Interactive Graphics
requirements, applications and implementation

**Simon URBANEK**
AT&T Labs - Research
Statistics Research, Shannon Lab
Florham Park, NJ, USA
Overview

- Types of statistical graphics
- Exploratory data analysis with interactive graphics
- Interactive graphics - the requirements
- Interaction in model building and analysis
- Synergy of two worlds: R and interactive graphics
- Creating interactive plots - behind the scenes
- Concluding remarks
Statistical graphics - example

Guinea Pigs' Tooth Growth

- Ascorbic acid
- Orange juice

Vitamin C dose mg vs. tooth length
Types of statistical graphics

- **Presentation graphics**
  - conveys findings
  - descriptive, self-contained (labels, legends...)
  - static, often printed
  - large audience

- **Exploratory graphics**
  - facilitates analysis
  - interactive, information on-demand
  - usually the audience is the analyst, but other persons can be involved in the analysis as well
Interactive Graphics
requirements, applications and implementation
Simon Urbanek, AT&T Labs - Research, Statistics

Daily use?

- Presentation graphics
  - R/S-plus, MATLAB, Excel, ...

- Exploratory graphics
  - ?
Exploratory Data Analysis (EDA)

EDA is an approach for data analysis that employs a variety of (mostly graphical) techniques to:

- maximize insight into a dataset
- uncover underlying structure
- extract important variables
- detect outliers and anomalies
- test underlying assumptions
- develop parsimonious model

*Coined by John W. Tukey (1977)*
Interactive statistical graphics

Key elements
- Linked highlighting
- Selection
- Queries
- Manipulation of plot parameters
- Pan/zoom

Optional elements
- Warnings
- Censoring
Highlighting

Is highlighting itself the same type of plot?
Extension: color brushing
Linked highlighting
Other possible properties to link

- **Order of categories**
  - barcharts, spineplots, mosaic plots, boxplots by group ...

- **Axis scales**
  - scatterplots, histograms, ...

- **Data points**
  - changes in data are reflected in all plots

- **Model properties**
  - nodes in a tree with histograms, scatterplots, ...
Linking types

Cold
- no direct linking, only creation of a new plot based on the same data constitutes a link

Warm
- a change is signaled, but the user must approve notification to other plots manually

Hot
- a change is immediately signaled to all linked plots and those are updated immediately without additional user interaction
Selection

- **Selection method**
  - rectangle, slice, lasso
  - brush
  - pointer

- **Logical combination of selections**
  - modifier (ad-hoc change of the operator)
  - selection mode (pre-selection of the operator)
  - selection sequence (dynamic modification)
Selection

**Selection modes**
- replace
- intersection (logical “and”)
- union (logical “or”)
- toggle (logical “exclusive or”)
- exclusion (logical “not and”)

**Selection sequences**
- arbitrary combination of selection modes
- implicit left-associativity
Queries

- Query information that is not necessarily visible in the plot
- Multiple query levels

(w/courtesy of M. Theus)
Manipulation of plot parameters

Example: modification of the bin width in histograms
Panning/zooming
Additional features

**Warnings**
- Visual notification when a property is not apparent or possibly misleading ("What you see is not what you think you see!")
- Examples: sub-pixel size or highlighting, selected case outside visible area (when not expected)

**Censoring**
- Non-proportional representation of too small or too large objects by enforcing minimal and maximal size
- Usually used in conjunction with zoom: "censored zooming"
Interactive ≠ dynamic!

Interactive graphics
- user performs actions that interacts with the data, plot or selection, resulting in updates of the plot
- supports at least selection and linked highlighting

Dynamic graphics
- the plot is updated periodically (animations, live data streams, ...)
- interactions change only the data view (e.g. 3d-rotation)
Interactive graphics and statistical models

Interactive graphics are not limited to pure data displays. Inclusion of statistical information:

- analytic aides (density estimators, conditional plots, ...)
- combination of data and model display (linear regression, smoothing, splines, ...)
- additional model attributes (confidence bands, parameters, residuals, ...)

Interaction with the model

- change of model parameters
- selection of training data subset
Example: smoothing, splines, ...
Example: tree models
Some software for interactive graphics

**DataDesk (Paul Velleman et al.)**
- most standard plots, fast, warm and hot linking, wide variety of selection tools, allows data manipulation, point-and-click modelling interface (commercial, Win/Mac only, not recent)

**XGobi/GGobi (Debbey Swayne, Andreas Buja et al.)**
- focused on multidimensional, continuous data and projection techniques, allows data manipulation/transformation (limited selection interface, Win/Mac support expected to improve in the upcoming new version)

**Mondrian (Martin Theus)**
- selection sequences, variety of continuous and categorical plots, supports spatial data, cross-platform, uses R for models (no projections or 3d rotation)
Other software with interactive graphics

- SAS/JMP
- SAS/INSIGHT
- Spotfire
The future: synergy of two worlds

Software for interactive graphics
- interactive user interface, multiple views
- stand-alone, not extendable
- limited data manipulation and modelling support

Software for statistical computing
- language-based
- rich data manipulation and modelling tools
- flexible, extendable
- static graphics, limited user interaction

combine both worlds!
Interactive Graphics for R

models, computing data manipulation printed output

visualization data analysis

DEMO
iPlots - Goals

- **Full interactivity**
  - linked highlighting, zoom, queries, multiple views, ...

- **Flexibility**
  - programmable from R (axes, colors, selection ...)

- **Extensibility**
  - modifications in R possible (custom objects, lines, ...)

- **Integration**
  - flat learning curve, just one application: R
iPlots - Design
iPlots - Implementation

- interactive graphics code in Java
  - object oriented
  - platform independent
  - good graphics and interaction support by Java

- R/Java interface
  - rJava interface (now on CRAN)

- Availability
  - tested on Windows, Linux and Mac OS X
Future Research

- Allow definition of new plots directly in R
- Leverage of new language features for interfaces
- Implement additional back-ends
  - AWT, Swing (current), OpenGL (in progress)
- Support a full callback-interface (within JGR)

New iPlots with above features expected for useR! 2006
Summary

Interactive statistical graphics are important tools for EDA

Key requirements are linked highlighting, change of plot parameters and queries

Free stand-alone software for interactive graphics include GGobi and Mondrian

Interaction is not limited to standard statistical plots, but is also helpful in the context of model selecting, fitting and evaluation

Statistical computing environments can be complemented by interactive graphics

iPlots project aims to provide interactive graphics for R
Contact and Acknowledgements

Simon URBANEK
urbanek@research.att.com
AT&T Labs - Research
Statistics Research, Shannon Lab
Florham Park, NJ, USA
http://research.att.com/~urbanek

Joint work with
Martin Theus,
Alexander Gouberman,
Tobias Wichrey
RoSuDA
University of Augsburg
Germany
Software links

- Mondrian, Klimt, JGR, iPlots (free)
  - http://www.rosuda.org/software/
- GGobi (free)
  - http://www.ggobi.org/
- DataDesk (commercial)
  - http://www.datadesk.com/
- SAS JMP (commercial)
  - http://www.jmp.com/