MULTIVARIATE DATA FOR METAGENOMICS USING R

Susan Holmes ©
http://www-stat.stanford.edu/~susan/

Bio-X and Statistics

SISMID14-Lecture 7, June 2011

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

```
>read.table()
>read.csv()
>read.delim2()
```

Getting Started: Data from qiime or phylochip

Rectangular data + Side Information.

- with a number of taxa or species (often as rows of the table).
- with a certain number of samples/patients recorded as columns of the table.
- Phylogenetic/family relationships between the rows of the table.
- Extra clinical/environmental information about the samples.

Challenges of this type of data: heterogeneity

'Homogeneous data are all alike;

all heterogeneous data are heterogeneous

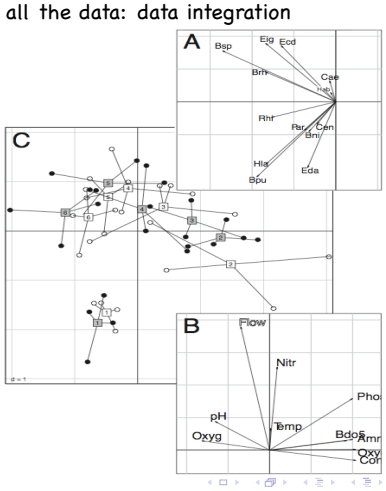
in their own way.'

Heterogeneity

- Status : response/ explanatory.
- Hidden (latent)/measured.
- Type :
  - Continuous
  - Binary, categorical
  - Graphs/ Trees
  - Images
  - Maps/ Spatial Information
  - Rankings
- Amounts of dependency: independent/time series/spatial.

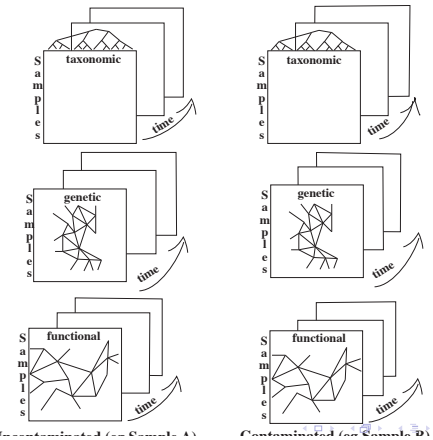
Goals in Modern Biology: Systems Approach

Look at the data/ all the data: data integration

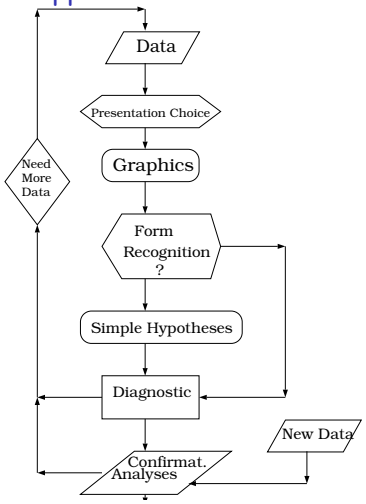


Goals in Modern Biology: Systems Approach

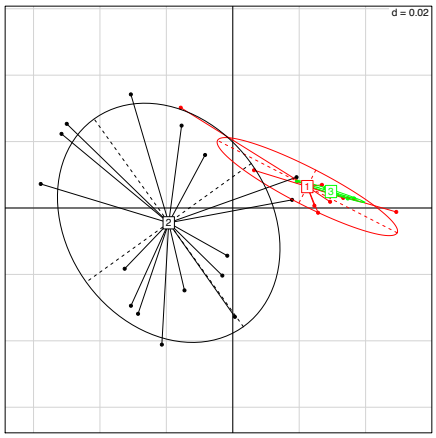
Look at the data/ all the data: data integration



First Approach: EDA?



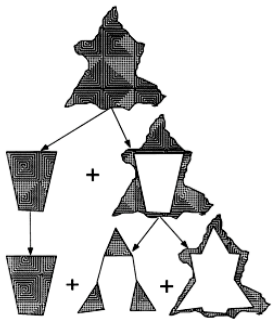
Discovery by Visualization



Basic Visualization Tools

- Boxplots, barplots.
- Scatterplots of projected data.
- Scatterplots with binning variable
- Hierarchical Clustering, heatmaps, Phylogenies.
- Combination of Phylogenetic Trees and data.

