Introduction to R

Eric Feigelson

Dept. of Astronomy & Astrophysics

Center for Astrostatistics

Penn State University

edf@astro.psu.edu

Space Telescope Science Institute statistics mini-course

Fall 2011

The R statistical computing environment

- R is the public-domain version of the commercial S-Plus statistical computing package. Integrates data manipulation, graphics and statistical analysis. Uniform documentation and coding standards.

- Fully programmable C-like language, similar to IDL. Specializes in vector or matrix inputs; not designed for maps, images or movies.

- Easily downloaded from http://www.r-project.org with Windows, Mac or UNIX binaries.

- Tutorials available in dozens of books (most since 2005) and on-line.

- >2600 user-provided add-on packages collected in Comprehensive R Archive Network http://www.cran.r-project.org.

Some functionalities of R

Base R

arithmetic & linear algebra, bootstrap resampling, empirical distribution tests, exploratory data analysis, generalized linear modeling, graphics, robust statistics, linear programming, local and ridge regression, maximum likelihood estimation, multivariate analysis, multivariate clustering, neural networks, smoothing, spatial point processes, statistical distributions & random deviates, statistical tests, survival analysis, time series analysis

Selected methods from Comprehensive R Archive Network (CRAN)

Bayesian computation & MCMC, classification & regression trees, genetic algorithms, geostatistical modeling, hidden Markov models, irregular time series, kernel-based machine learning, least-angle & lasso regression, likelihood ratios, map projections, mixture models & model-based clustering, nonlinear least squares, multidimensional anlaysis, multimodality test, multivariate time series, multivariate outlier detection, neural networks, non-linear time series analysis, nonparametric multiple comparisons, omnibus tests for normality, orientation data, parallel coordinates plots, partial least squares, periodic autoregression analysis, principal curve fits, projection pursuit, quantile regression, random fields, random forest classification, ridge regression, robust regression, self-organizing maps, shape analysis, space-time ecological analysis, spatial analyisis & kriging, spline regressions (MARS, BRUTO), tessellations, three-dimensional visualization, wavelet toolbox

Interfaces: BUGS, C, C++, Fortran, Java, Perl, Python, Xlisp, XML

I/O: ASCII, binary, bitmap, cgi, FITS, ftp, gzip, HTML, SOAP, URL

Graphics & emulators: Grace, GRASS, Gtk, Matlab, OpenGL, Tcl/Tk, Xgobi

Math packages: GSL, Isoda, LAPACK, PVM

Text processor: LaTeX

Since c.2003, R has been the premier public-domain statistical computing package.

History of R

Late 1980s: John Chambers developed C-basedprogrammable statistical analysis system, **S**

