Teaching Computers to See Patterns in Scatterplots with Scagnostics

by Harriet Mason, Stuart Lee, Ursula Laa, and Dianne Cook

Abstract An abstract of less than 150 words.

Introduction

Visualising high dimensional data is often difficult and requires a trade-off between the usefulness of the plots and maintaining the structures of the original data. Scagnostics (scatterplot diagnostics) are a set of visual features that can be used to identify interesting and abnormal scatterplots, and thus give a sense of priority to the variables we choose to visualise. This proposal will discuss the creation of an R package that will provide a user-friendly method to calculate these scagnostics. The package will be tested on datasets with known interesting visual features to ensure the scagnostics are working as expected, before finally being used to explore and describe a time series dataset.

As the number of dimensions in a dataset increases, the process of visualising its structure and variable dependencies becomes more tedious. This is because the number of possible pairwise plots rises exponentially with the number of dimensions. Datasets like Anscombe's quartet (Anscombe, 1973) or the datasaurus dozen (Locke and D'Agostino McGowan, 2018) have been constructed such that each pairwise plot has the same summary statistics but strikingly different visual features. This design is to illustrate the pitfalls of numerical summaries and the importance of visualisation. This means that despite the issues that come with increasing dimensionality, visualisation of the data cannot be ignored. Scagnostics offer one possible solution to this issue.

The term scagnostics was introduced by John Tukey in 1982 (Tukey, 1988). Tukey discusses the value of a cognostic (a diagnostic that should be interpreted by a computer rather than a human) to filter out uninteresting visualisations. He denotes a cognostic that is specific to scatter plots a scagnostic. Up to a moderate number of variables, a scatter plot matrix (SPLOM) can be used to create pairwise visualisations, however, this solution quickly becomes infeasible. Thus, instead of trying to view every possible variable combination, the workload is reduced by calculating a series of visual features, and only presenting the outlier scatter plots on these feature combinations.

There is a large amount of research into visualising high dimensional data, most of which focuses on some form of dimension reduction. This can be done by creating a hierarchy of potential variables, performing a transformation of the variables, or some combination of the two. Unfortunately none of these methods are without pitfalls. Linear transformations are subject to crowding, where low level projections concentrate data in the centre of the distribution, making it difficult to differentiate data points (Diaconis and Freedman, 1984). Non-linear transformations often have complex parameterisations, and can break the underlying global structure of the data, creating misleading visualisations. While there are solutions within these methods to fix these issues such as a burning sage tour for crowding (Laa et al., 2020a) or liminal package for maintaining global structure (Lee et al., 2020) all these methods still involve some transformation of the data. Scagnostics gives the benefit of allowing the user to view relationships between the variables in their raw form. This means they are not subject to the linear transformation issue of crowding, or the non-linear transformation issue of misleading global structures. That being said, only viewing pairwise plots can leave our variable interpretations without context. Methods such as those shown in *ScagExplorer* (Dang and Wilkinson, 2014) try to address this by visualising the pairwise plots in relation to the scagnostic measures distribution, but ultimately the lack of context remains one of the limitations of using scagnostics alone as a dimension reduction technique.

Scagnostics are not only useful in isolation, they can be applied in conjunction with other techniques to find interesting feature combinations of the transformed variables. The tourr projection pursuit currently uses a selection of scagnostics to identify interesting low level projections and move the visualisation towards them (Laa and Cook, 2020). Since scagnostics are not dependent on the type of data, they can also be used to compare and contrast scatter plots regardless of the discipline. In this way, they are a useful metric for something like the comparisons described in *A self-organizing, living library of time-series data*, which tries to organise time series by their features instead of on their metadata (Fulcher et al., 2020).

