Crash Course: R and Exploratory Data Analysis for Multivariate Data

Susan Holmes

Bio-X and Statistics

October 12, 2009

What is R?

- The S programming language developed by John Chambers at Bell Labs in 1976 to turn ideas into software. Developed after they made C.
  S was designed to allow people to do statistical analysis without having to write programs in a language like Fortran.
- R is an open source version of the S language described by Chambers et al. in the blue book. R was written initially by Robert Gentleman and Ross Ihaka and released under the GPL in 1995.
- Objects are geared towards visualization and open methods.

First Steps: Help

```
> help("plot")
> help("for")
> library(help = "stats")
> help(package = "stats")
> help.search()
> help.start()
> RSiteSearch("t test")
> apropos("package")
> help(mean)
> example(mean)
```

Getting Started: Data from the inside

```
> data()
> library(vsn)
> data(kidney)
> fit = vsn2(kidney)                          ## fit
vsn2: 8704 x 2 matrix (1 stratum).
Please use 'meanSdPlot' to verify the fit.
> meanSdPlot(fit)
> nkid = predict(fit, newdata=kidney)  ## apply fit
> plot(exprs(nkid), pch=".")
> abline(a=0, b=1, col="red")
```

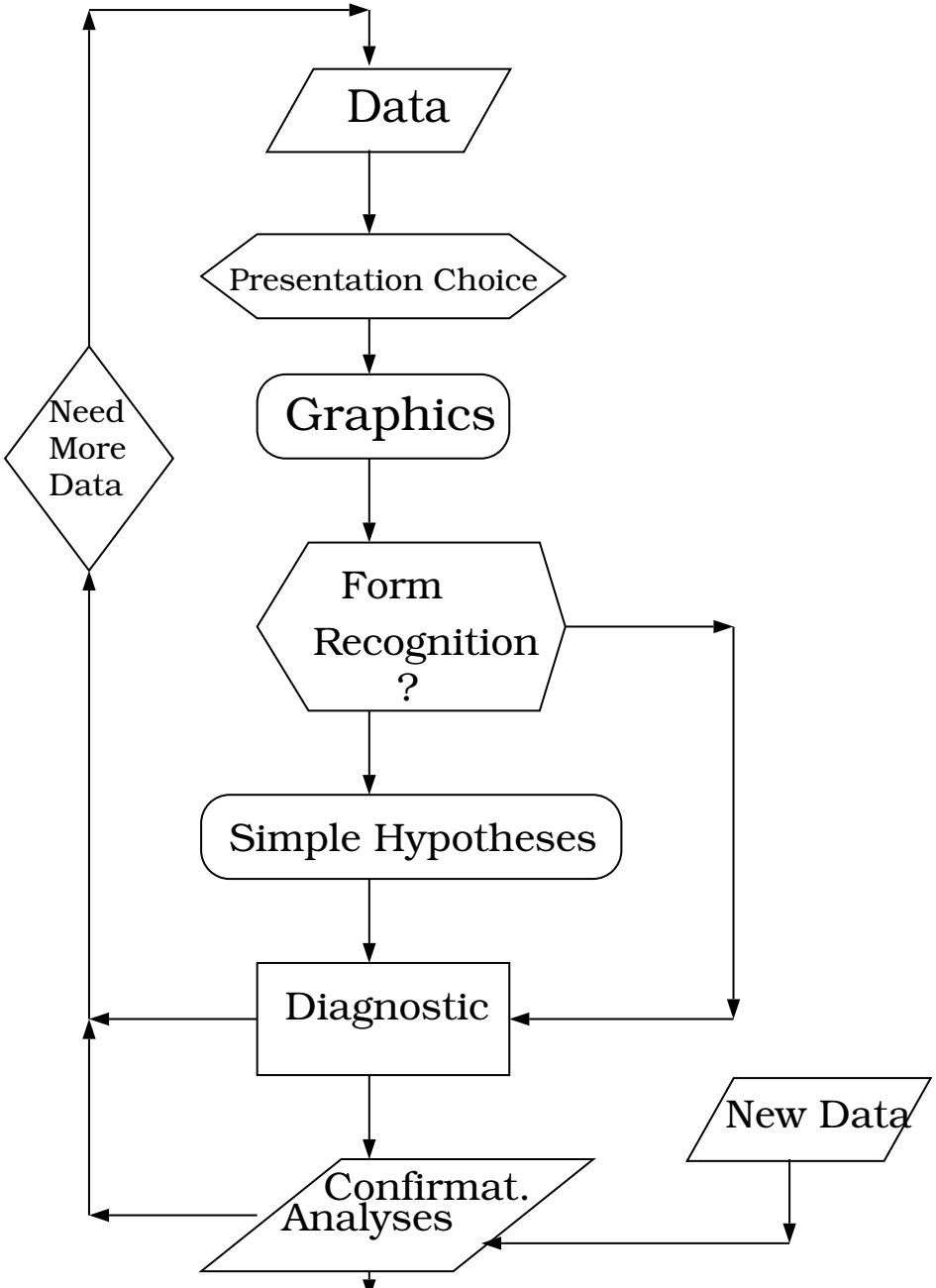
Getting Started: Data from the outside

```
>table1=read.table("Msig4.ascii")
>Norm=read.csv("Norm.csv")
>Msig4=read.delim2("Msig4.txt",sep=" ")
```

