R: statistics for all of us

R: an international statistical collaboratory

Prepared for part of the symposium on
Multivariate Statistical Methods in Individual Differences Research
International Society for the Study of Individual Differences
Biennial meeting, Adelaide, July , 2005

William Revelle, Northwestern University
personality-project.org/r/

R: statistics for all of us

- What is it?
- Why use it?
- Common (mis)perceptions
- Examples for personality and individual differences research

R: What is it?

- R: An international collaboration
- R: the open source - public domain version of S+
- R: written by statisticians (and all of us) for statisticians (and the rest of us)
- R: an extensible language

Common statistical programs

General	Specialized
R	AMOS
S+	EQS
SAS	LISREL
SPSS	Plus
STATA	Mx
SYSTAT	your favorite program.

Common statistical programs most are costly

General	Specialized
R	AMO$
$+	EQ$
A	LI$REL
$P$$	Maple$
$TATA	Mx
YTAT	your favorite program.

R: a way of thinking
(from the R point of view)

- "R is the lingua franca of statistical research. Work in all other languages should be discouraged."
- "This is R. There is no if. Only how."
- "Overall, SAS is about 11 years behind R and S-Plus in statistical capabilities (last year it was about 10 years behind) in my estimation."

Taken from the R.-fortunes (selections from the R.-help list serve)

But it is open source - how can you trust it?

- Q: When you use it [R], since it is written by so many authors, how do you know that the results are trustable?
- A: The R engine [...] is pretty well uniformly excellent code but you have to take my word for that. Actually, you don't. The whole engine is open source so, if you wish, you can check every line of it. If people were out to push dodgy software, this is not the way they'd go about it.

Taken from the R.-fortunes (selections from the R.-help list serve)

What is R? : Technically

- R is an open source implementation of S (S-Plus is a commercial implementation)
- R is available under GNU Copy-left
- The current version of R is 2.1.1
- R is group project run by a core group of developers (with new releases semiannually)
- (Adapted from Robert Gentleman)

R: History

- 1991-93: Ross Dhaka and Robert Gentleman begin work on R project at U. Auckland
- 1995: R available by ftp under the GPL
- 96-97: mailing list and R core group is formed
- 2000: John Chambers, designer of S joins the R core (wins a prize for best software from ACM for S)
- 2001-2005: Core team continues to improve base package
- Many (>400) others contribute "packages"

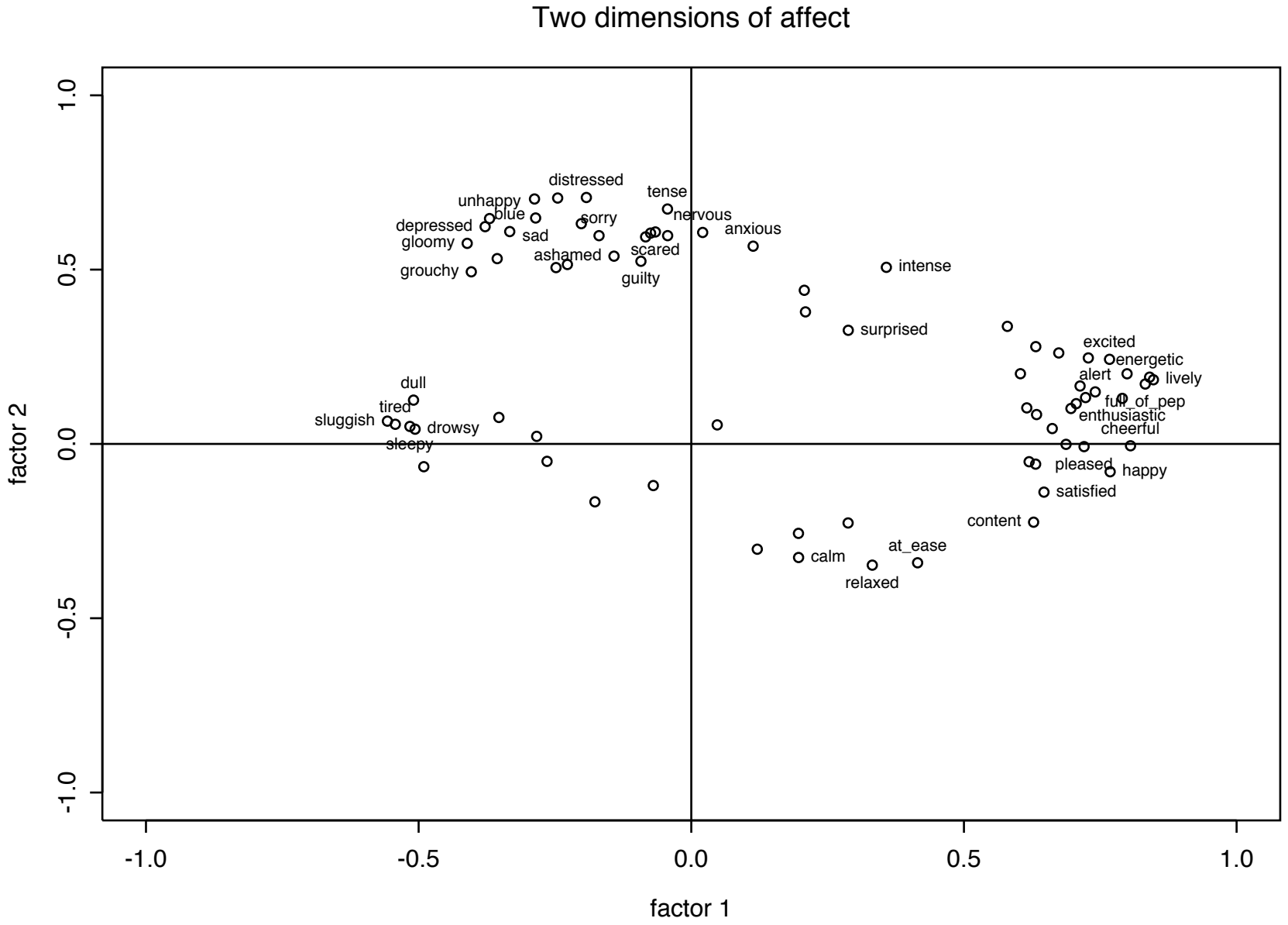
Why R?

- Graphics for data exploration and interpretation
- Data manipulation including statistics as data
- Statistical analysis
  - Standard univariate and multivariate generalizations of the linear model
  - Multivariate-structural extensions