Several scagnostics have been previously defined in *Graph-Theoretic Scagnostics* (Wilkinson et al., 2005), which are typically considered the basis of the visual features. They were all constructed to range [0,1], and later scagnostics have maintained this scale. The formula for these measures were revised in *Scagnostic Distributions* and are still calculated according to this paper (Wilkinson and Wills,

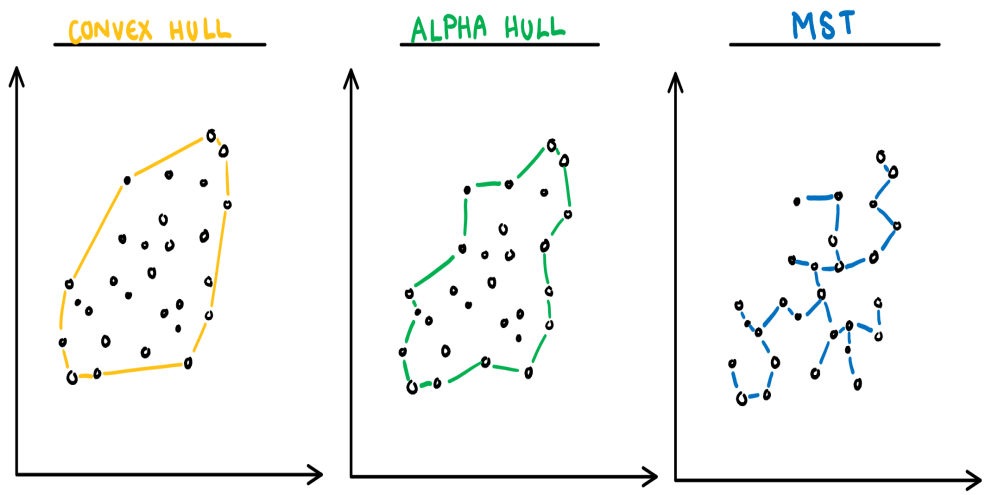


Figure 1: The building blocks for graph-based scagnostics
(#fig:building-blocks)

2008). In addition to the main nine, the benefit of using two additional association scagnostics were discussed in Katrin Grimm's PhD thesis (Grimm, 2016). These two association measures are also used in the tourr projection pursuit (Laa and Cook, 2020).

There are two existing scagnostics packages, *scagnostics* (Wilkinson and Wills, 2008) and the archived package *binostics* (Laa et al., 2020b). Both are based on the original C++ code from *Scagnostic Distributions* (**?**), which is difficult to read and difficult to debug. Thus there is a need for a new implementation that enables better diagnosis of the scagnostics, and better graphical tools for examining the results.

This paper describes the R package, `cassowaryr` that computes the currently existing scagnostics, and adds several new measures. The paper is organised as follows. The next section explains the scagnostics. This is followed by a description of the implementation. Several examples using collections of time series and XXX illustrate the usage.

Scagnostics

Building blocks for the graph-based metrics

In order to capture the visual structure of the data, graph theory is used to calculate most of the scagnostics. The pairwise scatter plot is re-constructed as a graph with the data points as vertices and the edges are calculated using Delaunay triangulation. In the package this calculation is done using the alphahull package (Pateiro-Lopez et al., 2019) to construct an object called a scree. This is the basis for all the other objects that are used to calculate the scagnostics (except for monotonic, dcor and splines which use the raw data). The graph (screen object) is then used to construct the three key structures on which the scagnostics are based; the convex hull, alpha hull and minimum-spanning tree (MST).

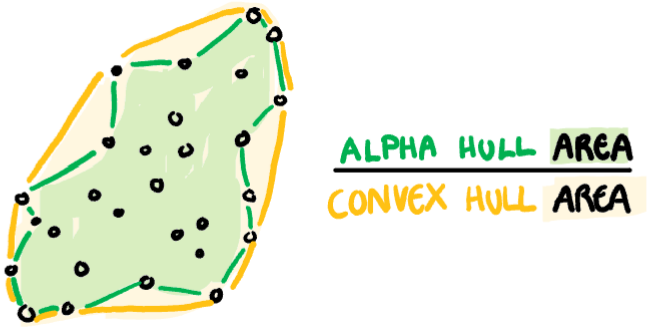
- **Convex hull:** The outside vertices of the graph, connected to make a convex polygon that contains all points. It is constructed usnig the tripack package.
- **Alpha hull:** A collection of boundaries that contain all the points in the graph. Unlike the convex hull, it does not need to be convex. It is calculated using the alphahull package (Pateiro-Lopez et al., 2019).
- **MST:** the minimum spanning tree, i.e the smallest distance of branches that can be used to connect all the points. In the package it is calculated from the graph using the igraph package (Csardi and Nepusz, 2006).

