Scagnostic in R (and GGobi)

Hadley Wickham
Outline

- Intro to Scagnostics
- R package and GGobi interface
- Speed
- Data expo data (how can we explore temporal/spatial relationships with scagnostics?)
Scagnostics

- **Scatterplot Diagnostics**
  (John & Paul Tukey 1985)

- Characterize 2d distributions according to density, skewness, shape, outliers, ...

- Wilkinson, Anand & Grossmann 2005: Graph-Theoretic Scagnostics
Graph-Theoretic Scagnostics

- Set of indices based on properties of convex hull, alpha hull & minimal spanning tree (MST)
Graph-theoretic Indices

- Outlying, Convex, Skinny, Stringy, Straight, Monotonic, Skewed, Clumpy, Striated
- Indices are bounded between 0 and 1
- ~ Uniformly distributed
R package

• Collaborative effort!
  • Lee Wilkinson and Anushka Anand: original Java code
  • Heike Hofmann: port to C++
  • Michael Lawrence and Duncan Temple Lang: bug fixing
  • Me: wrap C++ code in R functions
R interface

- `install.packages("scagnostics")`
- `library(scagnostics)`
- `scagnostics(x, y)`
- `scagnostics(data.frame)`
- `ggobi(scagnostics(data.frame))`
Example

- Prim 7 data
- Old physics data, but interesting high (7) D structure
- What do scagnostics reveal?
- Later, Heike will look at arbitrary projections
Speed

- $O(n \, d^2)$
- Points binned into 50x50 grid by default
- Problems with outlier detection?
- Quadratic increase in number of variables
- $d(d-1)/2$ plots in scatterplot matrix
Data expo

Biannual Data Competition sponsored by ASA and (in 2006) by NASA

Spatio-temporal longitudinal data:
Six years of monthly climate data on 24 x 24 regular grid over Central America
Data expo

- Large meteorological dataset:
  - 7 meteorological variables
  - 72 time points
  - 576 locations

- How can we explore interesting temporal/spatial relationships?
"Reshape" the data

- Can rearrange the data so that times occur in the columns
- Separate data frame for each variable: 72 x 576 data matrix
- Does it help find interesting temporal relationships?
Further work

• Make it faster?
• More/better scagnostics?
• Scagnostics for sections
• Are there times/locations which have different relationships between the original variables?