Teaching Computers to See Patterns in Scatterplots with Scagnostics

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

Abstract As the number of dimensions in a dataset increases, the process of visualising its structure and variable dependencies becomes more tedious. 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. Here, we will discuss the creation of the *cassowaryr* R package that will provide a user-friendly method to calculate these scagnostics, as well as the development of adjusted measures not previously defined in the literature. The package is be tested on datasets with known interesting visual features to ensure the scagnostics are working as expected,before being applied to time series, physics and AFLW data to show their value as a preliminary step in exploratory data analysis.

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. 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 which zooms in further on points closer to the middle of a tour to prevent crowding (Laa et al., 2020a), or the liminal package which facilitates linked brushing between a non-linear and linear data transformations to 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

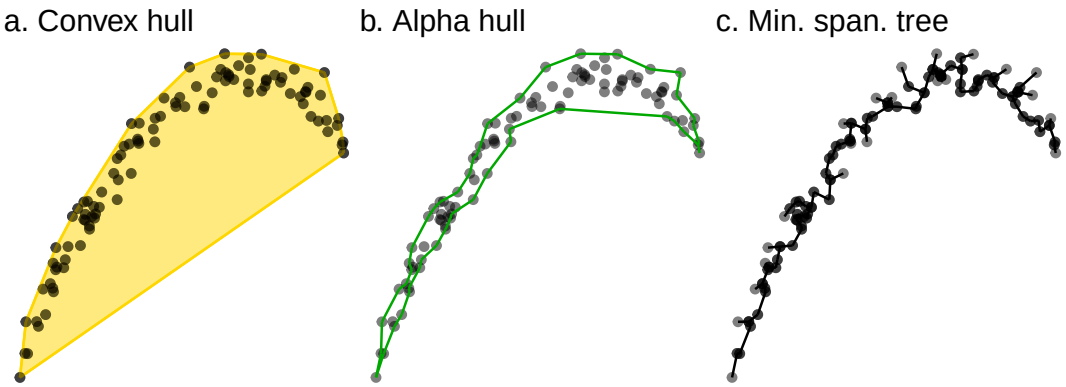


Figure 1: The building blocks for graph-based scagnostics

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, 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* (Wilkinson and Wills, 2008), 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) (Figure @ref(fig:building-blocks2)).

- **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.

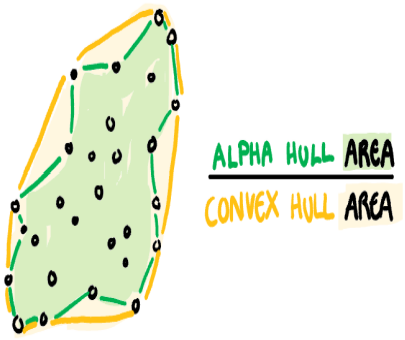


Figure 2: Convex Scagnostic Visual Explanation

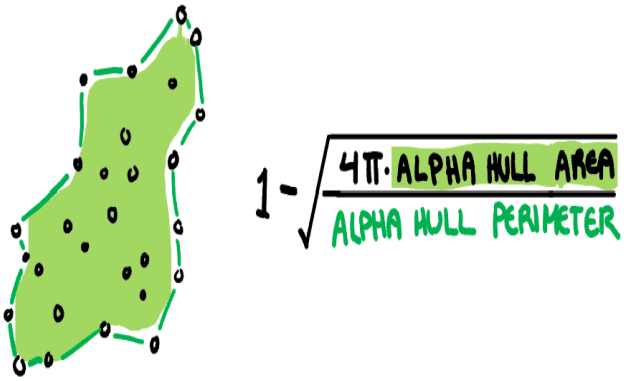


Figure 3: Skinny Scagnostic Visual Explanation

- **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.