Iterative Structuration (Tukey, 1977)



A distance-> projection

Modern Statistics: Non parametric, multivariate

- Exploratory Analyses: hypothesis generating.
  - Projection Methods (new coordinates)
    Principal Component Analysis
  - Principal Coordinate Analysis-Multidimensional Scaling (PCO,MDS)
  - Correspondence Analysis
  - Discriminant (C) Analysis
- Tree based methods
  - Phylogenetic Trees
  - Clustering Trees
  - Decision Trees
- Confirmatory Analyses: Hypothesis verification.
  - Permutation tests (Monte Carlo).
  - Bootstrap (Monte Carlo).
  - Bayesian nonparametrics (Monte Carlo).

Modern Methods: Robust Methods

Variance
Variability of one continuous variable $\longrightarrow$ the variance.
NOT ROBUST, low breakdown.
Solution: Take ranks, clumps, logs, or trimming (``windsorize'') the data.

```
Example (uses package psych):
>>>>>>>>>>>>>>>>>>>>>>>>>>>
wm=winsor(tab1,trim=0.05)
```

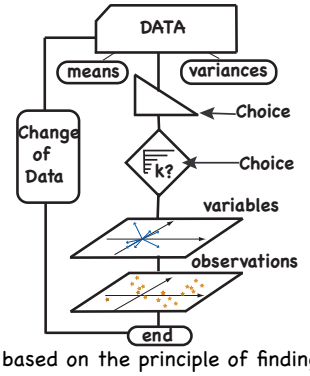
Part I

EDA: *Exploratory Data Analysis*

Data Checking
Hypothesis Generating

One table methods: PCA, MDS, PCoA, CA, ...

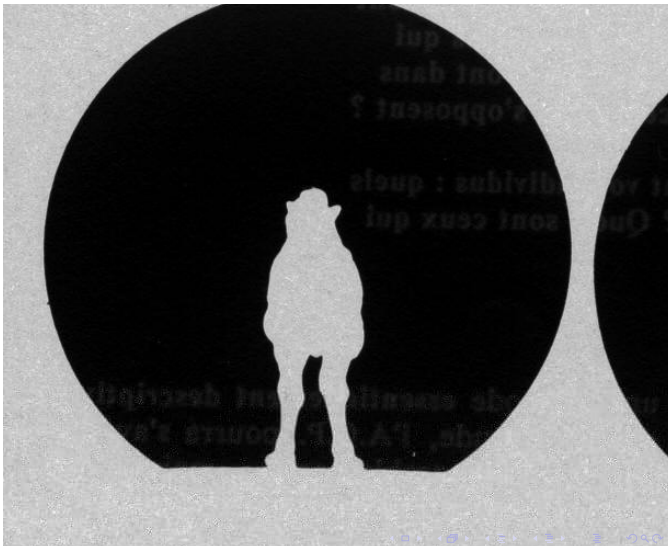
.

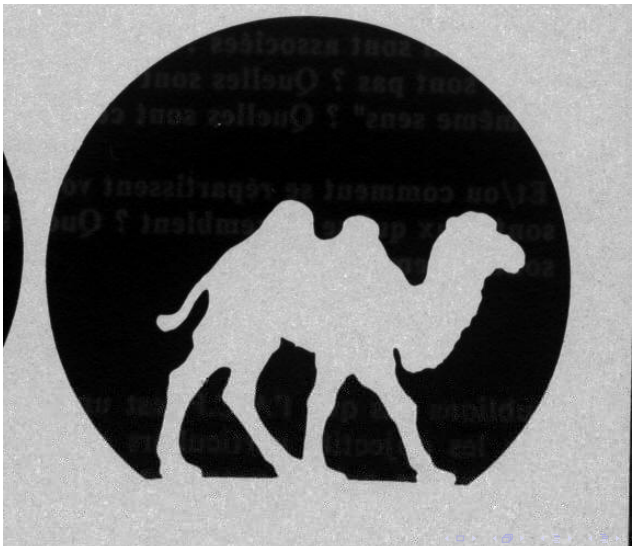


All based on the principle of finding the largest axis of inertia/variability.