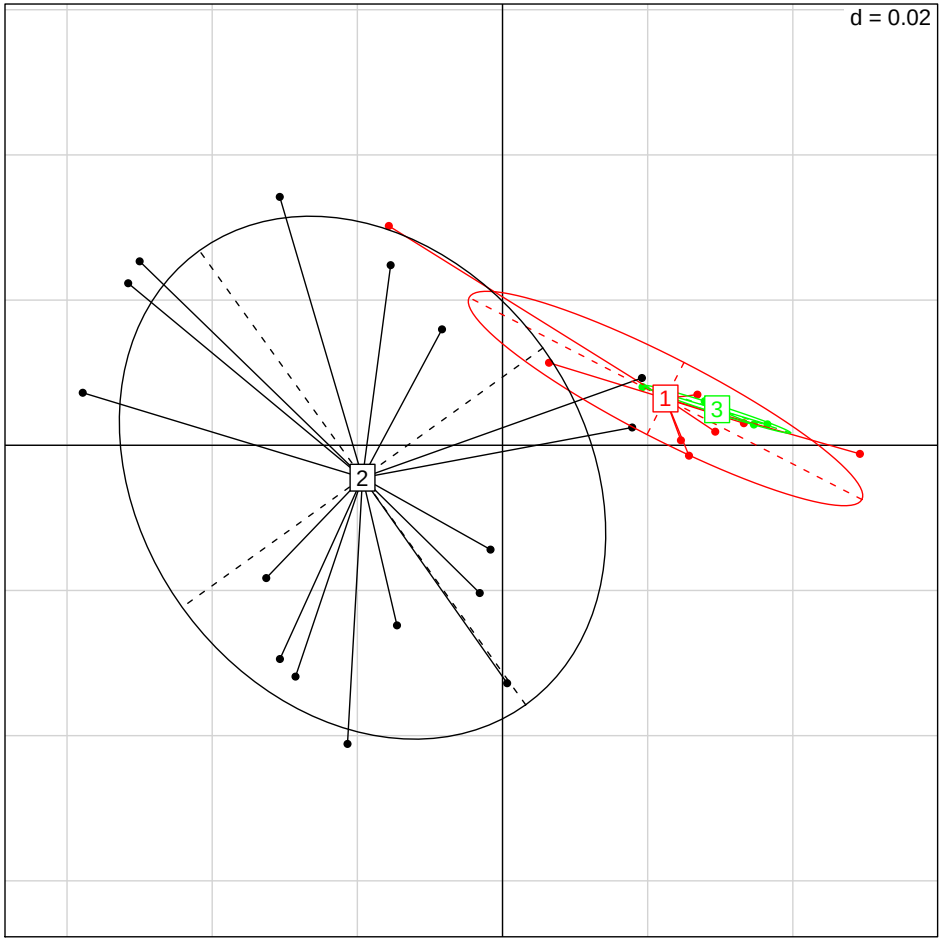
Getting Started: Data, specialized

```
library(Biobase)
RG = read.maimages(fileName, source = "agilent", columns=li
 Gb="gBGMedianSignal",Rb="rBGMedianSignal"),
 other.columns=list(gnonunif="gIsFeatNonUnifOL",rnonunif="r
 gIsSaturated="gIsSaturated",
 rIsSaturated="rIsSaturated",
 gIsFeatNonUnif="gIsFeatNonUnif",
 rIsFeatNonUnif="rIsFeatNonUnif",
 gIsFeatPopnOL="gIsFeatPopnOL",
 rIsFeatPopnOL="rIsFeatPopnOL"))
 allcdfs=readAffy()
```