Graph-based scagnostics

The nine scagnostics defined in *Scagnostic Distributions* are detailed below with an explanation, formula, and visualisation. We will let A= alpha Hull C= convex hull, M = minimum spanning tree, and s= the scagnostic measure. Since some of the measures have some sample size dependence, we will let w be a constant that adjusts for that.

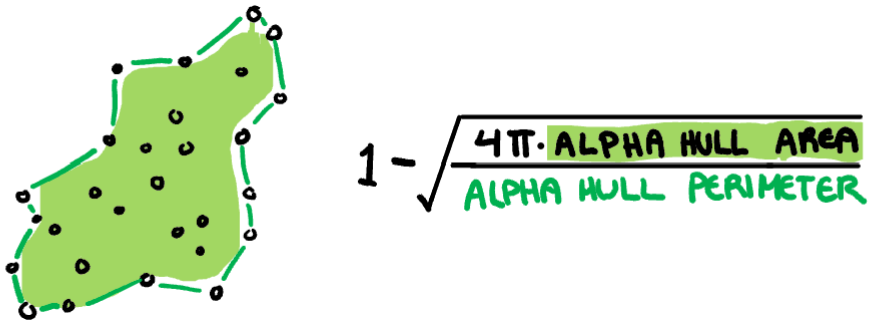
- **Convex:** Measure of how convex the shape of the data is. Computed as the ratio between the area of the alpha hull (A) and convex hull (C).

$$s_{convex} = w\frac{area(A)}{area(C)}$$



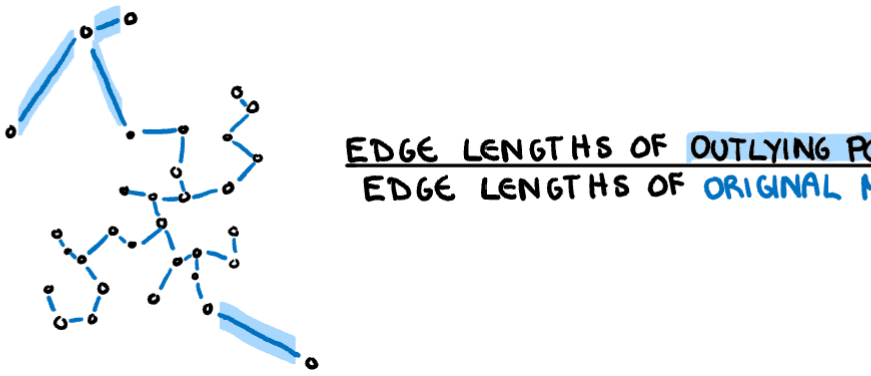
- **Skinny:** A measure of how "thin" the shape of the data is. It is calculated as the ratio between the area and perimeter of the alpha hull (A) with some normalisation such that 0 correspond to a perfect circle and values close to 1 indicate a skinny polygon.

$$s_{skinny} = 1 - \frac{\sqrt{4\pi area(A)}}{perimeter(A)}$$



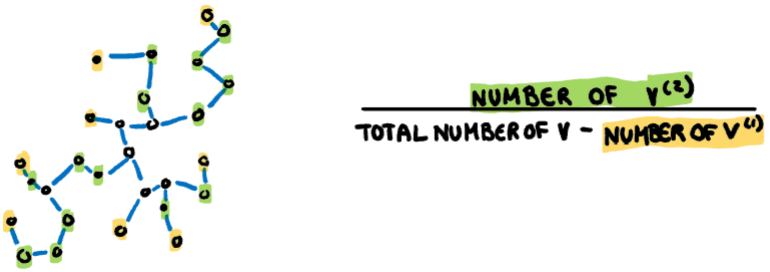
- **Outlying:** A measure of proportion and severity of outliers in dataset. Calculated by comparing the edge lengths of the outlying points in the MST with the length of the entire MST.