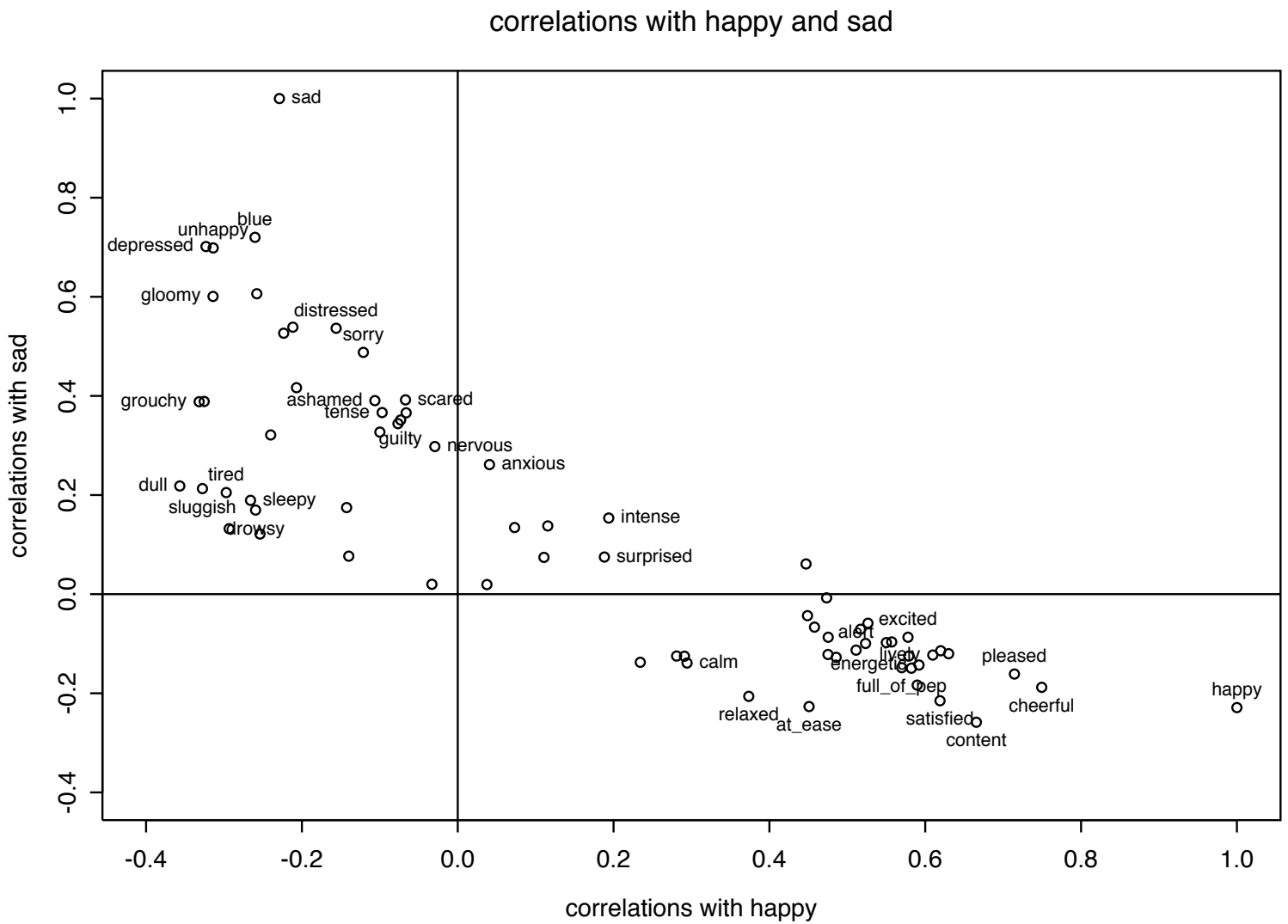
Why R? Graphics

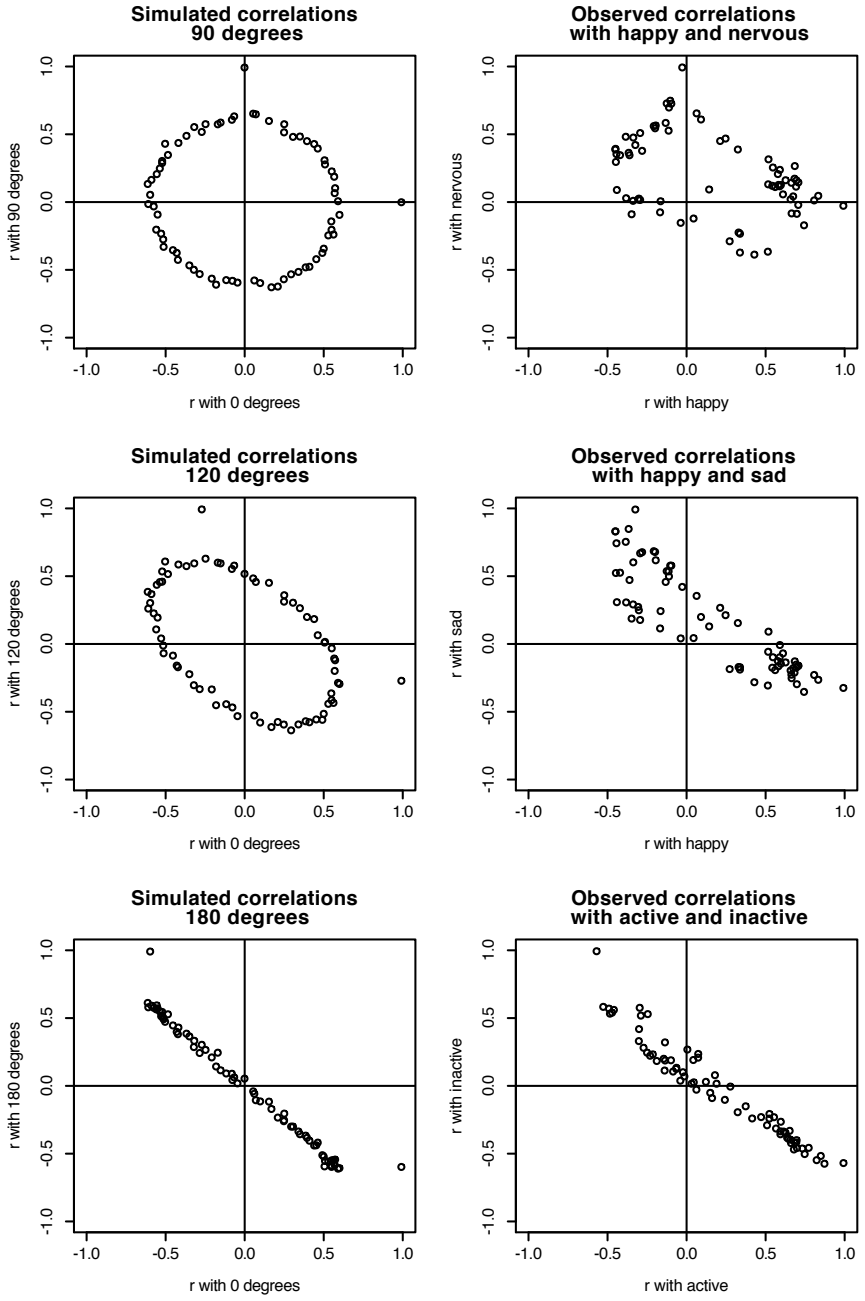
- Sample graphics taken from
  - http:personality-project.org/r/
    - showing what can be done by an amateur
  - http://addictedtor.free.fr/graphiques/
    - showing some most impressive graphs

Standard Plots of factor loadings



Data points can be dynamically identified

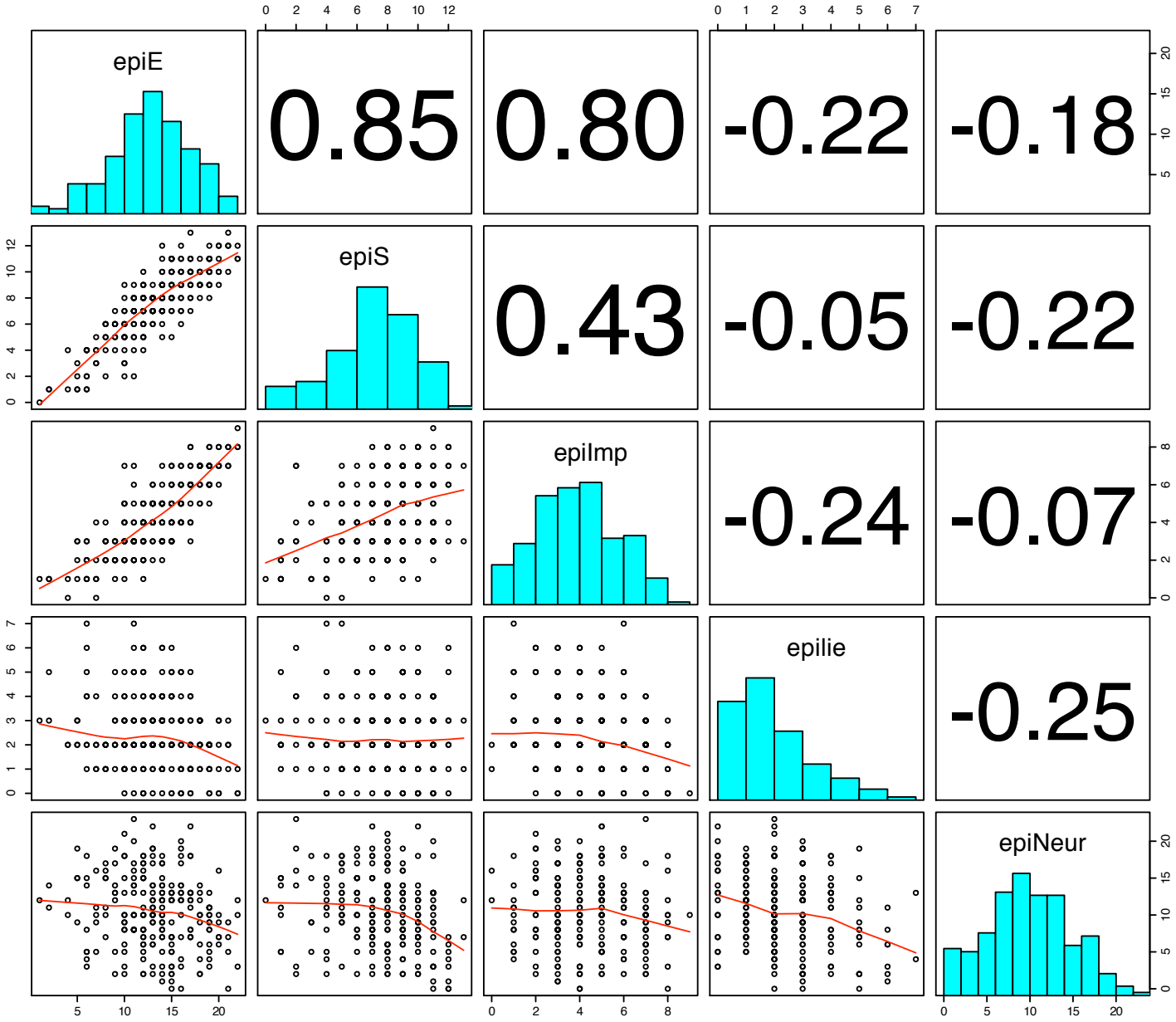




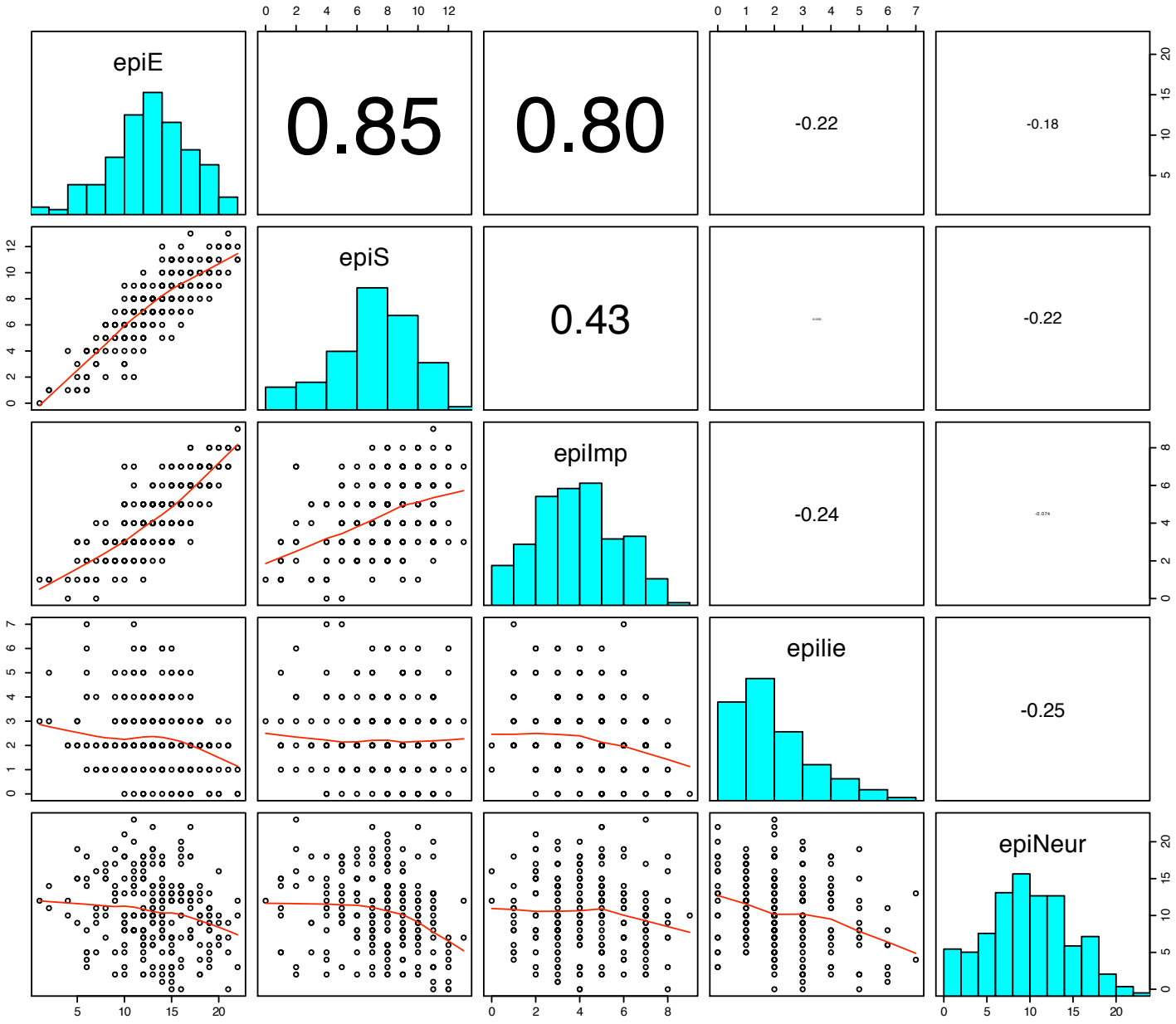
Multi-panel graphs can be labeled separately and organized vertically or horizontally

