Scagnostics - Characterising and Selecting Scatterplots

Katrin Grimm

Department of Computer-oriented Statistics and Data Analysis
University of Augsburg

June 29, 2015
Introduction

- Scagnostics = **Scatterplot + Diagnostics**
- First mentioned by Paul and John Tukey in the middle of the 1980’s
- Scagnostics should help characterize scatterplots and find interesting structures
- Increasing amounts of data make new approaches necessary
- Details were never published by Tukey and Tukey
- Wilkinson et al revived the topic and implemented concrete measures (2005) in the R package scagnostics
Motivation

Why are scatterplots important?

- Two-dimensional structures are a good starting point for discovering interactions between variables in data frames
- Interactions between two variables are visualisable and can be easily interpreted in scatterplots

How can we overview all two-dimensional structures visually?

1. Look at all in $\binom{p}{2}$ scatterplots
2. Look at all in one scatterplot matrix
3. Look at all in different scatterplot matrices with $q$ variables per matrix
4. Use scagnostics
299 cases (Germans electoral districts)
70 variables
- vote shares of the parties in 2009 and 2005
- demographic and economic information about the constituencies (for example unemployment rate, population density, birth rate)

68 numeric variables $\Rightarrow$ 2278 possible two-dimensional plots
How can we overview all 2D structures visually? -
1. Look at all in \( \binom{p}{2} \) scatterplots
How can we overview all 2D structures visually? - 2. One scatterplot matrix with all variables
How can we overview all 2D structures visually? -
3. Scatterplot matrices with for example 10 variables in each matrix
How can we overview all 2D structures visually? -

4. Use scagnostics

1. Calculate measures for each scatterplot
2. Cluster the plots by these measures
3. Based on the clustering, define so called *Exemplars* and *Outliers*, where
   - *Exemplars* represent a group of similar plots
   - *Outliers* are strongly different from the other scatterplots
4. Look at the *Exemplars* and *Outliers*

⇒ Look at only a few plots, but know all the different structures in the data
**scagnostics in R - An overview of the approach**

1. Data are scaled to the unit interval and binned by hexagonal binning

2. After calculating a measure for outlier (*Outlying*), outliers are excluded from the data

3. Calculation of the
   - Minimum spanning tree
   - Convex hull
   - Alpha hull

4. Calculation of the further eight measures:
   - Skewed, Clumpy, Sparse, Striated, Convex, Skinny, Stringy, Monotonic
Outliers and Exemplars are specified after a nine dimensional clustering of the measures

Each cluster stands for a group of similar plots and the Exemplars are representatives for these

Outliers should be plots, which are strongly different to the other plots

The idea is that you only have to look at the Outliers and Exemplars to grasp all the different structures in the data
Outlier detection can be dependent on the binning
- Far-off points in the same bin → treated as outliers
- Far-off points points in different bins → whether treated as outliers or not depends on the distance of the points to each other

The measure Convex is strongly influenced by non-detected outliers (because of using the convex hull)
The *scagnostics* package delivers three *Exemplars* and 144 *Outliers*
The cluster membership of the plots is not recorded by the package

A ward clustering with three clusters based on the 2134 plots which are not classified as *Outliers* delivers clusters with sizes of 243 (two clusters), 1541 (linear dependency), 350 (“nothing special“)

- The *Exemplars* are in the different clusters
The *Exemplars* (1) - German election
Cluster structures caused by bimodal variables

Scagnostics - Characterising and Selecting Scatterplots

Katrin Grimm

Introduction

Scagnostics in R

Scagnostics in advance

Conclusions and further potential
The *Exemplars (2)* - German election
(Small) positive or negative correlation
The *Exemplars* (3) - German election “Nothing special“

Scagnostics - Characterising and Selecting Scatterplots

Katrin Grimm

Introduction

scagnostics in R

scagnostics in advance

Conclusions and further potential
What about more clusters? - German election
5 clusters
What about more clusters? - German election
7 clusters
The *Outliers* - German election
20 randomly chosen

Scagnostics - Characterising and Selecting Scatterplots

Katrin Grimm

Introduction
scagnostics in R
scagnostics in advance

Conclusions and further potential
## An overview of all measures - German election

<table>
<thead>
<tr>
<th>Measure</th>
<th>Example Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outlying</td>
<td><img src="image" alt="Outlying Example" /></td>
</tr>
<tr>
<td>Skewed</td>
<td><img src="image" alt="Skewed Example" /></td>
</tr>
<tr>
<td>Clumpy</td>
<td><img src="image" alt="Clumpy Example" /></td>
</tr>
<tr>
<td>Sparse</td>
<td><img src="image" alt="Sparse Example" /></td>
</tr>
<tr>
<td>Striated</td>
<td><img src="image" alt="Striated Example" /></td>
</tr>
<tr>
<td>Convex</td>
<td><img src="image" alt="Convex Example" /></td>
</tr>
<tr>
<td>Skinny</td>
<td><img src="image" alt="Skinny Example" /></td>
</tr>
<tr>
<td>Stringy</td>
<td><img src="image" alt="Stringy Example" /></td>
</tr>
<tr>
<td>Monotonic</td>
<td><img src="image" alt="Monotonic Example" /></td>
</tr>
</tbody>
</table>

### Details

- **Outlying**: Images showing extreme values.
- **Skewed**: Images showing data with asymmetry.
- **Clumpy**: Images showing data clustered together.
- **Sparse**: Images showing data spread out thinly.
- **Striated**: Images showing data with repeated patterns.
- **Convex**: Images showing data forming a convex shape.
- **Skinny**: Images showing data with narrow spread.
- **Stringy**: Images showing data forming elongated clusters.
- **Monotonic**: Images showing a linear relationship.

---

**Scagnostics** - Characterising and Selecting Scatterplots

Katrin Grimm

**Introduction**

**scagnostics in R**

**scagnostics in advance**

**Conclusions and further potential**
What did we learn about the election data?

Looking at the *Exemplars* showed:
- There are a few plots with two clusters (due to bimodal variables)
- There are associations between some of the variables

Looking at some *Outliers* showed:
- There are unusual structures caused by discrete and skew variables
- Additional clusters

Looking at different clusterings and also at smallest and highest values in each measure provides additional help to overview the data structure.
A possible extension - Univariate filtering

- Some of the observed unusual structures are caused by univariate anomalies
- Univariate filtering can bring some advantages
- Excluding eight variables with univariate anomalies brings
  - 1770 instead of 2278 plots
  - Time saving about 17%
  - 108 Outliers instead of 144
  - More purely bivariate anomalies within the Outliers
Conclusions

- The `scagnostics` package is very helpful to get a first overview of unknown data
- It could be useful in some situations to have more flexibility
  - for the binning
  - for outliers (different ways of outlier detection, show outliers, show plots without outliers)
  - for univariate filtering

Further potential
- Use parallel computing to speed up computation
- Extend the idea to other plots and other dimensions