$$s_{outlying} = \frac{length(M_{outliers})}{length(M)}$$



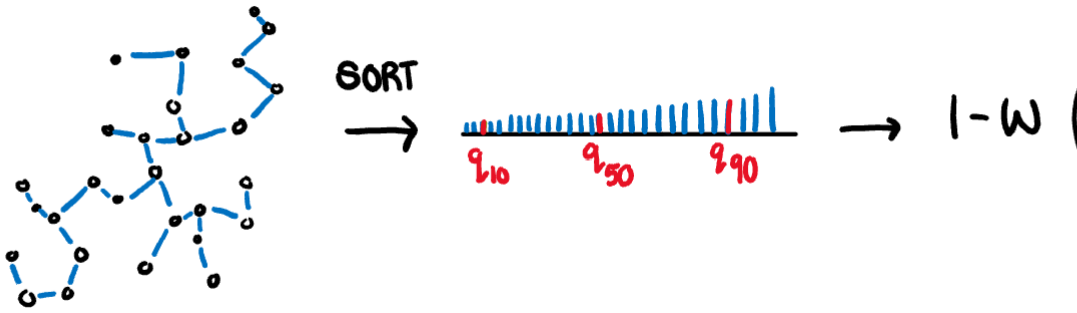
- **Stringy:** This measure identifies a "stringy" shape with no branches, such as a thin line of data. It is calculated by comparing the number of vertices of degree two ($V^{(2)}$) with the total number of vertices (V), dropping those of degree one ($V^{(1)}$).

$$s_{stringy} = \frac{|V^{(2)}|}{|V| - |V^{(1)}|}$$



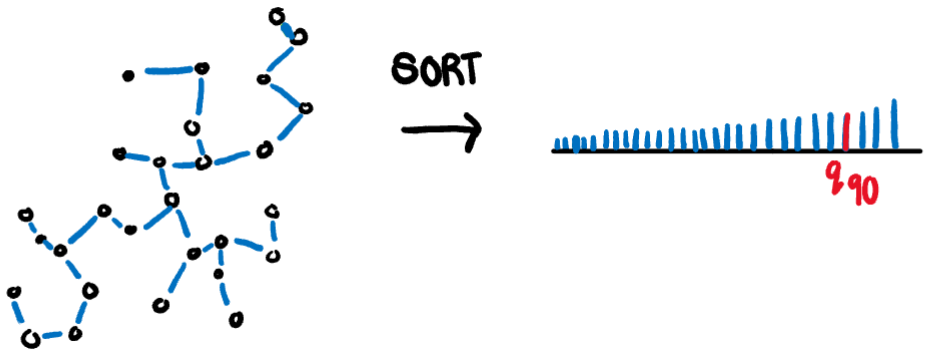
- **Skewed:** A measure of skewness in the edge lengths of the MST (not in the distribution of the data). It is calculated as the ratio between the 40% IQR and the 80% IQR, adjusted for sample size dependence.

$$s_{skewed} = 1 - w(1 - \frac{q_{90} - q_{50}}{q_{90} - q_{10}})$$



- **Sparse:** Identifies if the data is sporadically located on the plane. Calculated as the 90th percentile of MST edge lengths.

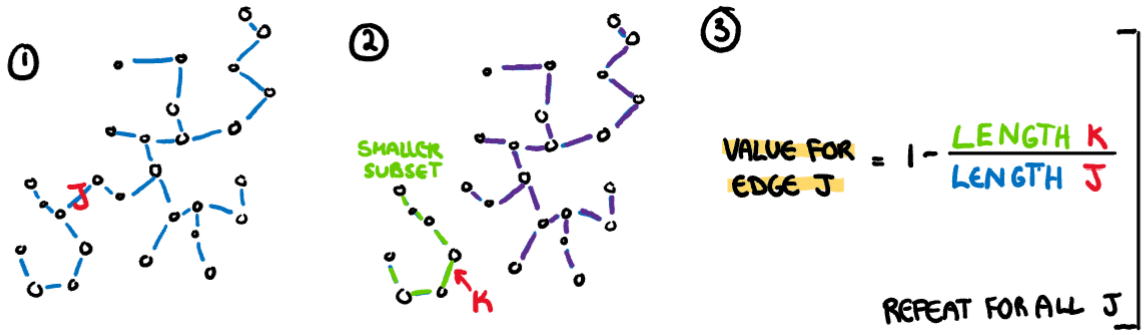
$$s_{sparse} = wq_{90}$$



- **Clumpy:** This measure is used to detect clustering and is calculated through an iterative process. First an edge J is selected and removed from the MST. From the two spanning trees that are created by this break, we select the largest edge from the smaller tree (K). The length of this edge (K) is compared to the removed edge (J) giving a clumpy measure for this edge. This process is

