binostics: computing scagnostics measures in R and C++

by Ursula Laa, Dianne Cook, Hadley Wickham, Heike Hofmann

Abstract An abstract of less than 150 words.

Introduction

- scatter plot matrix, need for variable selection with big data
- scagnostics (Wilkinson et al., 2005, Wilkinson and Wills (2008)): characterise scatterplots with 8 summaries (scatterplot diagnostics), select interesting plots from a SPLOM of the scagnostics measures and interactively look at those
- similar methods (e.g. MIC(Reshef et al., 2011), available via minerva package (Albanese et al., 2012))
- alternative methods to see combinations of variables: tours (Asimov, 1985, Wickham et al. (2011)), PCA, LDA, PP (Friedman, 1987)

things to mention: scagnostics package (CRAN) (Urbanek et al., 2012), mbgraphic (Grimm, 2017) (archived!) and Katrins thesis work (Grimm, 2016)

Scagnostics measures

Following Wilkinson et al. (2005) the scagnostics measures are evaluated on the data after hexagon binning, and based on the Delaunay triangulation for comutational efficiency. In addition, outlying points are removed before computing the measures (with exception of the Outlying measure), to make the measures more robust.

Underlying definitions

All measures are based on graphs, i.e. a set of vertices and edges, that can all be extracted from the Delaunay triangulation, namely:

- the convex hull, i.e. the outer edges of the Delaunay triangulation
- the alpha hull (Edelsbrunner et al., 1983)
- the minimum spanning tree (MST) (Kruskal, 1956)

In the following, the length of the MST is defined as the sum of lengths of all edges, and q_x is defined as the xth percentile of the MST edge lengths.

When computing the alpha hull, α is set to q_{90} (Wilkinson et al., 2005). The definition of outlying points is also based on edge length, with points being tagged as outlying if all adjacent edges in the MST have a lenght larger than

$$w = q_{75} + 1.5(q_{75} - q_{25}). \tag{1}$$

Several of the measures are corrected for dependence on sample size using the weight

$$\omega = 0.7 + \frac{0.3}{1 + t^2}, \tag{2}$$

where $t = n/500$ and n the sample size. This correction weight was determined by comparing the scagnostics measures over a large number of datasets (Wilkinson et al., 2005).

Measure definitions

- **Outlying**: compares the edge lengths of the MST of outlying points with the length of the original MST T

$$c_{\text{outlying}} = \frac{length(T_{\text{outliers}})}{length(T)} \tag{3}$$

- **Skewed**: is measuring skewness in the distribution of MST edge lengths (and thus large values might not correspond to skewed distributions of points) as

$$c'_{\text{skewed}} = \frac{q_{90} - q_{50}}{q_{90} - q_{10}}, \tag{4}$$

and is corrected to adjust for dependence on the sample size as

$$c_{\text{skewed}} = 1 - \omega(1 - c'_{\text{skewed}}). \tag{5}$$

- **Sparse**: detects if points are only found in small number of locations in the plane, as is the case for categorical variables,

$$c_{\text{sparse}} = \omega q_{90}. \tag{6}$$

- **Clumpy**: to detect clustering we split the MST in two parts by removing a single edge j, and compare the largest edge lenght within the smaller of the two subsets to the length of the removed edge j. The clumpy measure is defined by maximising over all edges in the MST as

$$c_{\text{clumpy}} = \max_j[1 - \max_k[\text{length}(e_k)]/\text{length}(e_j)], \tag{7}$$

with k running over all edges in the smaller subgraph found after removing edge j from the MST.

- **Striated**: aims to detect patterns like smooth algebraic functions or parallel lines, by evaluating the angles between adjacent edges,

$$c_{\text{striated}} = \frac{1}{|V|} \sum_{v \in V^{(2)}} I(\cos\theta_{e(v,a)e(v,b)} < -0.75), \tag{8}$$

where V is the set of vertices, $|V|$ the total number of vertices in the triangulation, $V^{(2)}$ the subset of vertices with two edges (i.e. vertices of degree two), and I the indicator function.