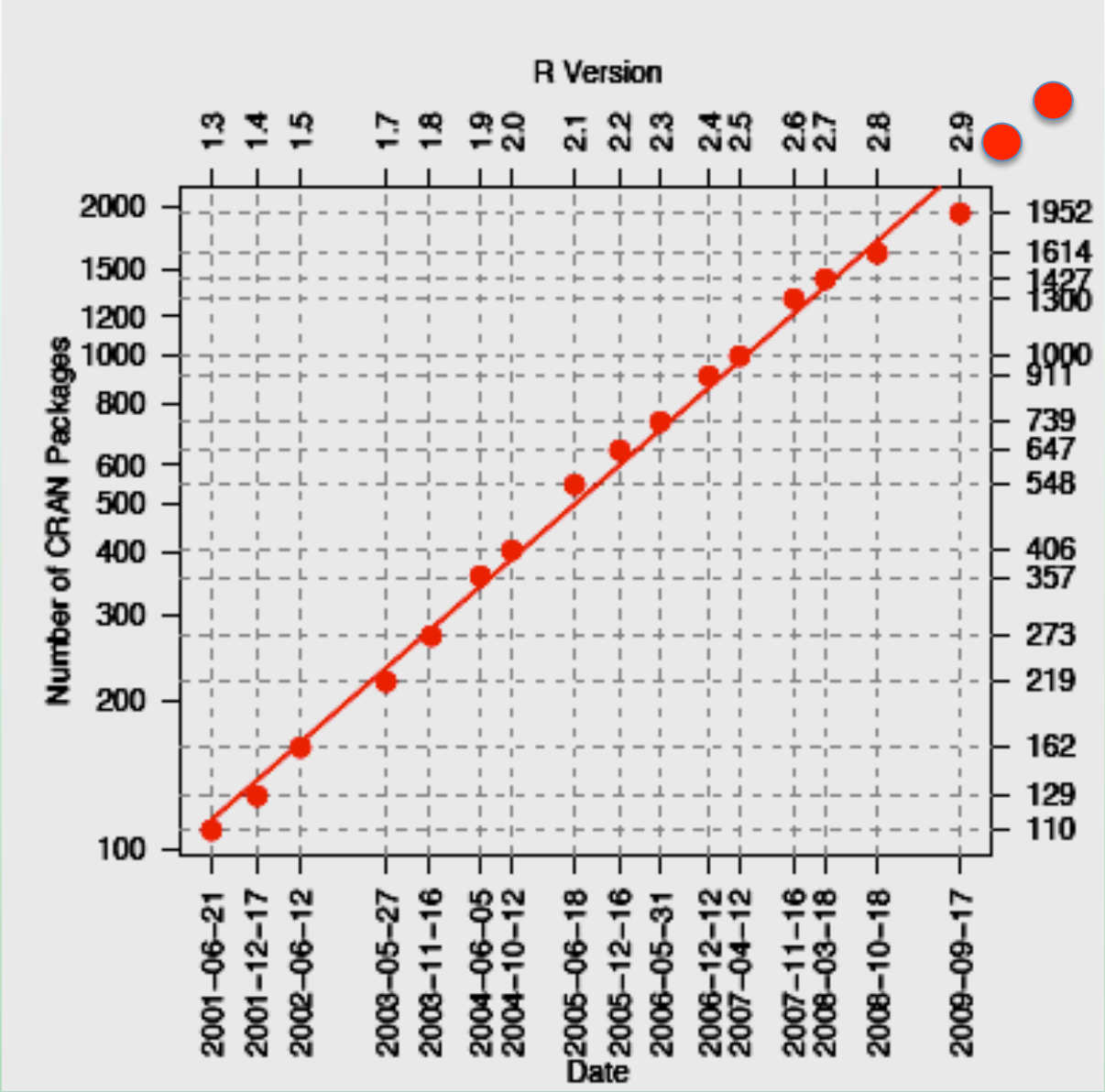
Early 1990s: Ross Ihaka & Robert Gentleman mimic **S** in an GNU/GPL system, **R**

Late 1990s: **R** core development group expands, **CRAN** opened for contributed packages

2000s: Dramatic growth in content (esp **CRAN**) and usage (~2M users by 2010)

Today: **R** is the principal software environment for the development and promulgation of new statistical methodology. Dominates ASA Section on Statistical Computing, J. Stat. Comput., etc.

Growth of CRAN contributed packages



Oct 1, 2011 count:

3,320 packages

Aspects of the social organization and trajectory of the R project, J. Fox, ***The R Journal***, 1/2, 5 (2009)

Some features of R

- Designed for individual use on workstation, exploring data interactively with advanced methodology and graphics. Very similar experience to IDL.

- **R** objects placed into `classes': *numeric, character, logical, vector, matrix, factor, data.frame, list,* and dozens of others designed by **CRAN** packages. *plot, print, summary* functions are adapted to class objects. The *list* class allows a hierarchical structure of heterogeneous objects (like IDL *sav* file).

- Extensive graphics based on SVG, RGTK2, JGD, and other GUIs. See graphics gallery at http://www.oga-lab.net/RGM2.