What is EDA?


Data
Presentation Choice
Graphics
Need More Data
Form Recognition ?
Simple Hypotheses
Diagnostic
New Data
Confirmat. Analyses

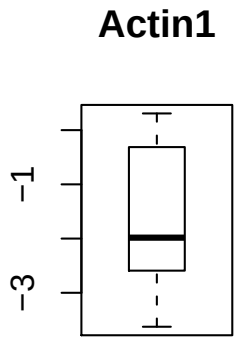
Discovery by Visualization



Basic Visualization Tools

- Stem and Leaf Plots
- Boxplots, histograms.
- Scatterplots.
- Scatterplots with binning.
- Smoother Scatterplots.
- High dimensional plots, by projection.
- Hierarchical Clustering.

```
-3 | 6
-3 | 0
-2 | 977666
-2 | 3333210
-1 | 99775
-1 |
-0 | 75
-0 | 331
 0 | 01133
```



How to Look at continuous variable distributions.

- Histograms
- Normal Probability Plot (`qqplot`, `qqnorm`).
- Residual Plots.
- Glyph Plots.

Data Pre-processing

- Dealing with missing data.
- Centring, Scaling, Transforming Data.
- Combining variables to simplify output.
- Using meta-information.

EDA and Data Mining: the downside

- Data Snooping and Pattern Finding.
- Multiple Testing.
- Validation: separate data, bootstrap and cross-validation.

Modern Statistics

- Multivariate
- Robust
- Computer Intensive
- Non parametric
- Bayesian

A systematic review of Multivariate Methods

Dichotomy of data-mining and multivariate statistical methods into two groups:

- Special Status for one variable or set of variables. Regression, multiple response regression, discriminant analysis (LDA), analysis of variance (ANOVA), Redundancy Analysis (RA) depending on whether the explanatory variables are categorical or continuous.
- Same status for all variables, some may be distances, some may be categorical some may be continuous.

Table of Methods for Studying Links between Variables

Techniques	Vars. to explain (response)	Explanatory Var.
Multiple Regression	1 continuous	p continuous
Analysis of Variance	1 continuous	p categorical
Analysis of Covariance	1 continuous	p_1 continuous, p_2 categorical
Correspondence Analaysis	1 categorical	1 categorical
Canonical Correl. Analysis	q continuous	p continuous
Redundancy Analysis	q continuous	p continuous
Discriminant Analysis	1 categorical	p continuous
Multidimensional Analysis of Variance (MANOVA)	p continuous	p categorical
Multidimensional Analysis of Covariance	p continuous	p_1 categorical p_2 continuous
Regression Tree	1 continuous	p_1 continuous, p_2 categorical
Classification Tree	1 categorical	p_1 continuous, p_2 categorical

Table of Methods for Representing Data

Techniques	Variables
Principal Components	p continuous
Multiple Correspondence Analysis	p categorical or
	p categorical and q continuous
Multidimensionnal Scaling (PCoA)	categoricals and continuous
Double PCoA	distances and distances
Clustering (either hierarchical or not)	continuous and categorical

Other dichotomies are possible:

- Bayesian/Frequentists.
- Parametric/ Nonparametric.
- Robust.
- Supervised/ Unsupervised.
- Exploratory/ Confirmatory.