- **Convex**: convexity is computed as the ratio of the area of the alpha hull A and the convex hull H, adjusting for sample size dependence,

$$c_{\text{convex}} = \omega \frac{area(A)}{area(H)}. \tag{9}$$

- **Skinny**: The ratio of the perimeter to the area of the alpha hull A, with normalization such that zero corresponds to a full circle and values close to 1 indicate a skinny polygon,

$$c_{\text{skinny}} = 1 - \frac{\sqrt{4\pi area(A)}}{perimeter(A)}. \tag{10}$$

- **Stringy**: a stringy distribution should have no branches in the MST. This is evaluated by counting the number of vertices of degree two and comparing them to the total number of vertices (dropping those of degree one),

$$c_{\text{stringy}} = \frac{|V^{(2)}|}{|V| - |V^{(1)}|}. \tag{11}$$

- **Monotonic**: monotonicity is evaluated via the squared Spearman correlation coefficient, i.e. the Pearson correlation between the ranks of x and y,

$$c_{\text{monotonic}} = r^2_{\text{Spearman}}. \tag{12}$$

XXX maybe there is a problem with the implementation of outlying, because this sometimes gives values larger than one? (e.g. vs vs am on mtcars, also gives NaN for several other measures: clumpy, striated, stringy, monotonic)

Implementation

- input/output structure, how are we combining variables
- c++ interface for efficient binning, triangulation
- computation of measures
- how to work with the resulting output

Input is either: as raw function takes two vectors to calculate the scagnositcs for, or a 2D array (matrix or data frame) for which all combinations of variables will be considered

Output: the raw function (called on two vecotrs) returns a list which contains the scagnostics measures (`s`) and the binning information (`bins`), when called on a 2D array the output is of the "scagdf" class. The main result is stored in matrix format, with each column corresponding to one of the scagnositcs measures, and each row a combination of input variables. The row names specify this combination, e.g. `scagdf["x vs y",]` returns all scagnostics measures computed on the scatterplot of

variable x vs variable y. Additional attributes are "vars" (specifying the assignment of combinations of input variables to rows in the output) and "data" (a copy of the input data).

The R interface handels the reading the input and output formatting, and for each combination of variables we call C++ functionalities to compute the measures. The steps are:

- hexbinning of the data (here number of bins is free parameter, but if we find more than 250 non-empty bins in the result, redo binning with half the number of bins on each axis)
- computing the Delaunay triangulation and the MST
- use cutoff on edge length (based on MST) to identify outlying points, and recompute DT after removing outliers
- compute scagnostics measures

Example

A simple example showcasing how to use scagnostics measures

```
# FIXME need better examples!

library(tidyverse)
s <- binostics::scagnostics(mtcars)
s_tibble <- tibble::as_tibble(s[1:nrow(s),]) %>% #get tibble for plotting
  dplyr::mutate(vars = rownames(s))
GGally::ggpairs(select(s_tibble, -vars))
filter(s_tibble, Skinny==1 & Convex==1)$vars # these are discrete values only on the "outside"
filter(s_tibble, Skinny==1 & Convex==0)$vars # other discrete values give Convex=0
filter(s_tibble, Outlying==2 & Skewed==0)$vars # also points to discrete values only on the "outside"
```

Summary

- scagnostics measures useful when exploring large datasets
- the binostics implementaton is efficient thanks to c++ interface, and portable (no java dependence)
- most useful for interactive exploration (maybe good to use in Shiny app?)
- connection with PP (cite PPI paper)

Bibliography

D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello. minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics*, 29 (3):407–408, 12 2012. ISSN 1367-4803. [p1]

D. Asimov. The Grand Tour: A Tool for Viewing Multidimensional Data. *SIAM Journal of Scientific and Statistical Computing*, 6(1):128–143, 1985. [p1]

H. Edelsbrunner, D. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983. [p1]

J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266, 1987. [p1]

K. Grimm. *Kennzahlenbasierte Grafikauswahl*. doctoral thesis, Universität Augsburg, 2016. [p1]

K. Grimm. *mbgraphic: Measure Based Graphic Selection*, 2017. URL https://CRAN.R-project.org/package=mbgraphic. R package version 1.0.0. [p1]

J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956. [p1]

D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062): 1518–1524, 2011. ISSN 0036-8075. [p1]

S. Urbanek, L. Wilkinson, and A. Anand. scagnostics: Compute scagnostics - scatterplot diagnostics. https://cran.r-project.org/web/packages/scagnostics/index.html, 2012. [p1]

H. Wickham, D. Cook, H. Hofmann, and A. Buja. tourr: An R package for exploring multivariate data with projections. *Journal of Statistical Software*, 40(2):1–18, 2011. [p1]

L. Wilkinson and G. Wills. Scagnostics distributions. *Journal of Computational and Graphical Statistics*, 17 (2):473–491, 2008. [p1]

L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 157–164, Oct 2005. [p1]

Ursula Laa
Affiliation
line 1
line 2
author1@work

Dianne Cook
Affiliation
line 1
line 2
author2@work

Hadley Wickham
Affiliation
line 1
line 2
author3@work

Heike Hofmann
Affiliation
line 1
line 2
author4@work