- Uni- or bi-directional interfaces to other languages: BUGS, C, C++, Fortran, Java, JavaScript, Matlab, Python, Perl, Xlisp, Ruby.

- Only one astronomy **CRAN** package to date: FITSio (limited functionality)

Computational aspects of R

R scripts can be very compact

IDL: temp = mags(where(vels le 200. and vels gt 100, n))

*uq = temp((sort(temp))(ceil(n*0.75)))*

R: *uq = quantile(mags[vels>100. & vels<200.], probs=0.75)*

Vector/matrix functionalities are fast (like C); e.g. a million random numbers generated in 0.1 sec, a million-element FFT in 0.3 sec.

Some **R** functions are much slower; e.g.

for (I in 1:1000000) x[i] = x[i-1] + 1

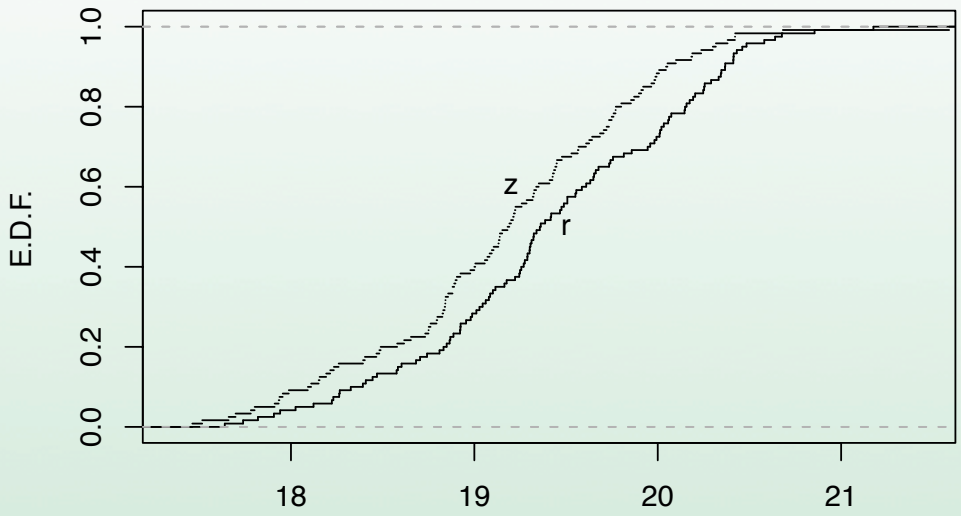
The **R** compiler is now being rewritten from `parse tree' to `byte code' (similar to Java & Python) leading to several-fold speedup.

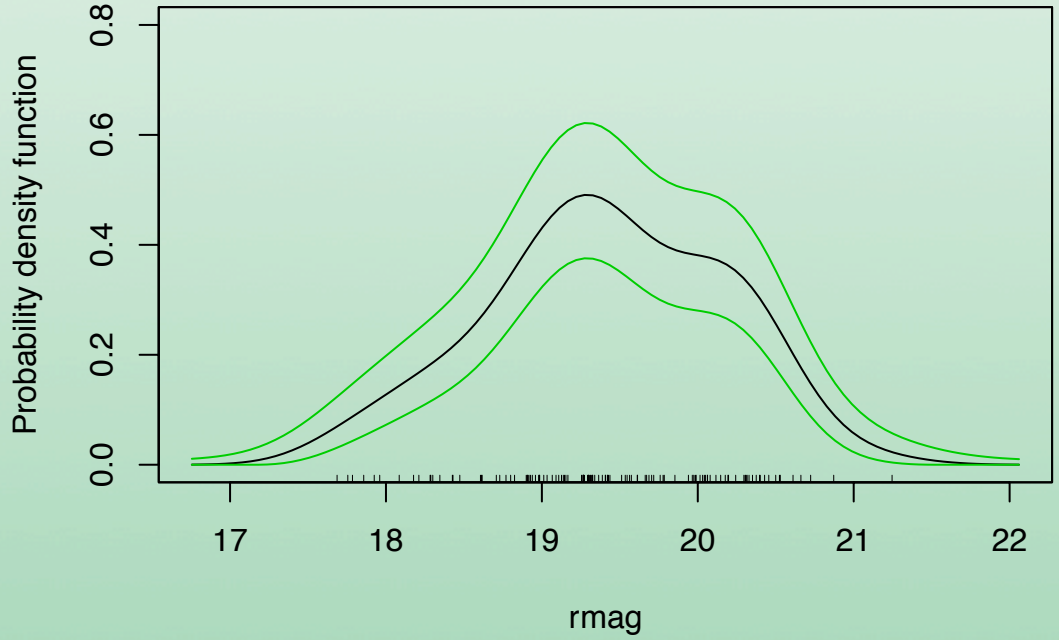
Several dozen **CRAN** packages are devoted to high-performance computing, parallelization, data streams, grid computing, GPUs, (PVM, MPI, NWS, Hadoop, etc).