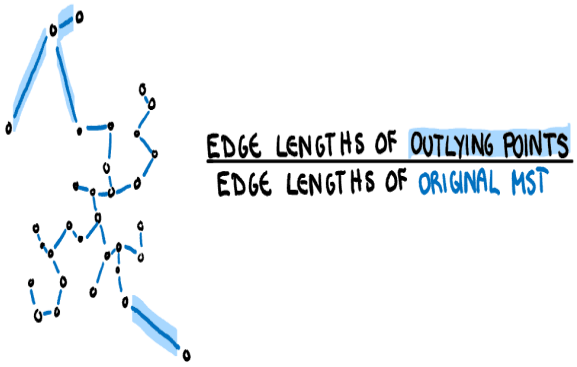


Figure 4: Outlying Scagnostic Visual Explanation

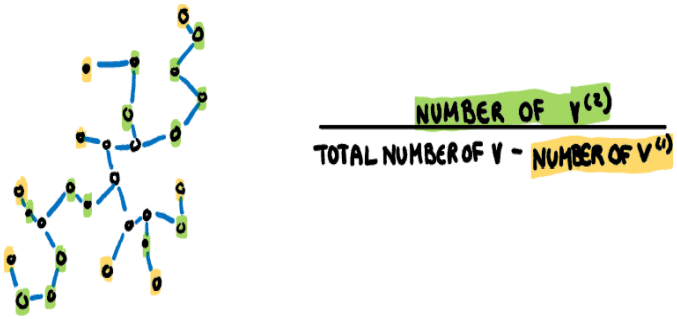


Figure 5: Stringy Scagnostic Visual Explanation

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

- **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.

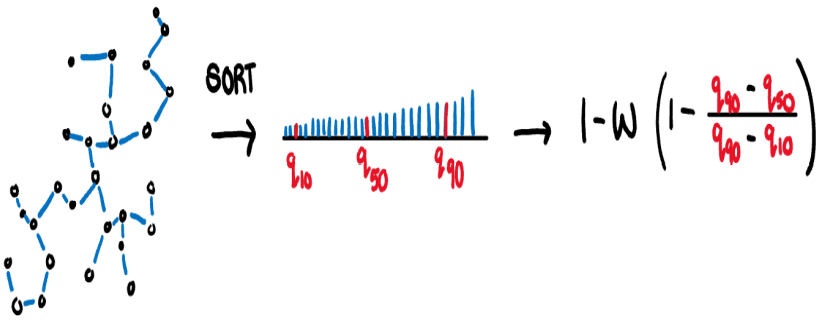


Figure 6: Skewed Scagnostic Visual Explanation

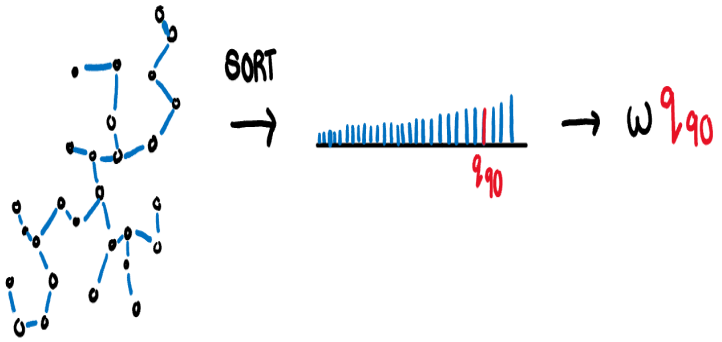


Figure 7: Sparse Scagnostic Visual Explanation

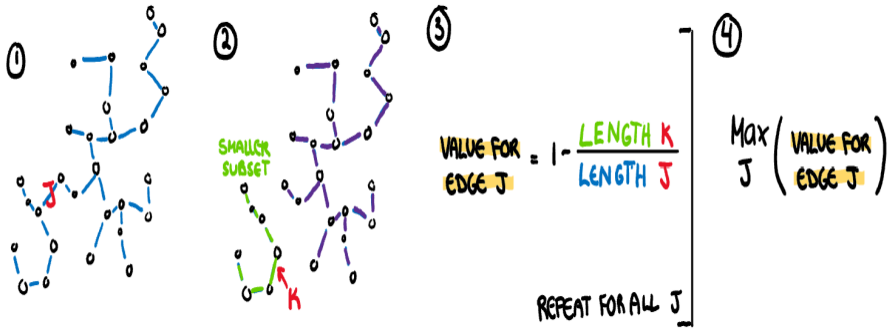


Figure 8: Clumpy Scagnostic Visual Explanation

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

- **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.