Simulated data can be generated to fit normal, rectangular, binomial, poissen, exponential, etc. distributions

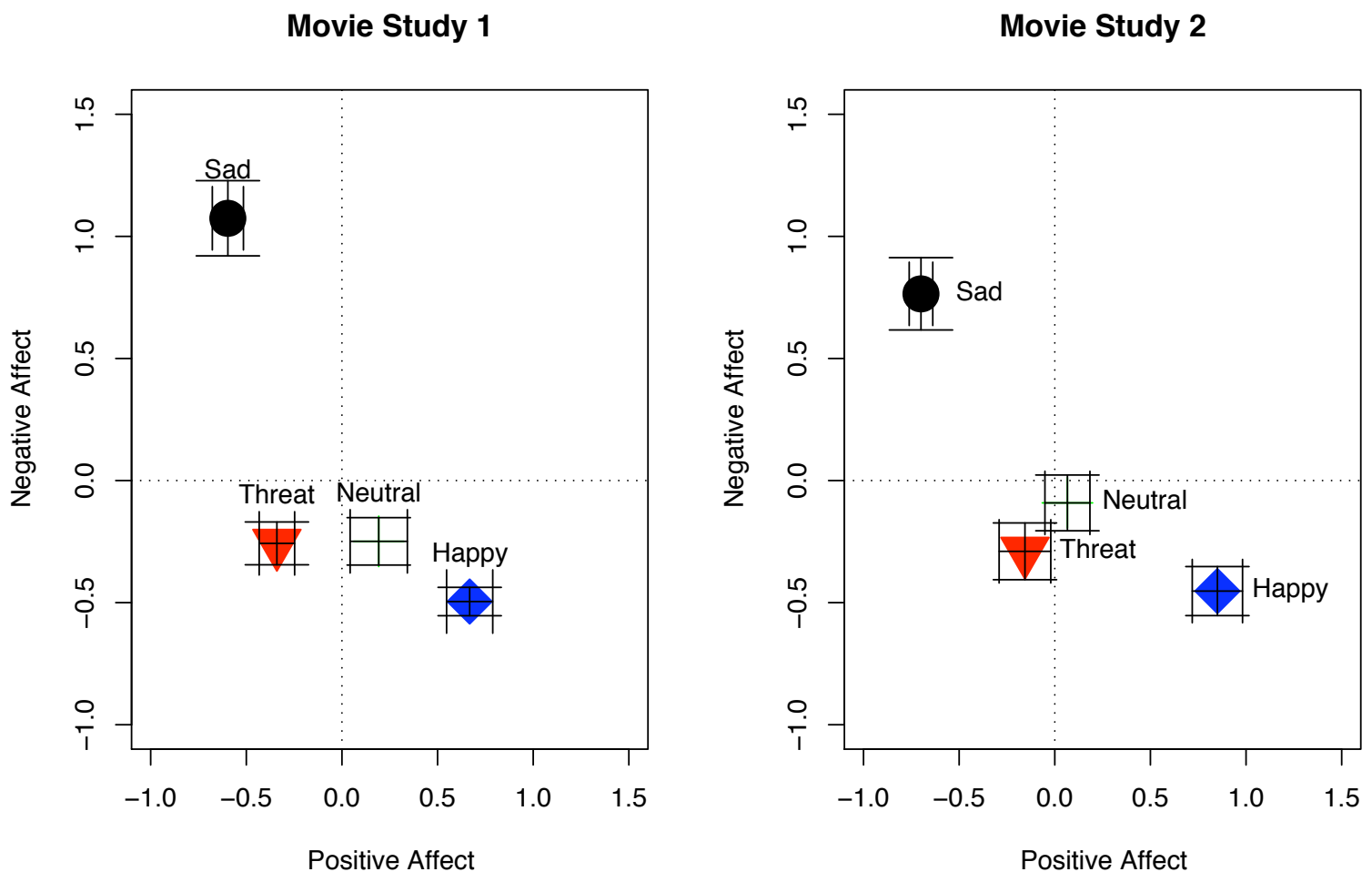
Scatter Plot Matrices can show smoothed fits

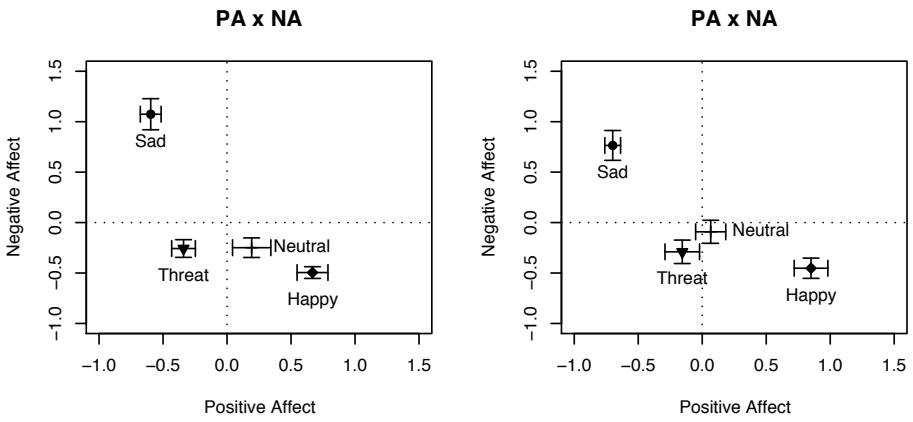


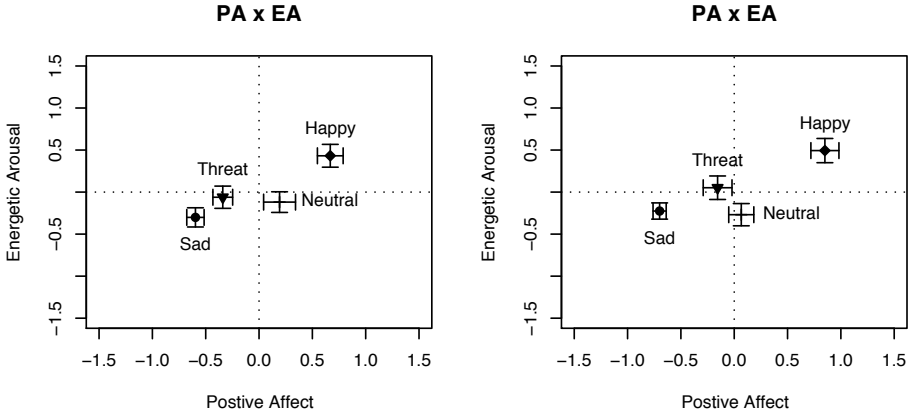
Can scale font size of correlations by absolute size of r