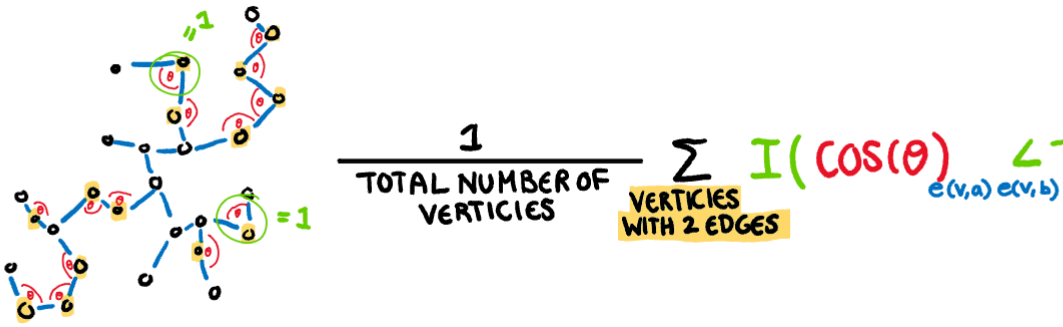
repeated for every edge in the MST and the final clumpy measure is the maximum of this value over all edges.

$$\max_j[1 - \frac{\max_k[length(e_k)]}{length(e_j)}]$$



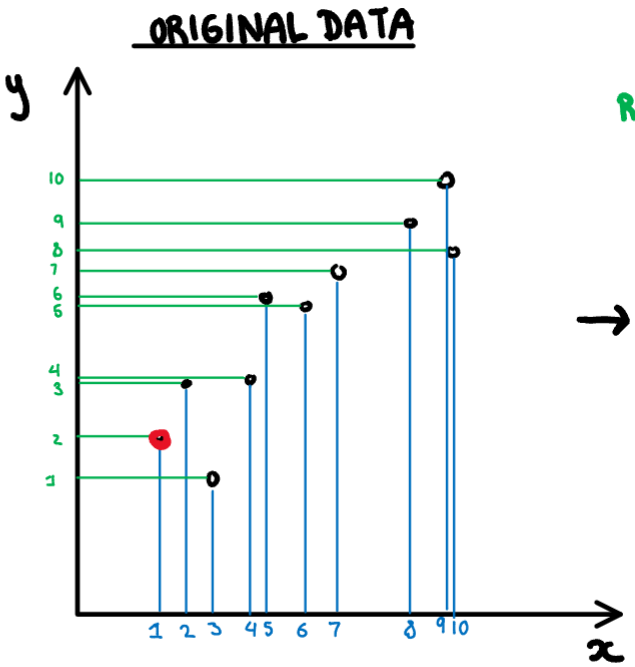
- **Striated:** This measure identifies features such as discreteness by finding parallel lines, or smooth algebraic functions. Calculated by counting the proportion of acute (0 to 40 degree) angles between the adjacent edges of vertices with only two edges.

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



- **Monotonic:** Checks if the data has an increasing or decreasing trend. Calculated as the Spearman correlation coefficient, i.e. the Pearson correlation between the ranks of x and y.