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

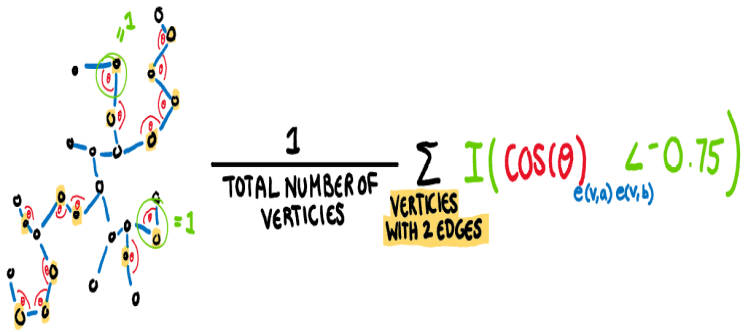


Figure 9: Striated Scagnostic Visual Explanation

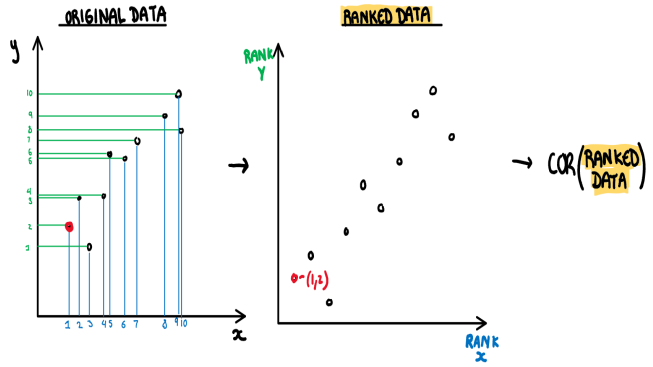


Figure 10: Monotonic Scagnostic Visual Explanation

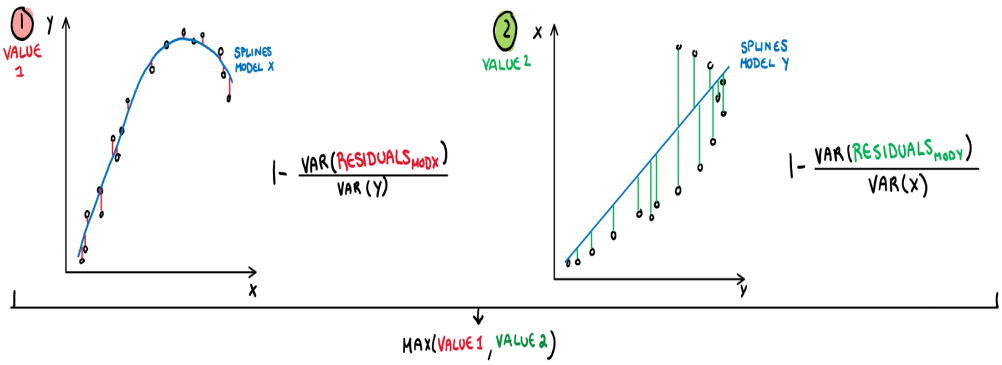


Figure 11: Splines Scagnostic Visual Explanation

Association-based scagnostics

- **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.