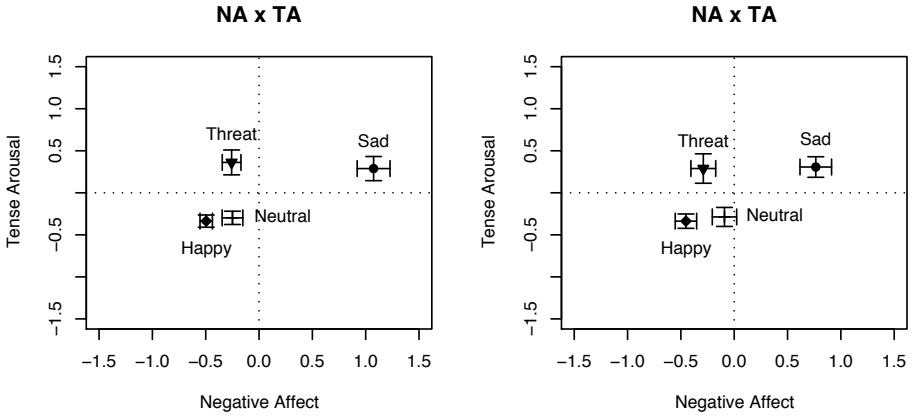


Error bars on two dimensions





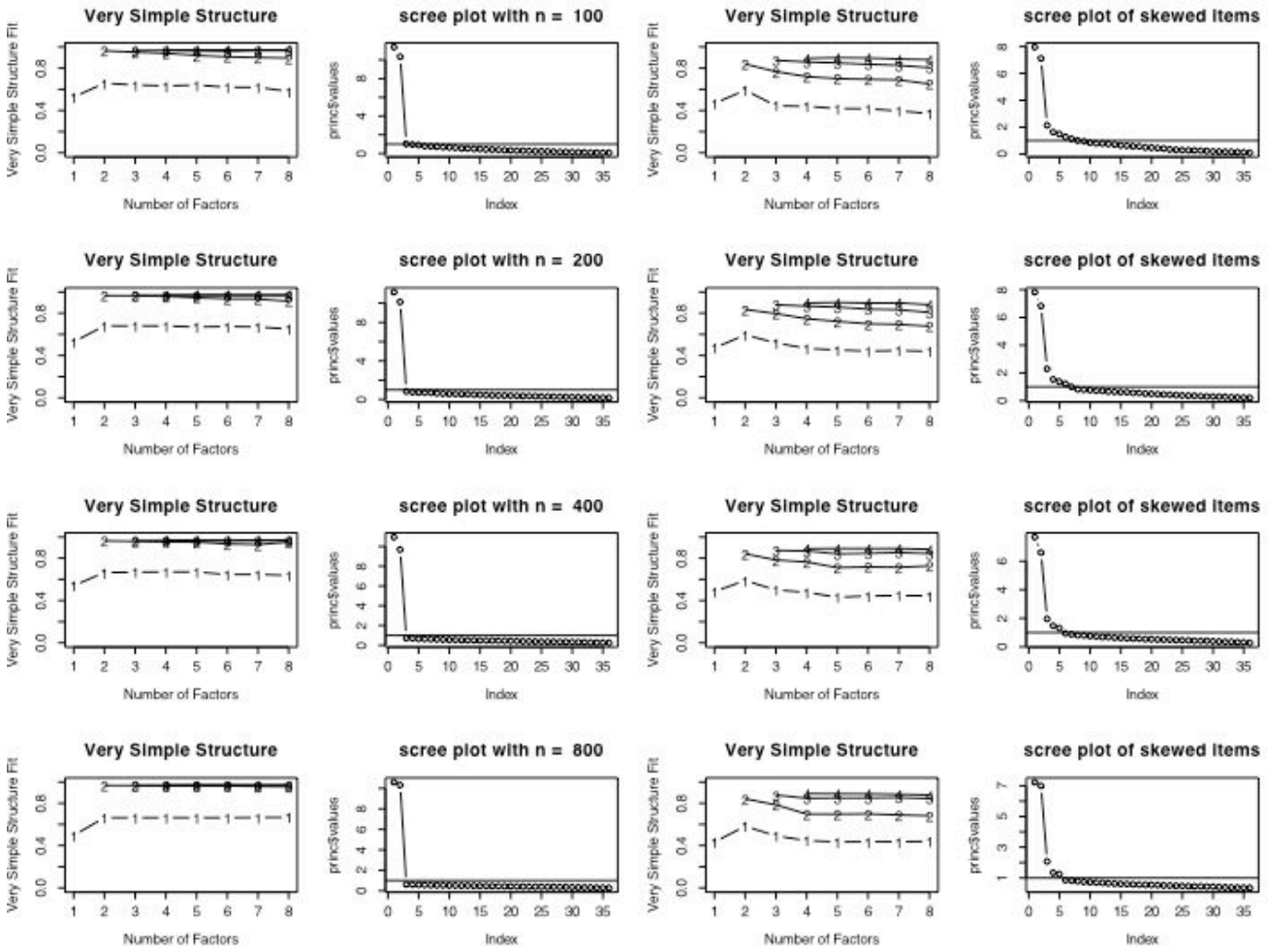




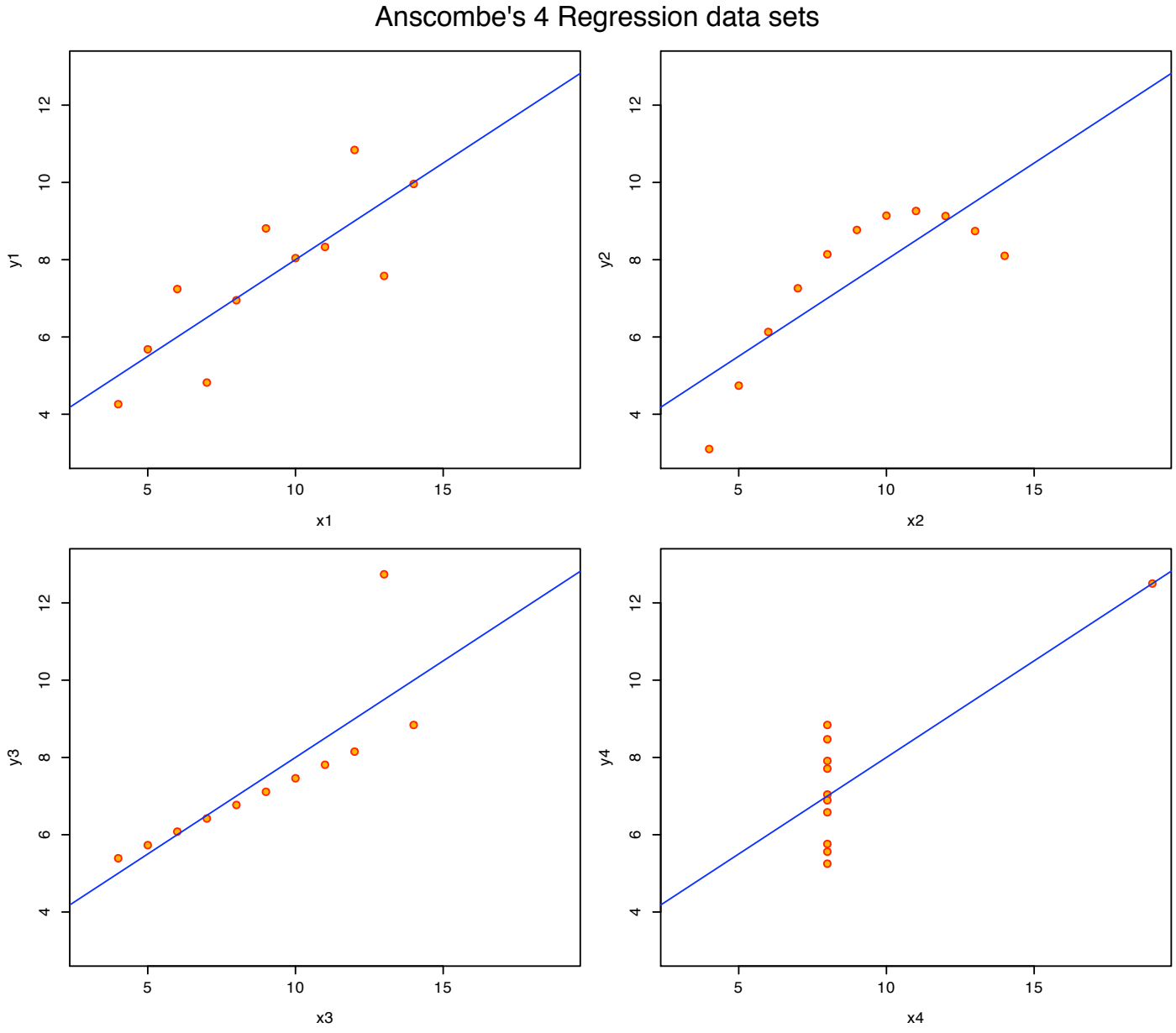
Presentation graphics scale to fit page and produce output for pdf presentations

Window or page size is controllable

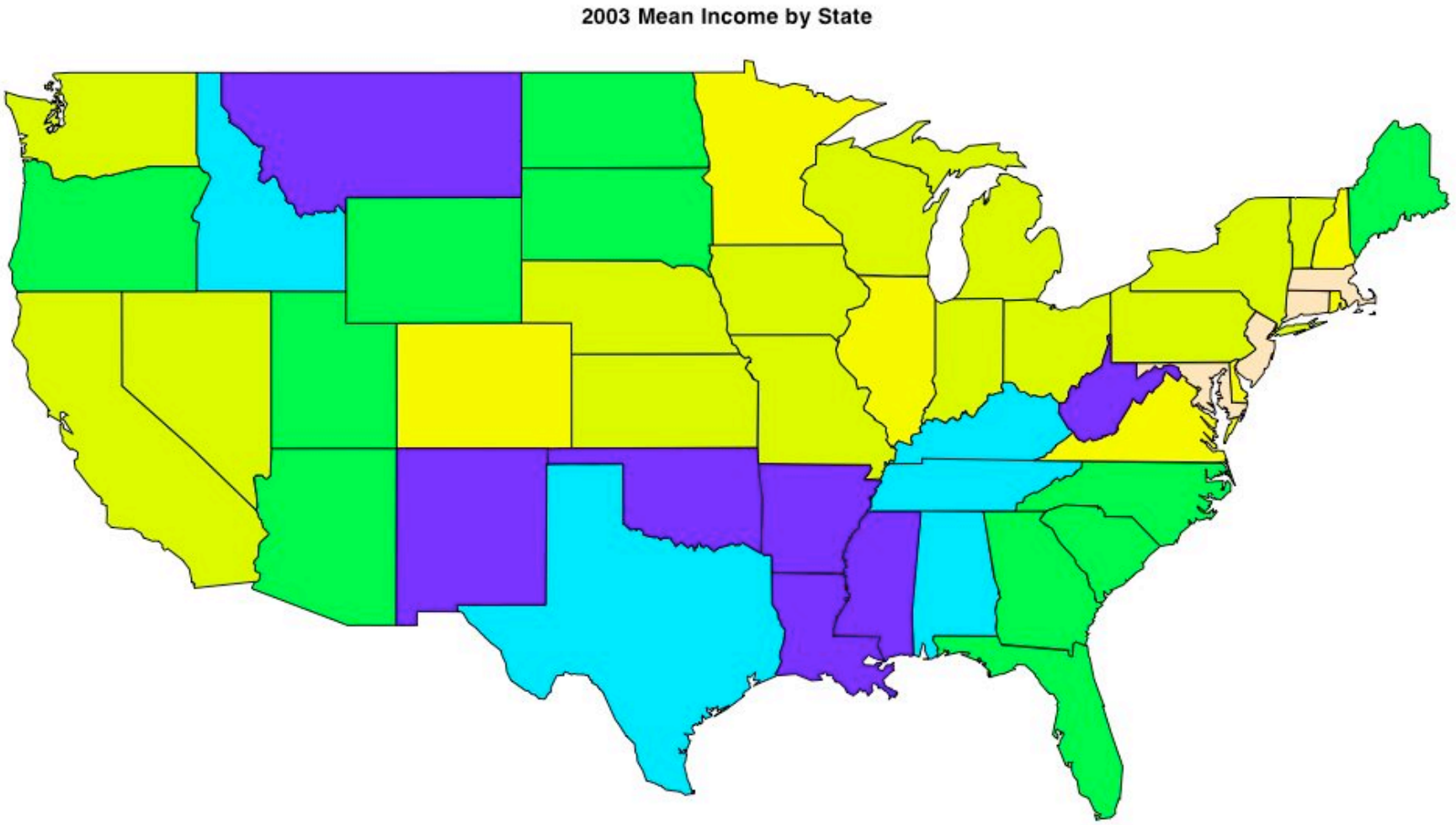
Graphic output can be taken from multiple programs (e.g. Very Simple Structure, factor analysis