$$s_{monotonic} = r^2_{spearman}$$

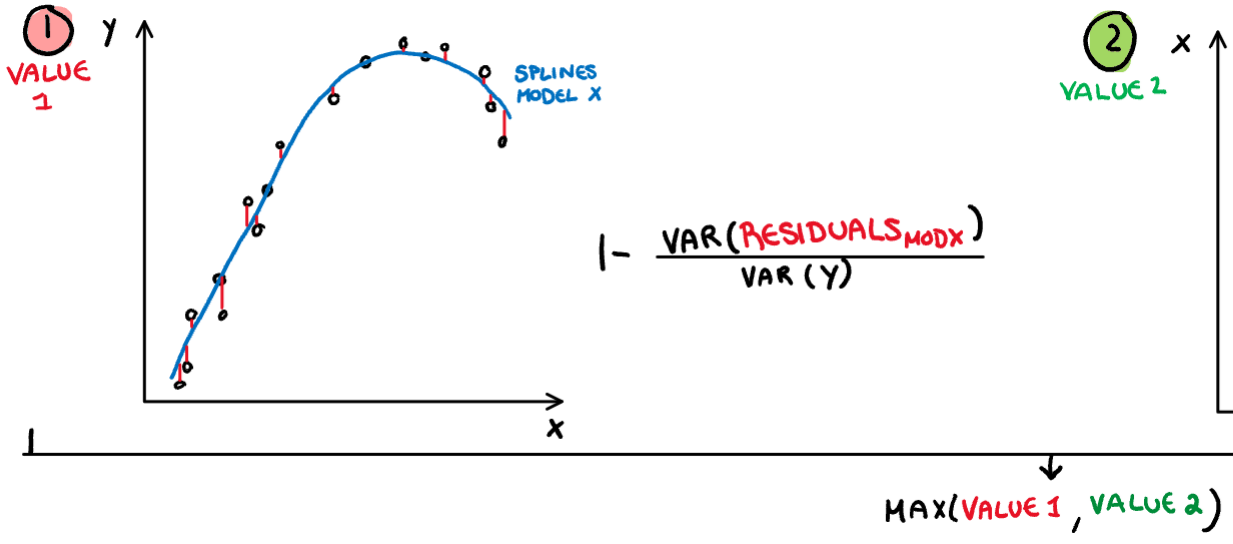


Association-based scagnostics

The two additional scagnostics discussed by Katrin Grimm are described below.

- **Splines:** Measures the functional non-linear dependence by fitting a penalised splines model on X using Y, and on Y using X. The variance of the residuals are scaled down by the axis so they are comparable, and finally the maximum is taken. Therefore the value will be closer to 1 if either relationship can be decently explained by a splines model.

$$s_{splines} = \max_{i \in x,y}[1 - \frac{Var(Residuals_{model\ i=.})}{Var(i)}]$$



- **Dcor:** A measure of non-linear dependence which is 0 if and only if the two variables are independent. Computed using an ANOVA like calculation on the pairwise distances between observations.

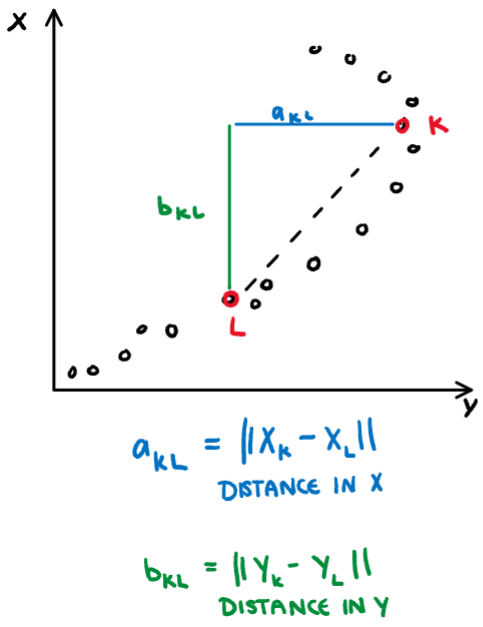
$$s_{dcor} = \sqrt{\frac{\mathcal{V}(X,Y)}{\mathcal{V}(X,X)\mathcal{V}(Y,Y)}}$$