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

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

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

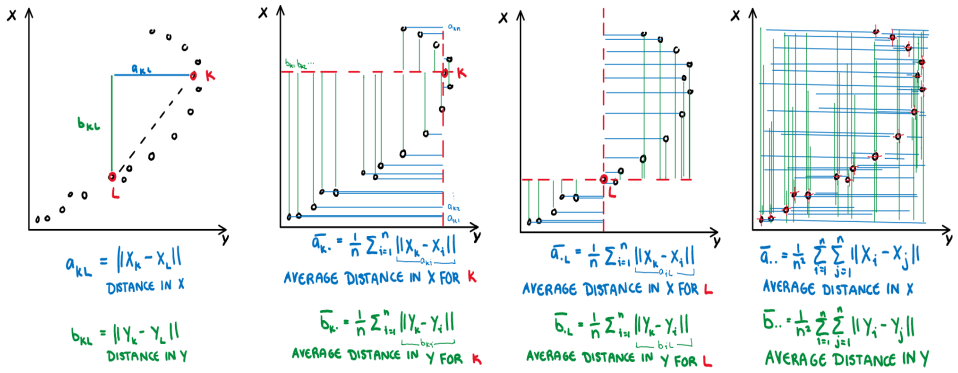


Figure 12: Dcor Scagnostic Visual Explanation

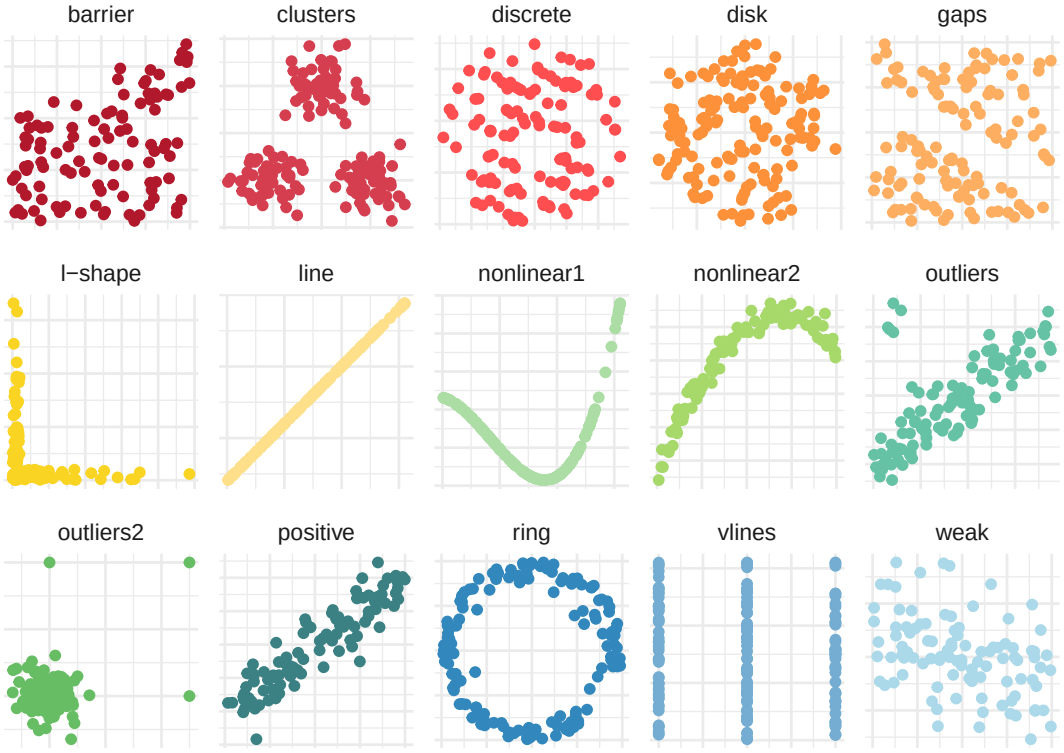


Figure 13: The Scatter Plots of the Features Dataset

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

Checking the scagnostics calculations

Once we have working functions that correctly calculate the scagnostics according to their definition, we can assess how well they identify the visual features of scatter plots. To test the packages ability to differentiate plots, we have creates a dataset called "features" (that is also in the Cassowaryr package) that contains a series of interesting and unique scatter plots which we can run our scagnsotics on.