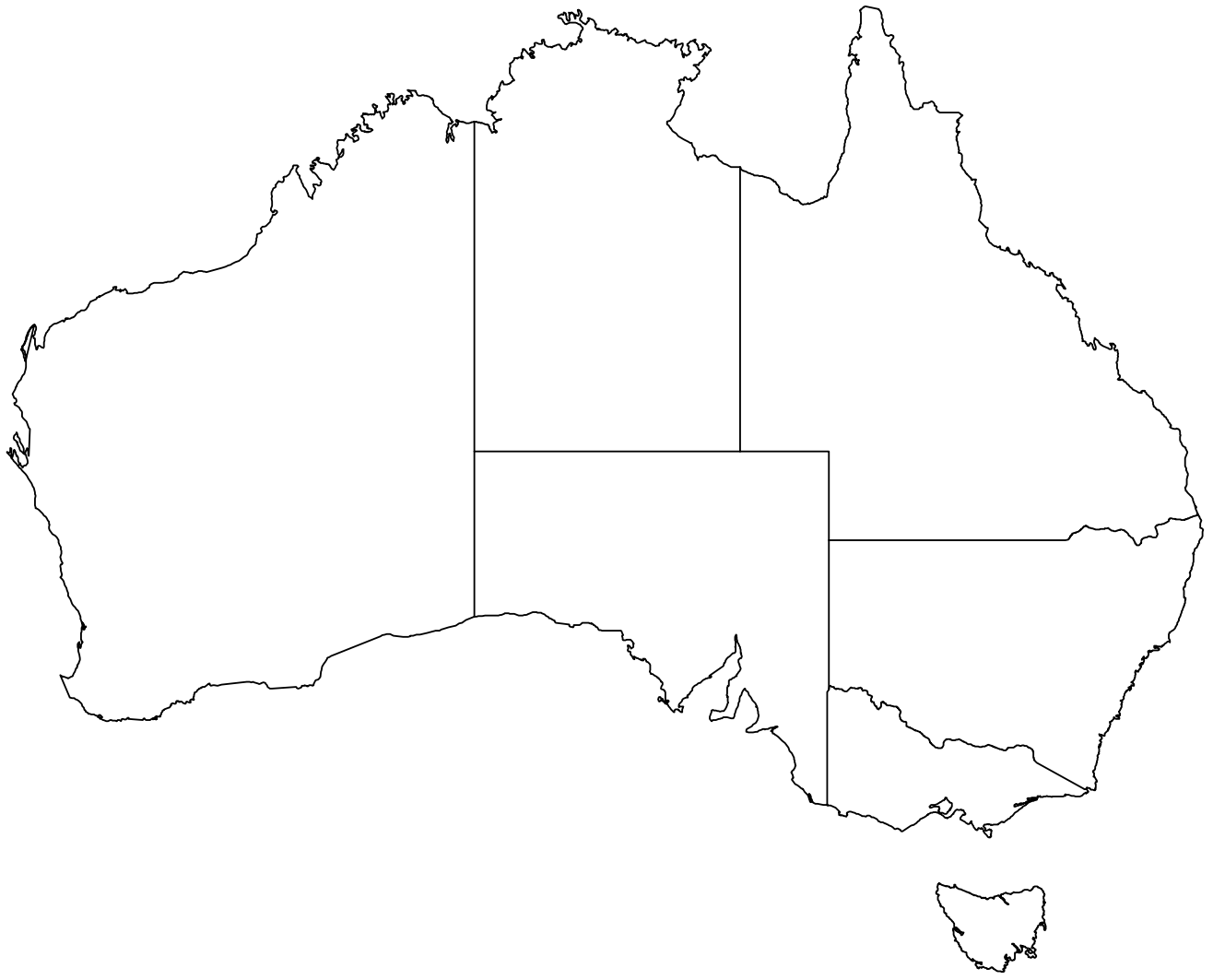
Built-in data sets provide useful demonstrations of stats



Mapping data (GIS) may be combined with descriptive data

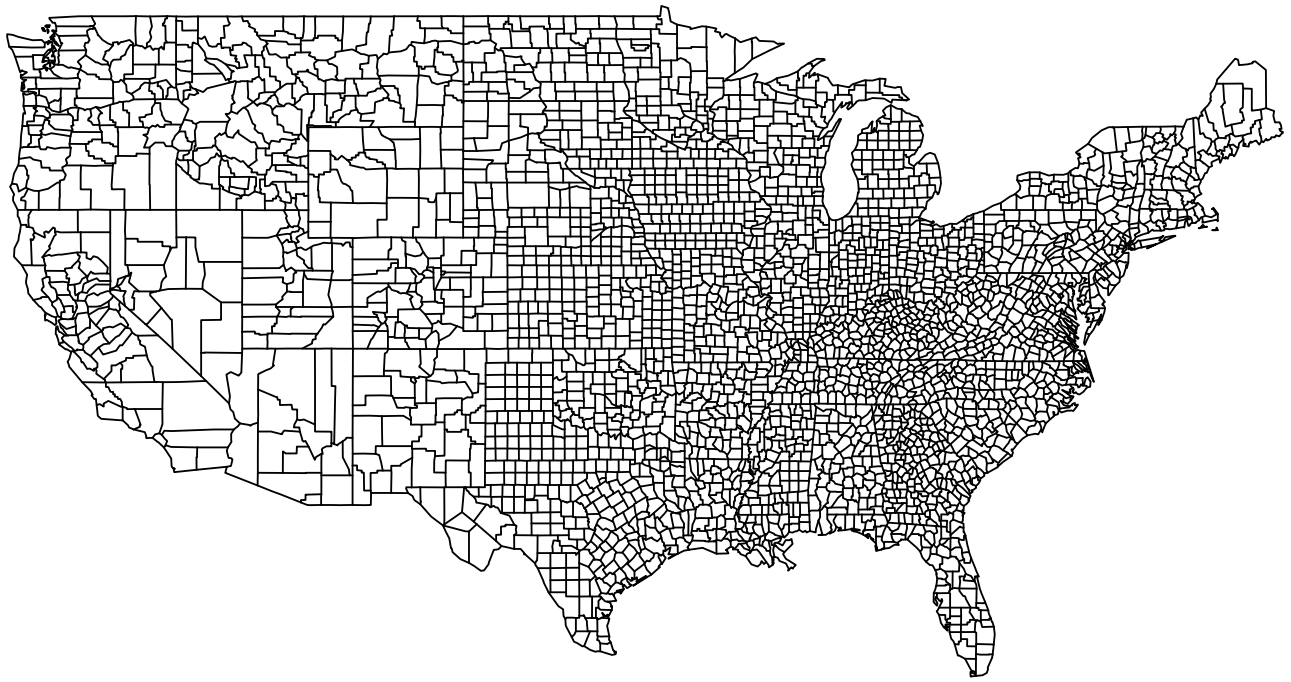


Many GIS map files are available for download from ESRI



GIS maps detail regional boundaries

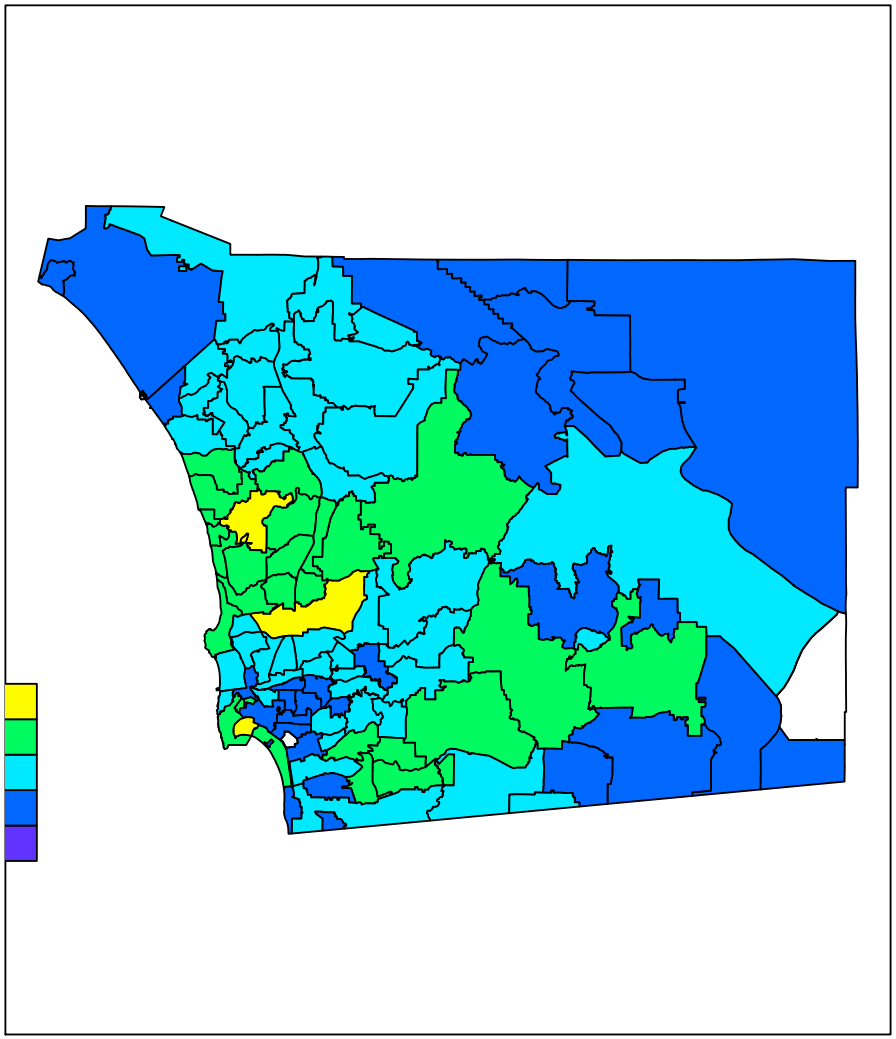
US counties



Counties of California

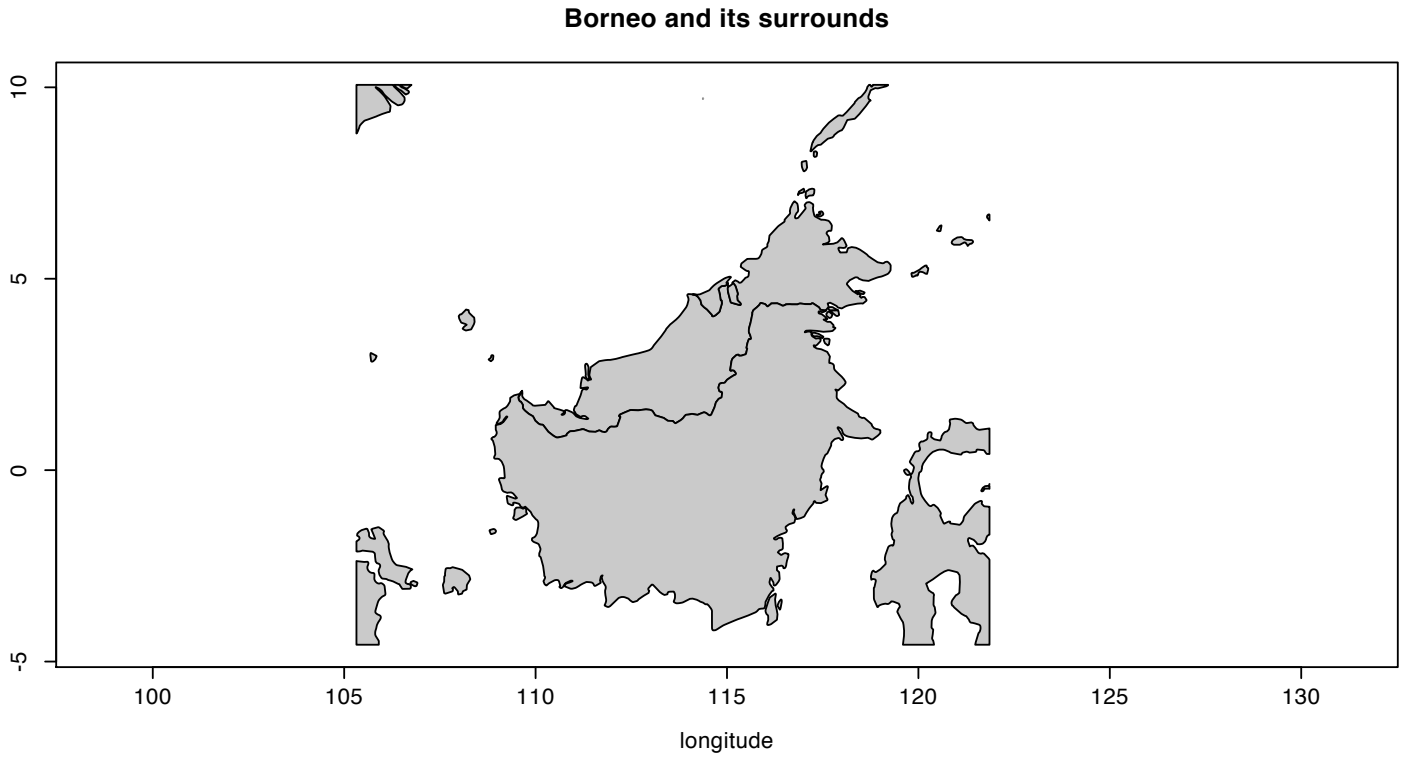
Combine income data from 2000 census with zipcode map of San Diego County