New Variables/coordinates from old or distances

Best Projection Directions??



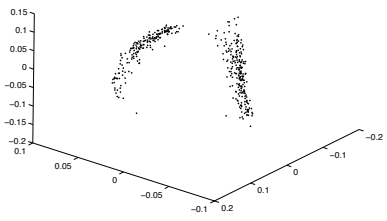


Ordination Methods

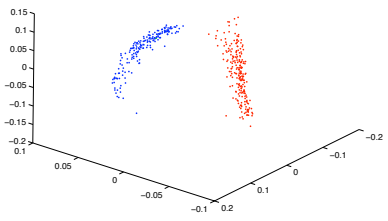
Many discrete measurements $\longrightarrow$ Gradients.
Data from 2005 U.S. House of Representatives roll call votes.
We further restricted our analysis to the 401 Representatives that voted on at least 90% of the roll calls (220 Republicans, 180 Democrats and 1 Independent) leading to a 401×669 matrix V of voting data.

The Data

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	−1	−1	1	−1	0	1	1	1	1	1
2	−1	−1	1	−1	0	1	1	1	1	1
3	1	1	−1	1	−1	1	1	−1	−1	−1
4	1	1	−1	1	−1	1	1	−1	−1	−1
5	1	1	−1	1	−1	1	1	−1	−1	−1
6	−1	−1	1	−1	0	1	1	1	1	1
7	−1	−1	1	−1	−1	1	1	1	1	1
8	−1	−1	1	−1	0	1	1	1	1	1
9	1	1	−1	1	−1	1	1	−1	−1	−1
10	−1	−1	1	−1	0	1	1	0	0	0



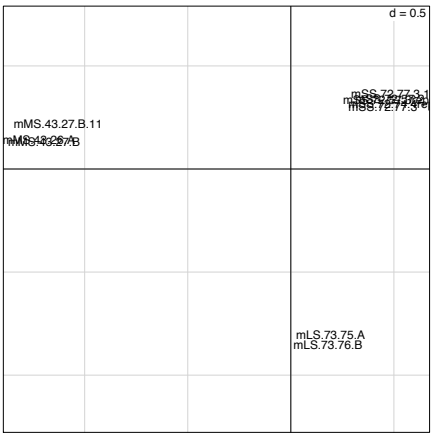
3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes.



3-Dimensional MDS mapping of legislators based on the 2005 U.S. House of Representatives roll call votes. Color has been added to indicate the party affiliation of each representative.

Projection Methods

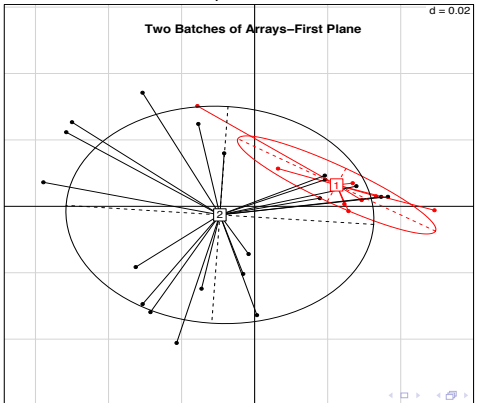
Project Various Factors as class labels onto the first few coordinates (components, factors,...).
Projection of supplementary group centers (means) and ellipses of variation (variance) as in the function s.class in the ade4[?] package
Example: Explore batch effects in the laboratory methods used to generate the data.
See quality of replicates



Projection of a categorical variable on a PCA

In the case of PCA: The ellipses are computed using the means, variances and covariance of each group of points on both axes, and are drawn with these parameters: the center of the ellipse is centered on the means, its width and height are given by the variances, and the covariance sets the slope of the main axis of the ellipse.

First two batches (in black and red)
(although both balanced with regards to IBS and healthy rats) were extremely different in variability and overall multivariate location. Batches were done different days with different sets of arrays.



A third batch was generated with the same arrays as batch 2 but the same experimental protocol as batch 1. The third group faithfully overlaps with batch 1 thus showing that the batch effect was not due to a difference in arrays but to the experimental protocol. This shows the utility of PCA in quality