These scatter plots typify certain visual features we want to look for in scatter plots, be it deterministic relationships (such as that shown in the nonlinear feature), discreteness in variables (vlines), or clustering (clumpy), we should be able to use scagnostics to identify each of these scatter plots. Below is a visual table of an example of a high, a moderate, and a low value, on each scagnostic. The scagnostics are supposed to range from 0 to 1 however in some cases the values are so compressed that a moderate value would not fit, indicating that the scagnostics do not work quite as intended. We suspect the reason for these warped distributions is the removal of binning as a preliminary step in calculating the scagnostics. We wanted the package to have binning as an optional method, considering choices in binning can lead to bias as noted in "Scagnostic Distributions" (Wilkinson and Wills, 2008) or unreproducible results as noted in "Robustness of Scagnostics" . Therefore the current

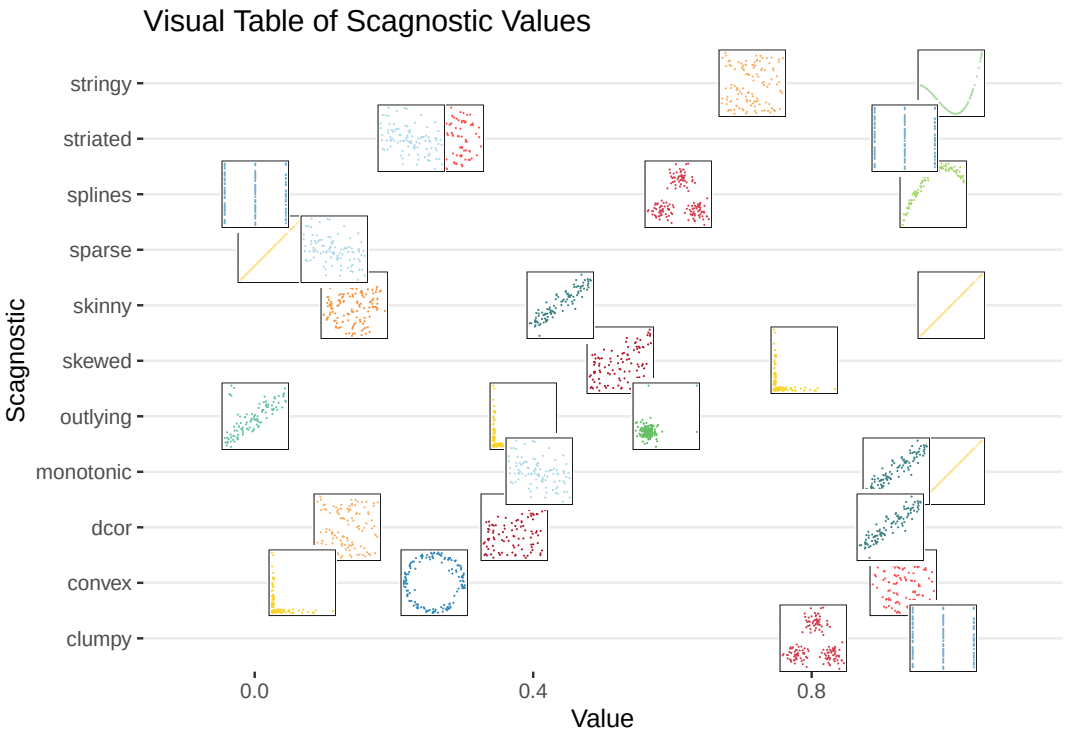


Figure 14: The Features Scatterplots in a Visual Table

scagnostics will be assessed without binning (Wang et al., 2020).

This plot gives a slight idea of the issues some of the scagnsotics face in their current state. The scagnostics based upon the convex hull (i.e. skinny and convex) work fine, as do the association measures such as montonic, dcor and splines. The main issue comes from the measures based on the MST, and their issues largely come from binning. The MST measures and their issues are:

- **Striated**: striated can identify the specific case of one discrete variable and one continuous variable (which alone is not particularly interesting) but will not identify two discrete variables. Since by definition it is a subset of the stringy measure, they are highly correlated, which means most variables that score highly on striated already score highly on stringy, making the measure less useful.
- **Sparse**: While sparse does seem to identify spread out distributions, it rarely returns a value higher than 0.1. As this measure is the 90th percentile of MST edge lengths, and the removal of binning allows for a large number of arbitrarily small edges. In addiiton to this, a larger number of observations will arbitrarily make this value smaller. The addition of new points will increase the number of small edges and decrease the number of large edges, and it is rare that a significantly large edge will be at or below the 90th percentile.
- **Skewed**: this measure can identify skewed edge lengths (such as the L shape in the visual table) however the values on real data rarely drop below 0.5 or rise above 0.8. Skewed seems to suffers from the same issue as sparse reguarding the binning issue and is also heavily influenced by the number of observations in the scatter plot.
- **Outlying**: the disctinction of outlying points described in the scagnostic literature is certainly strange. By definition an outlier must have *all* its adjacent edges in the MST above this threshold, and the visual table displays three interesting cases of this. The first plot (outliers2) returns a 0 even though the handful of points in the top corner would likely be considered to be outliers by a human. This is because within that group the points are close enough that all of them have at least one edge that is below the outlying threshold. Even if we changed the measure such that only one edge needed to be above the outlying threshold, it would only remove a single point. The l-shape shows an increasing spread of the points as they move away from the bottom left corner, as such, the larger edge lengths make sense within the distribution. Outlying does not take this into account, and identifies a large number of the spread out points to be outliers and removes them before computing the other scagnsotics. The value that scores the highest on the outlying measure is, without question, a highly outlying distribution, however the outlying measure only returns a 0.5, this is again due to the removal of binning as a preprocessing step.