Median Household Income for 2000

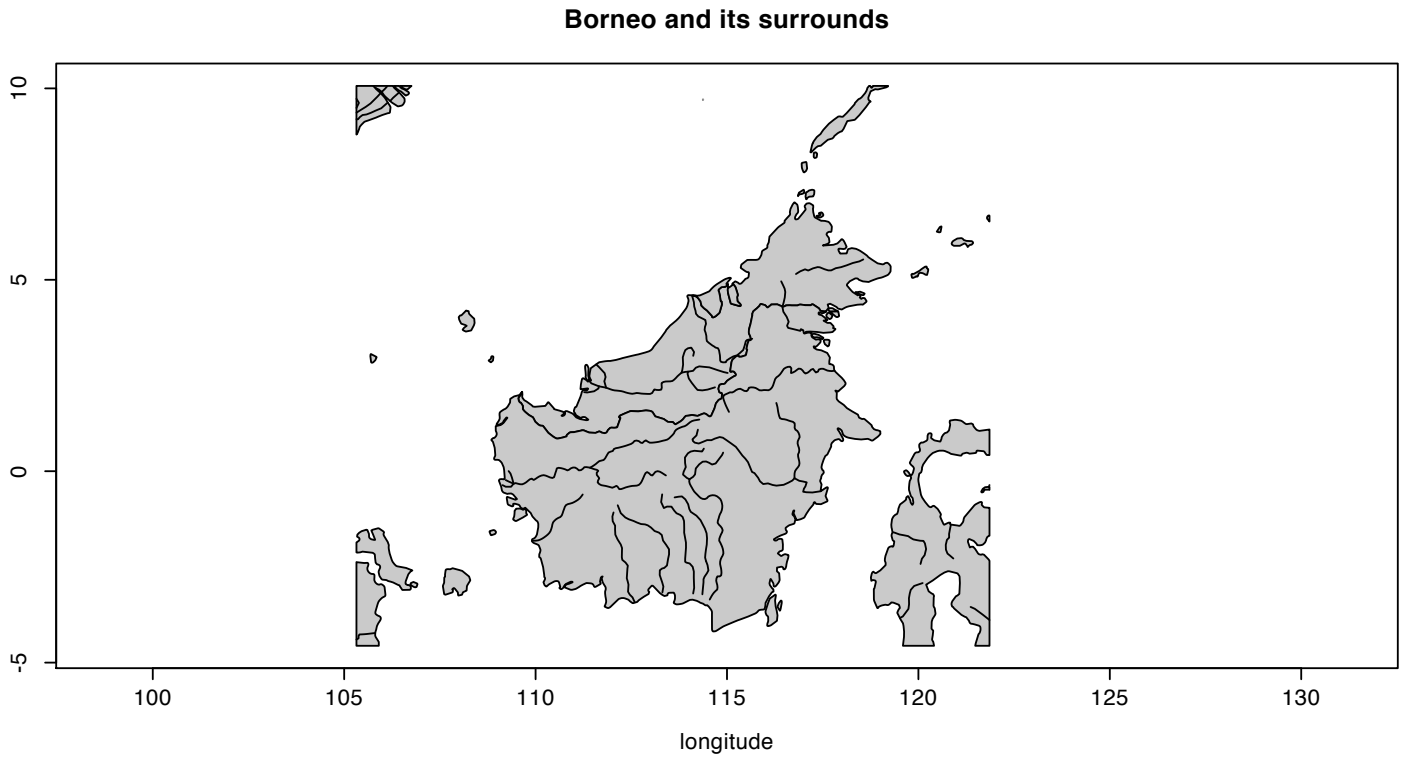


2000 median household income for zip codes

GIS files of Borneo can show country boundaries (e.g., Malaysia, Brunei, Indonesia)



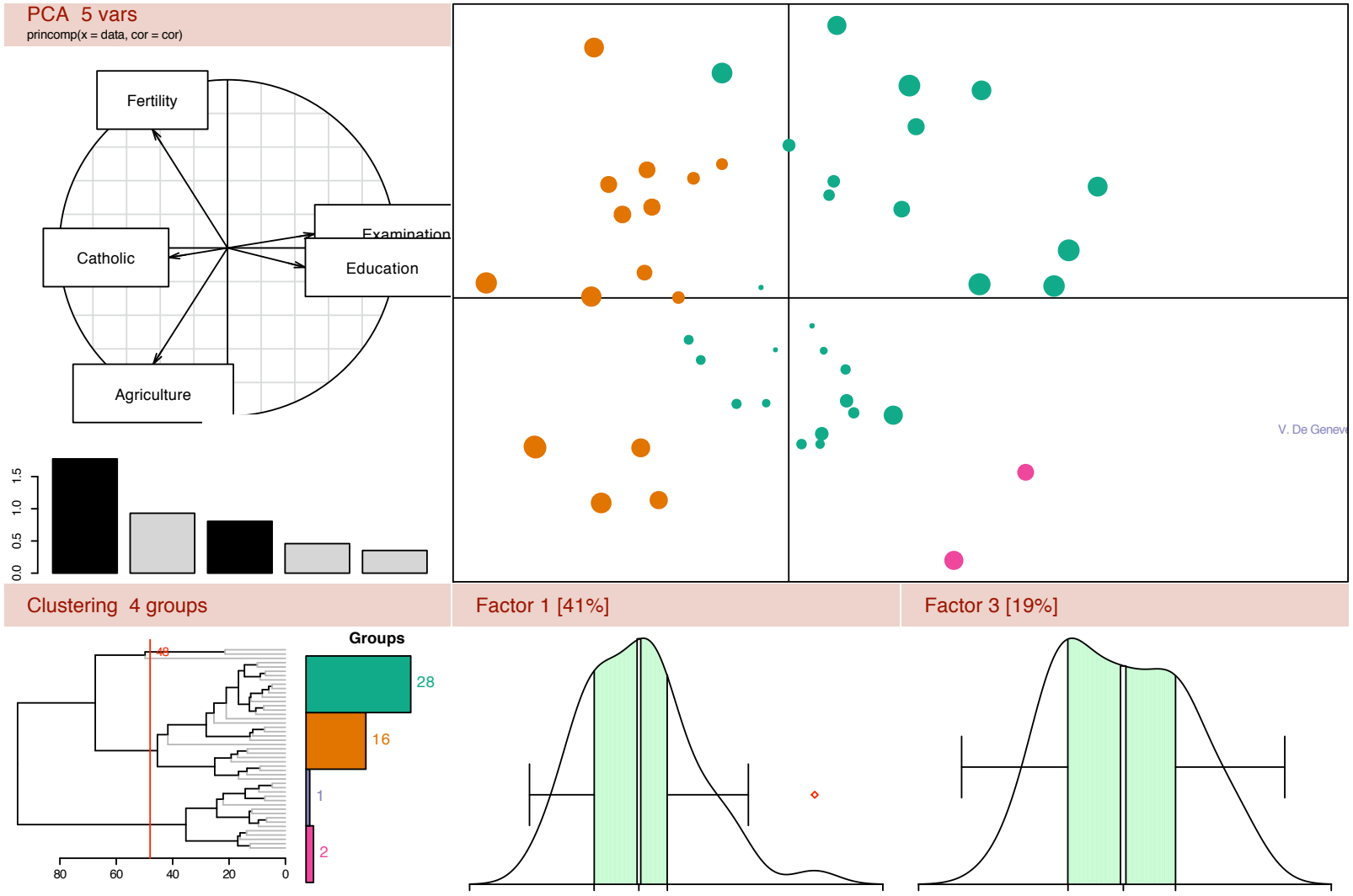
Public access GIS files include rivers and roads