Sample R Script

```
setwd('/Users/e5f/Desktop')

# Read dataset of 120 SDSS quasar r & z band magnitudes
qso <- read.table('http://astrostatistics.psu.edu/datasets/SDSS_QSO.dat', head=T)
dim(qso)
names(qso)
summary(qso)
rmag <- qso[1:120,9]
zmag <- qso[1:120,13]

# Plot e.d.f. with confidence bands
install.packages('sfsmisc')
library('sfsmisc')
ecdf.ksCI(rmag)

# Plot e.d.f.'s
plot(ecdf(rmag),cex.points=0, verticals=T, main='', xlab='Magnitude', ylab='E.D.F.')
plot (ecdf(zmag), cex.points=0, add=T)
text(19.5,0.5,lab='r') ; text(19.2,0.6,lab='z')
dev.copy2eps(file='R_test.eps')

# Are the shapes the same?
wilcox.test(rmag,zmag, conf.int=T)
wilcox.test(rmag,zmag+0.249, conf.int=T)
```

```
# Plot histograms and kernel density estimators
hist(rmag,breaks=30)
plot(density(rmag, bw=bw.nrd0(rmag)))

# Plot k.d.e. with confidence bands
install.packages('sm') ; library(sm)
help('sm.density')
sm.density(rmag) ; tt <- sm.density(rmag)
lines(tt$eval.points,tt$upper,col=3) ;
lines(tt$eval.points,tt$lower,col=3)
dev.copy2eps(file='R_test2.eps')
```


E.D.F.
0.0
0.2
0.4
0.6
0.8
1.0
z
r
18
19
20
21
Magnitude


Probability density function
0.0
0.2
0.4
0.6
0.8
17
18
19
20
21
22
rmag

Selected books on R

Modern Statistical Methods for Astronomy with R Applications
E. Feigelson & G. J Babu 2012

An Introduction to R (online at CRAN mirror sites under `Manuals', dozens of tutorials)

Introductory Statistics with R P. Dalgaard, 2nd ed. 2008

R in a nutshell: A desktop quick reference J. Adler 2009

The R Book, M. Crawley 2007

A Handbook of Statistical Analyses Using R, B. S. Everitt & T. Hothorn 2nd ed, 2009

Software for data analysis: Pogramming with R, J. Chambers 2008

Introductory Time Series with R Cowpertwait & A. V. Metcalfe 2009
(one of dozens in Springer ***Use R!*** series)

ggplot2: Elegant Graphics for Data Analysis H. Wickham 2nd ed, 2009

+ ~1 additional arriving monthly at Penn State's libraries

Possible topics for stat/**R** tutorials

- Nonparametric statistics
- Density estimation (data smoothing)
- Regression
- Multivariate analysis
- Multivariate classification (data mining)
- Censoring & truncation (nondetections)
- Time series analysis
- Spatial point processes

Let us browse the R Web site:

http://r-project.org