list of values returned
\item {*abc*} (Description of 'abc')
*\references{ to literature }, \author{}, \note{}, \ seealso{ other packages }
*\examples{WORKING ONES with random or custom data }, \keywords{ONE PER LINE }

Final Steps

- R CMD build redct # Then Debug
- R CMD check redct # For Examples
- R CMD build redct

ESSENTIALS: Simple Project

- Competency in the subject(s) in question
- For example, R with its packages cannot teach you data clustering
- Detailed flowcharts for the work-flow in the project
- Detailed Information about all relevant data sets and scope of the project
- Identify relevant packages
- Some contexts require review of research literature as well

NEXT STEPS

- Be prepared for deficiences in supplied specifications
- Revise specifications, method, relevant packages and more
- Estimate Required System Resources
- Compute Cost
- License: Avoid signing anything restrictive for R development
- Optimize Documentation and Reports

MIGRATION

- From Other Statistical Softwares: SPSS, SPLUS, Matlab
- Mathematical Software: Octave, Matlab, Mathematica, GAP, HOL
- Databases
- Other Econometrics and Financial Software
- Generally Easy

CHEERS!