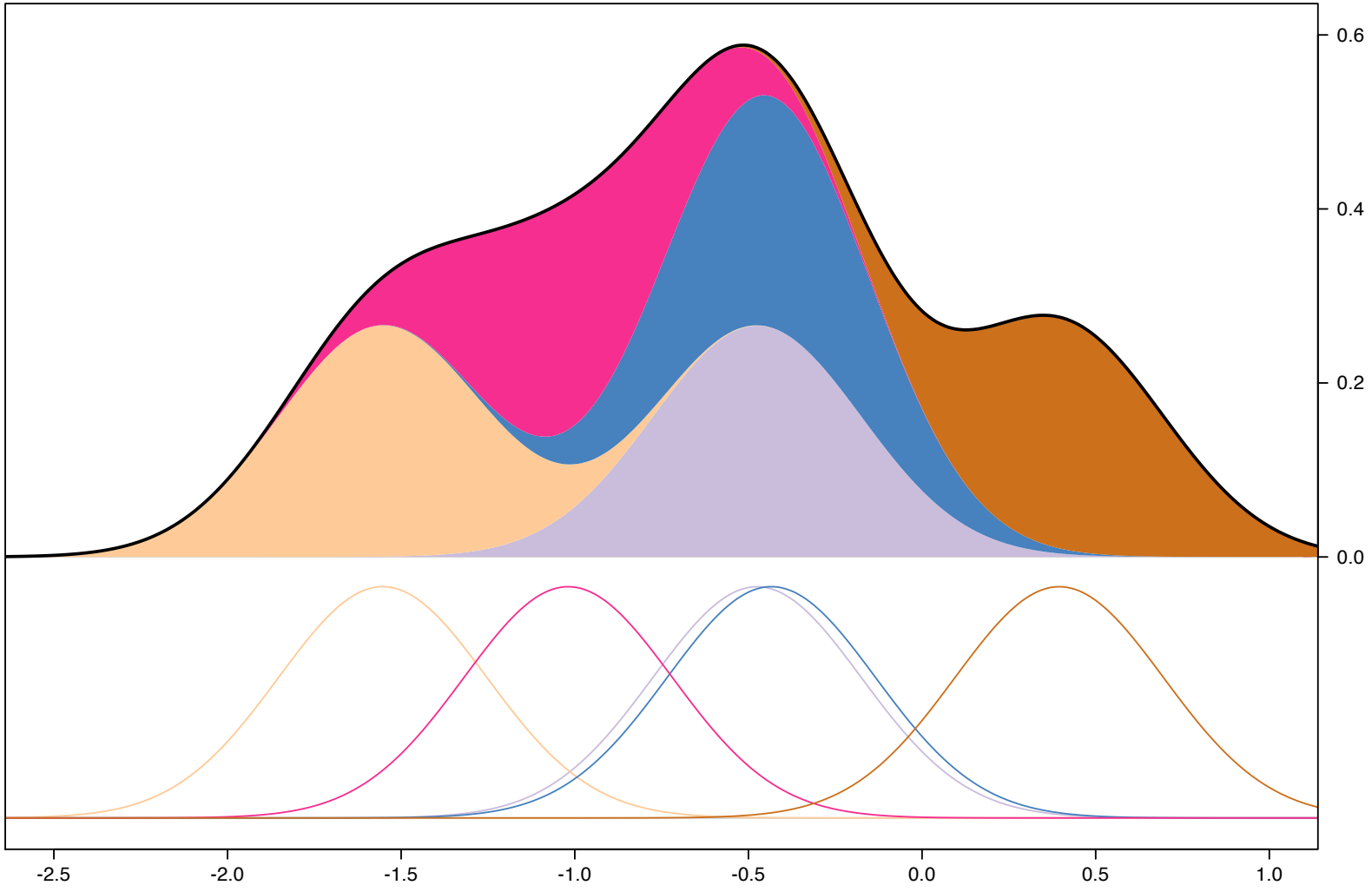
Even more graphics

- Taken from a collection of R demonstrations and graphics
  - http://addictedtor.free.fr/graphiques/

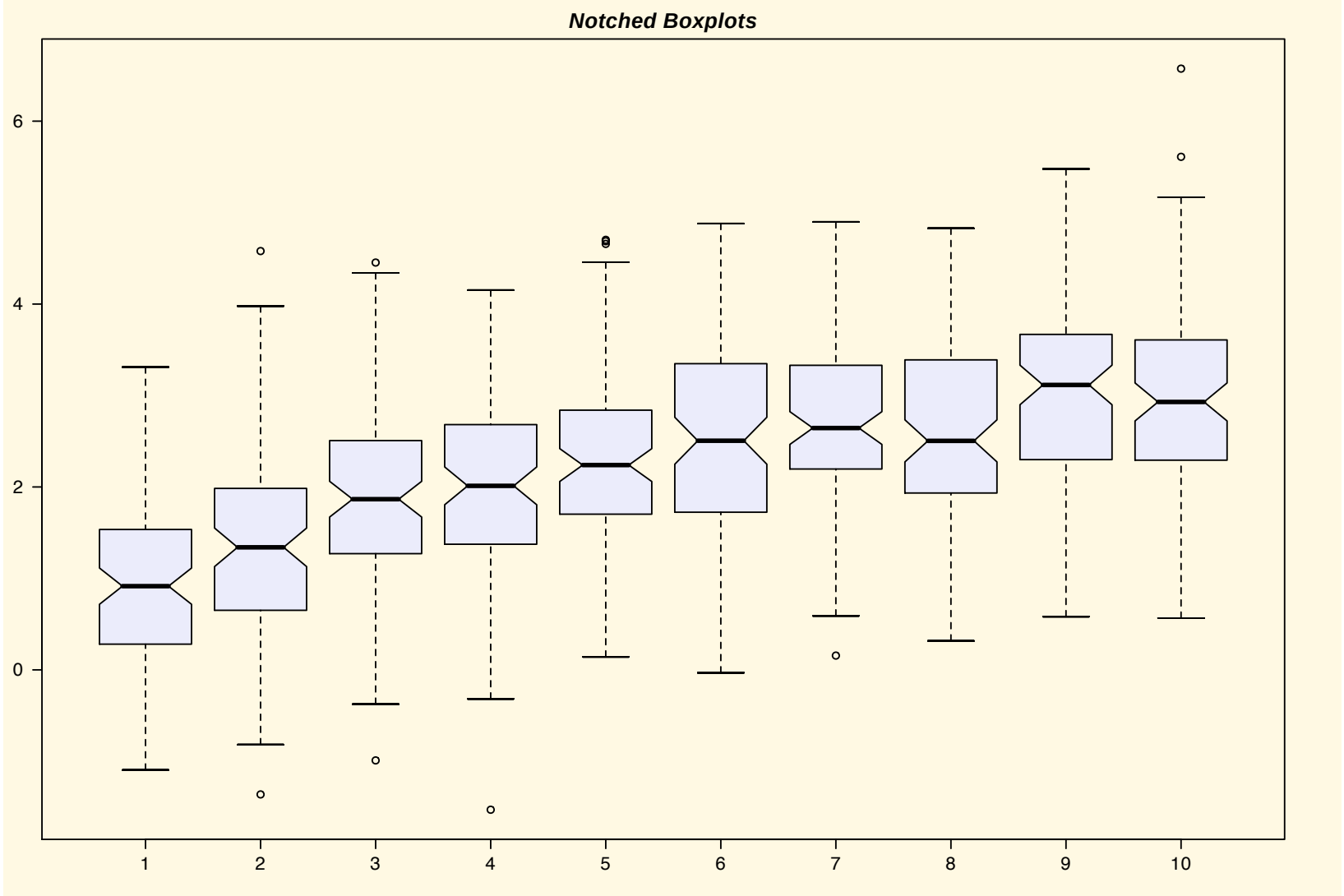
Principal components and clustering of sources of variance in USA arrest data