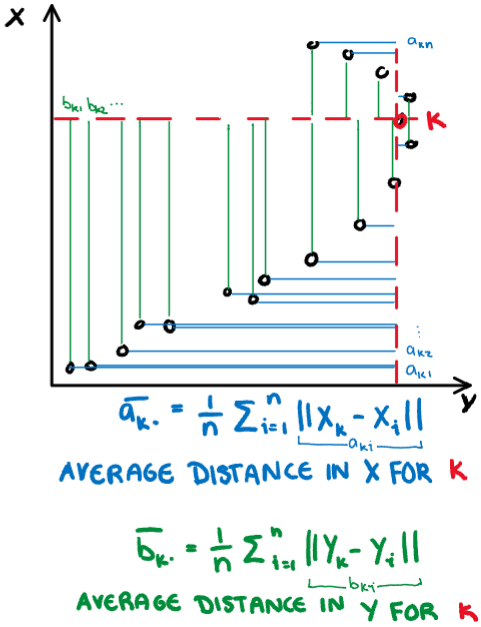
where

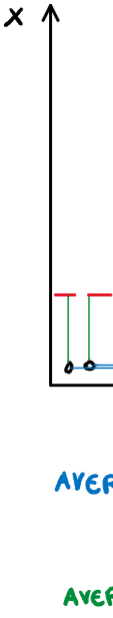
$$\mathcal{V}(X,Y) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} A_{kl}B_{kl}$$

where

$$A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.j} - \bar{a}_{..}$$
$$B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.j} - \bar{b}_{..}$$