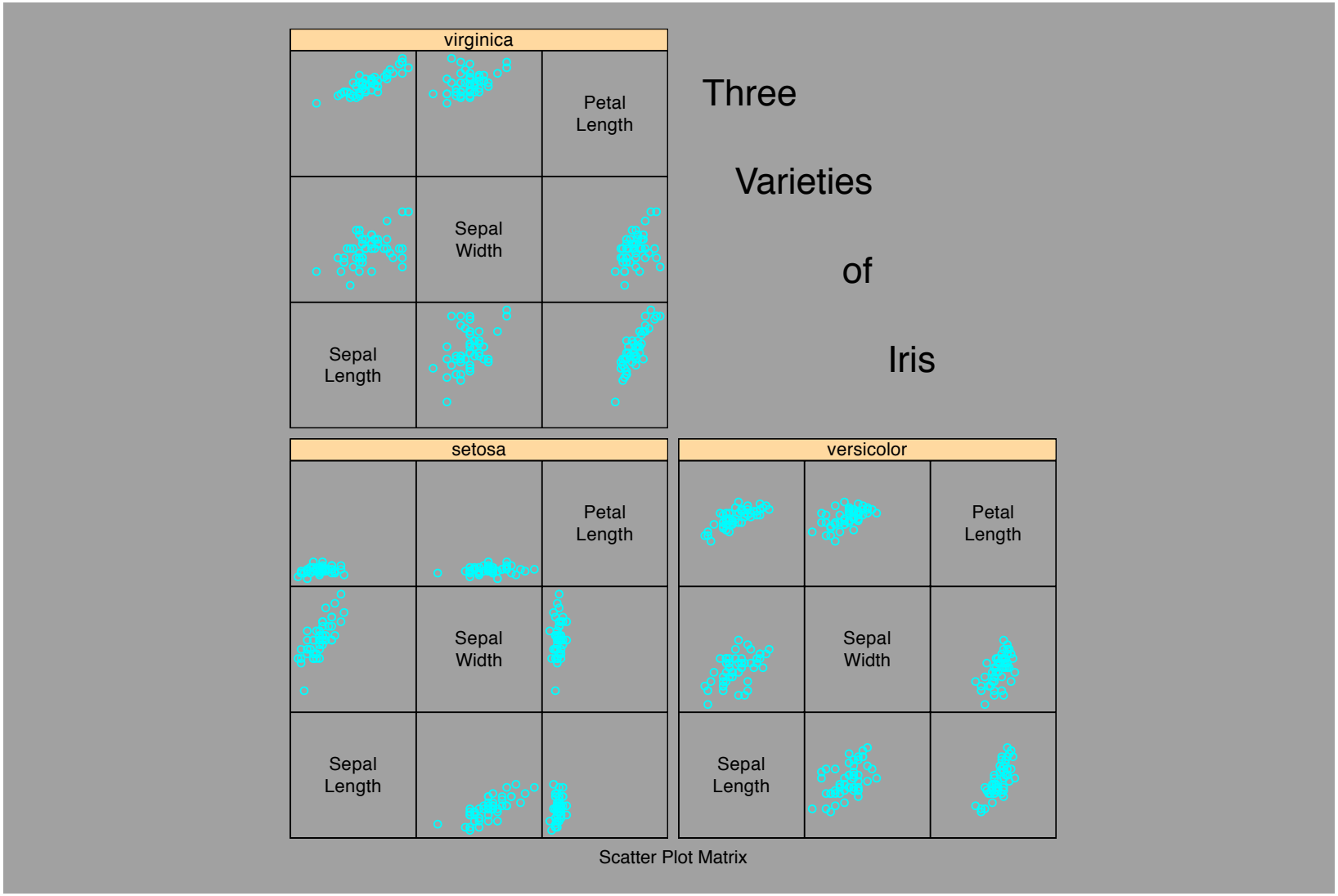
Mixture models



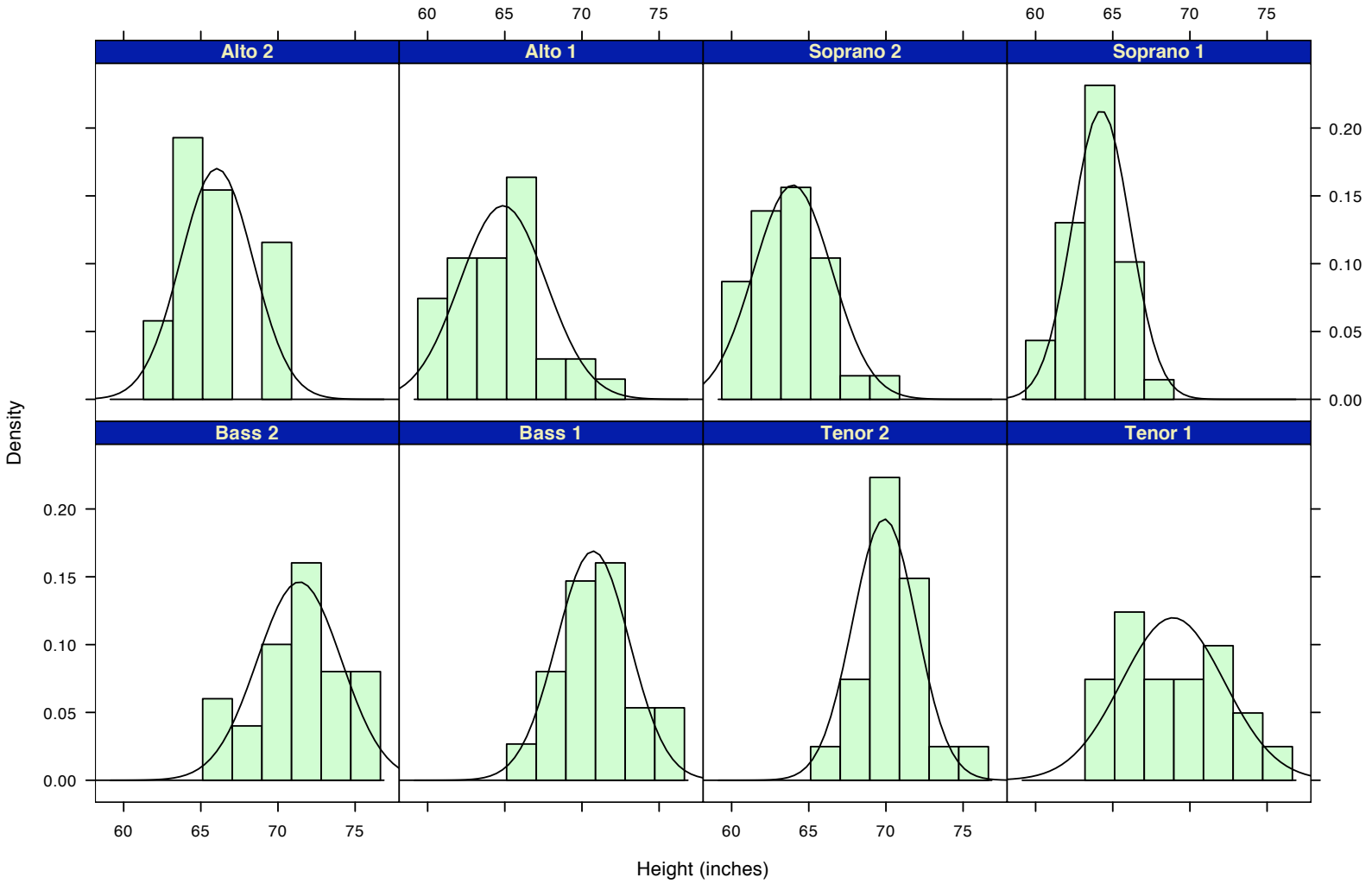
Notched Boxplots show confidence regions



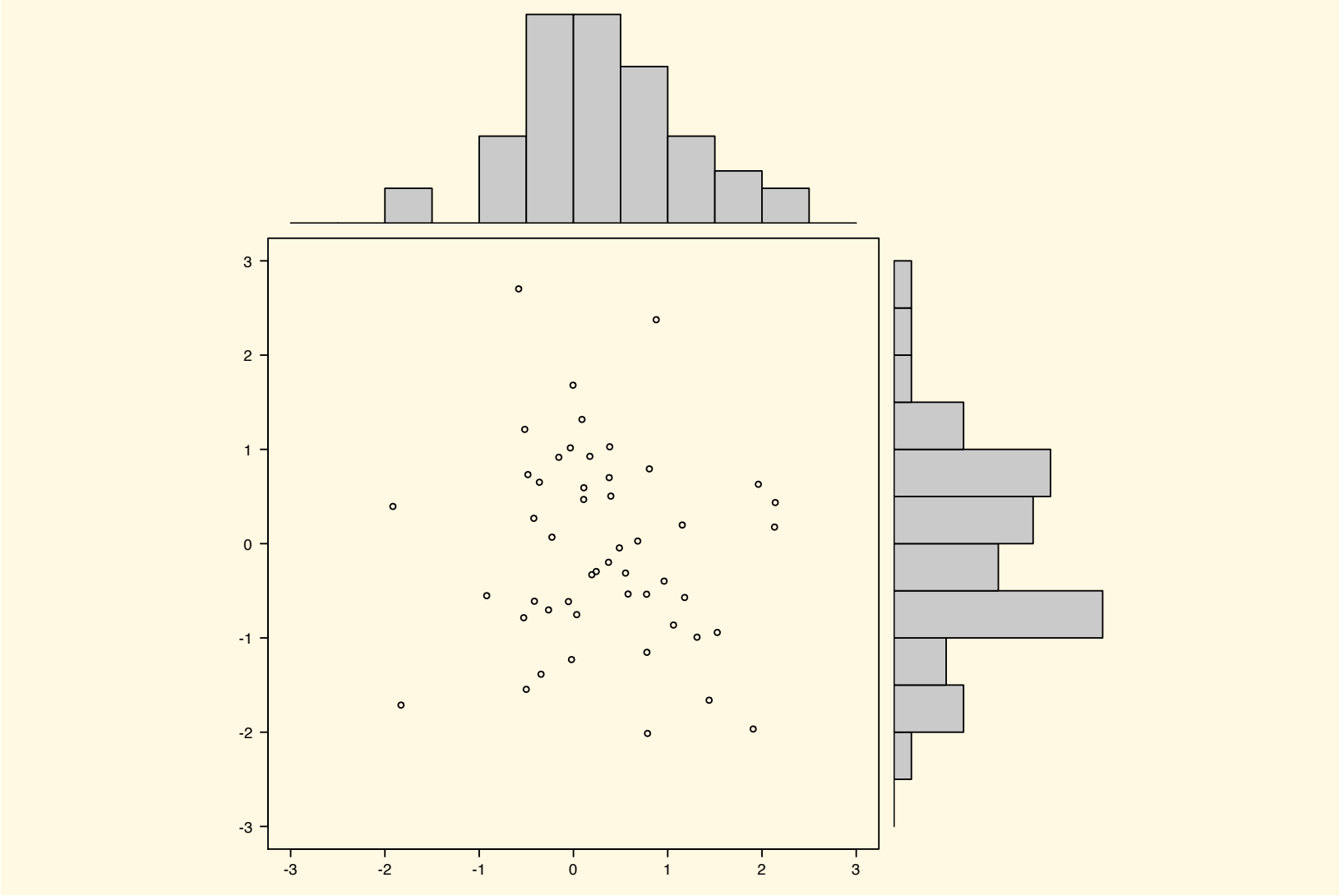
Multipanel graphs



Histograms and fitted distributions



Combine scatter plot with histograms



Why R?

- Graphics for data exploration and interpretation
- Data manipulation including statistics as data
- Statistical analysis
  - Standard univariate and multivariate generalizations of the linear model
  - Multivariate-structural extensions

Data Manipulation
Data Entry

- from console
- from clipboard (copied from other programs)
- from file (text files, csv, SPSS, Excel, MySQL)
- from the web

Data Manipulation: Data Structures

- Data types: integer, real, logical, character, string
- Vectors of any data type
- Matrices of any data type
- Data Frames (similar to matrix of mixed type)
- Lists of any mixture of types
- All operations are functions and the returned values may be used in any data structure (e.g., as an element of a data frame or of a list)

Data Manipulation

- standard arithmetic and logical operations
- matrix operations including transpose, inner product, outer product, diagonal, trace, invert
- searching, sorting, merging
- data cleaning by logical commands

Why R?

- Graphics for data exploration and interpretation
- Data manipulation including statistics as data
- Statistical analysis
  - Standard univariate and multivariate generalizations of the linear model
  - Multivariate-structural extensions

Standard Statistical packages in R

- Descriptive and exploratory statistics
- The general linear model and its special cases
  - t-test, ANOVA, MANOVA, regression, logistic regression, cox models, etc.
  - multilevel models (mixed models, hierarchical models)
  - time series, econometrics
  - circular statistics, environmental-geographical statistics

Psychometric packages

- Structural Equation Modeling
- Rasch Modeling of one parameter IRT
- Factor and Principal Component Analysis
  - Rotations (Orthogonal: Varimax, Oblique: quartimax, quartimin, Promax, etc.)
  - singular value decomposition
  - eigen vector - eigen value decomposition
- Multidimensional scaling

R is extensible

- Functions can be defined easily and then stored for later use.
- Packages of functions are written by "all of us" to solve particular problems
  - e.g. sem, rasch, rotation, lattice graphics
- Packages can be developed and tailored for a specific lab or problem area, or can be shared through the CRAN repository of (currently) > 400 packages

Programming in R-- personal examples

- Very Simple Structure
  - 6 weeks in Fortran for a mainframe
  - 3 weeks in Pascal for a Mac
  - 2 days in R for Mac/Unix/Windows
- Schmid Leiman decomposition and estimation of coefficient omega
  - 1 day

Popular misconceptions

- R is hard to learn
- R is not user friendly
- I can't figure R out
- R is free -- it can't be very good

Popular Misconceptions

- R is hard to learn
  - With a brief web based tutorial, 2nd and 3rd year undergraduates in psychological methods and personality research courses were using R for descriptive and inferential statistics and producing publication quality graphics
  - R is easy to learn, hard to master
    - R-help newsgroup is very supportive
    - Multiple web based and pdf tutorials

Popular Misconceptions

- R is not user friendly
  - Does not have a GUI, but those are being developed
  - Help and examples are embedded within program,
  - ? function produces help with examples for any function
  - R-help newsgroup is very supportive

Popular Misconceptions

- R is free, it can not be any good
  - Developed by some of the best statisticians around
  - vetted by all users, bugs when they exist, are rapidly fixed
  - R community provides excellent and timely statistical and programming help (if somewhat acerbic to those who have not done their homework)

R: an international collaboratory

- Most appropriate for an International Society to be aware of the power of international collaboration to produce cutting edge software

R: statistics for all of us

R: an international statistical collaboratory

Prepared for part of the symposium on
Multivariate Statistical Methods in Individual Differences Research
International Society for the Study of Individual Differences
Biennial meeting, Adelaide, July , 2005

William Revelle, Northwestern University
personality-project.org/r/
personality-project.org/r.short.html