Checking the scagnostics calculations

Maybe use Anscombe and datasaurus and the features data here

Software implementation

Installation

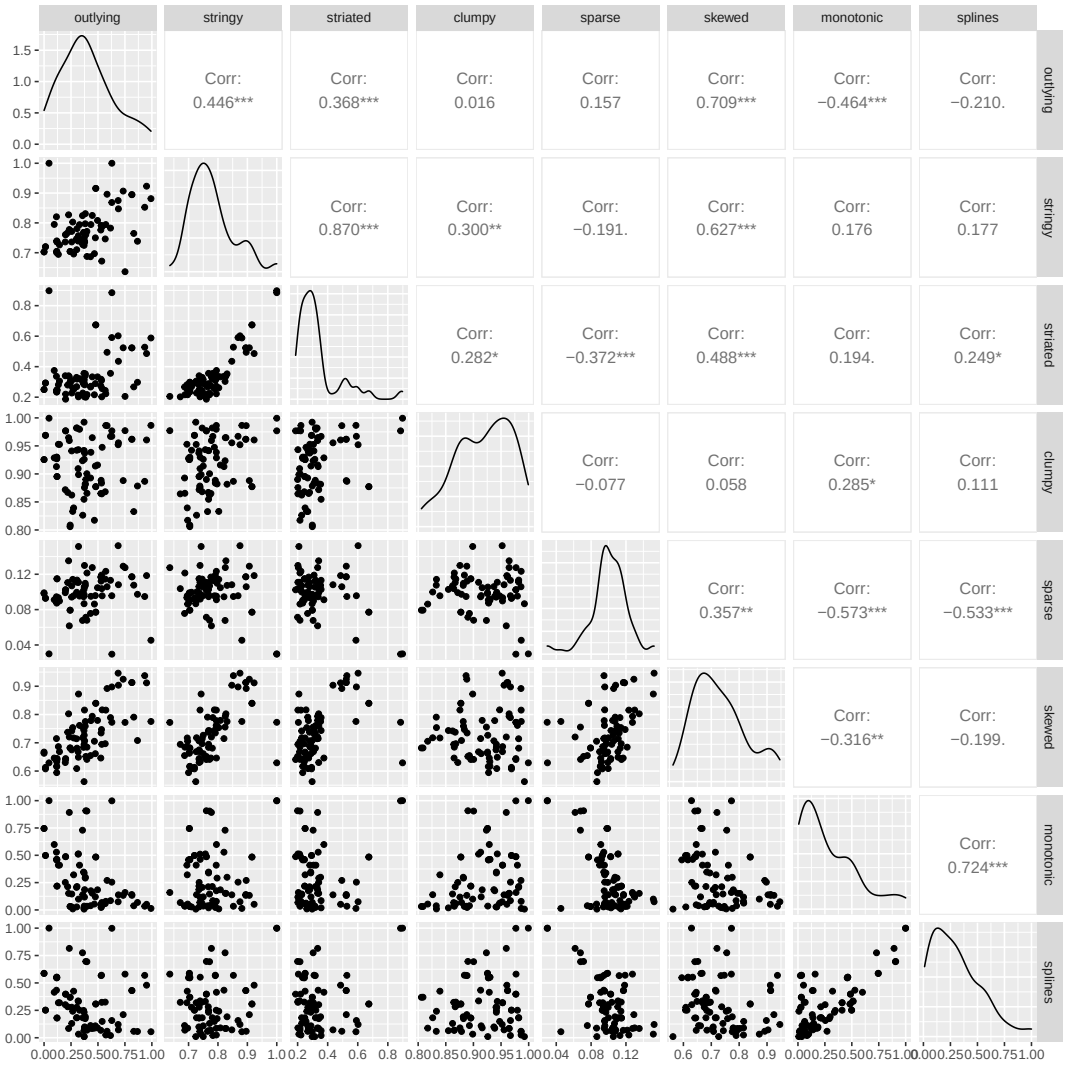
Functions

Tests

Examples

Collections of time series

A paragraph describing the compenginets data



Parkinsons

Black holes and netron star mergers?

Summary

Bibliography

F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973. doi: 10.1080/00031305.1973.10478966. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.1973.10478966. [p1]

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL https://igraph.org. [p2]

T. N. Dang and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *2014 IEEE Pacific Visualization Symposium*, pages 73–80, 2014. doi: 10.1109/PacificVis.2014.42. [p1]

P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12 (3):793–815, 1984. URL http://www.jstor.org/stable/2240961. [p1]

B. D. Fulcher, C. H. Lubba, S. S. Sethi, and N. S. Jones. A self-organizing, living library of time-series data. *Scientific data*, 7(1):213–213, 2020. [p1]

K. Grimm. *Kennzahlenbasierte Grafikauswahl*. doctoral thesis, Universität Augsburg, 2016. [p2]

U. Laa and D. Cook. Using Tours to Visually Investigate Properties of New Projection Pursuit Indexes with Application to Problems in Physics. *Computational Statistics*, 35:1171–1205, 2020. URL https://doi.org/10.1007/s00180-020-00954-8. [p1, 2]

U. Laa, D. Cook, and S. Lee. Burning sage: Reversing the curse of dimensionality in the visualization of high-dimensional data. *arXiv: Computation*, 2020a. [p1]

U. Laa, H. Wickham, D. Cook, and H. Hofmann. binostics: Computing scagnostics measures in r and c++. 2020b. URL https://github.com/uschiLaa/paper-binostics. [p2]

S. Lee, U. Laa, and D. Cook. Casting multiple shadows: High-dimensional interactive data visualisation with tours and embeddings, 2020. [p1]

S. Locke and L. D'Agostino McGowan. *datasauRus: Datasets from the Datasaurus Dozen*, 2018. URL https://CRAN.R-project.org/package=datasauRus. R package version 0.1.4. [p1]

B. Pateiro-Lopez, A. Rodriguez-Casal, and . *alphahull: Generalization of the Convex Hull of a Sample of Points in the Plane*, 2019. URL https://CRAN.R-project.org/package=alphahull. R package version 2.2. [p2]

J. Tukey. *The Collected Works of John W. Tukey*, pages 411,427,433. Chapman and Hall/CRC, 1988. [p1]

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

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]

Harriet Mason
Monash University
Department of Econometrics and Business Statistics
Melbourne, Australia
https://www.britannica.com/animal/quokka
ORCiD: 0000-1721-1511-1101
hmas0003@student.monash.edu

Stuart Lee
Genentech

https://stuartlee.org
ORCiD: 0000-0003-1179-8436
stuart.andrew.lee@gmail.com

Ursula Laa
University of Natural Resources and Life Sciences
Institute of Statistics
Vienna, Austria
https://uschilaa.github.io
ORCiD: 0000-0002-0249-6439
ursula.laa@boku.ac.at

Dianne Cook
Monash University
Department of Econometrics and Business Statistics
Melbourne, Australia
https://dicook.org
ORCiD: 000-0002-3813-7155
dicook@monash.edu