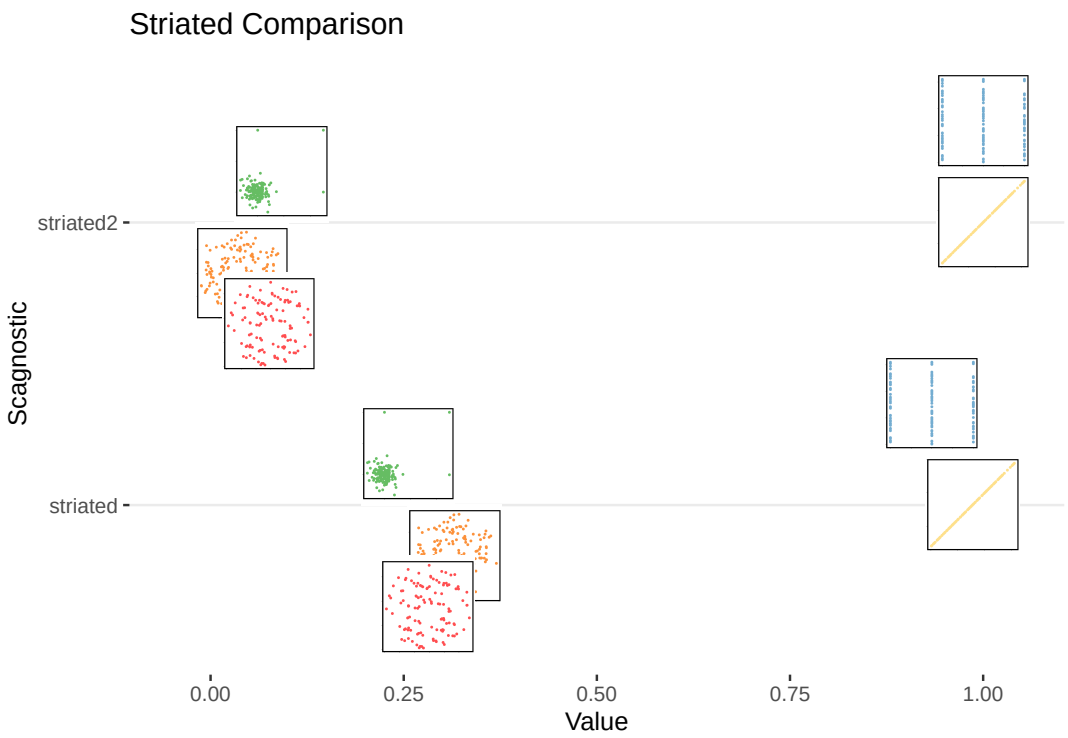


Figure 15: A Visual Table Comparison of Striated and Striated 2

- **Clumpy**: the clumpy scagnostic is probably the one that suffers the most with the removal of the binning step. Due to it being a ratio between an edge and its longest adjacent edge, it does not identify the largest edge, but rather an edge that is connected to an arbitrarily small edge. Because of this, this scagnsotic reliably returns an arbitrarily high value and scatter plots that actually contains clusters (such as clusters) scores low on this measure, while a continuous variable plotted against a discrete variable score arbitrarily high.
- **Stringy**: This measure rarely drops below 0.5 even on data generated from a random normal distribution (which should intuitively return a 0). Unlike the other scagnostics on this list, stringy does not depend upon the edge lengths of the MST, so it is hard to say if this issue stems from binning.

With these issues in mind, we have defines and written several new scagnostics that work even without the pre-processing seto of binning.

The Adjusted Scagnostics Measures

The measures that need an adjusted version are striated, sparse, skewed, and clumpy. The outlying and stringy measure could possibly be left as they are, as they are not as badly damaged by the removal of binning.

Striated Adjusted The issues surrounding the striated scagnostic are:

1. By only counting vertices with 2 edges, the set of vertices counted in this measure are a subset of those counted in stringy, thus the two meaures are highly correlated.
2. In order for the vertex to be counted, the angle between the edges needs to be approximately 135 to 220 degrees. The original idea seems to have been to identify the predominantly 180 degree angles that come with a discrete variable plotted against a continuous one, however the large margin of error just makes the measure almost identical to stringy.

To account for these two issues the striated adjusted measure considers all vertices (not just those with two adjacent edges), and makes the measure strict around the 180 and 90 degree angles. With this we can see the improvements on the measure.

While these two measures may seem similar at a glace, there are a few minor things that make the striated2 scagnsotic an improvement on the stirated scagnsotic. First of all, the perfect 1 value on

striated goes to the "line" scatter plot. While this does fulfil the definition, it is not what the measure is supposed to be looking for, rather supposed to be identifying the "vlines" scatter plot. Since striated does not count the right angles that go between the vertical lines, a truely striated plot will never get a full 1 on this measure, striated adjusted fixes this. After that there is a large gap in both measures because none of the other scatter plots have a strictly discrete measure on the x or y axis. The lower plots show that striated2 is also better at identifying discrete relationships with a rotation and noise added as shown in the "discrete" plot. In striated "discrete" is lower in the order than "outlying" which would indicate that striated has finished looking at discreteness. In striated2, after the plots with strict discreteness in "vlines" or strict rotated disceteness in "line", is the noisey and rotated "discrete" plot. Therefore in terms of ordering the plots in how well they represent the feature of discreteness, striated2 outperforms striated.

The scagnsotics need to be used and interpreted with the type of dataset you are working with in mind. For if we are looking at a dataset that is discrete, a very low value on striated2 would indicate some strange relationship in the scatter plot. Since the old striated measure is specifically trying to find a continuous variable against a discrete variable, its highest values are also identified by the striated2. The lowest values on striated simply identify a plot where all the variables are at right angles, once again a measure of disceteness but one that is not identified by striated. Striated2 encapsulates both versions of discreteness in the values that get exactly a 1.

Clumpy Adjusted The issues that need to be addressed with the new clumpy measure are:

1. It needs to consider more than 1 edge in its final measure to make the measure more robust
2. The impact of the ratio between the long and short edges need to be weighted by the size of their clusters so the measure does not simply identify outliers
3. It should not consider vertices that's adjacent angles form a straight line (to avoid identifying the angles striated identifies)

Before we calculated a new clumpy measure, we looked into applying a different adjustment defined in the *Improving the robustness of scagnsotics* that is a robust version of the original clumpy measure (Wang et al., 2020). This version of clumpy has been included in the package as "clumpy_r" however it is not included as an option in the higher level functions such as calc_scags() because its computation time is too long. This measure tries to build multiple clusters, each having their own clumpy value, and then returns the weighted sum, where each value is weighted by the number of observations in that cluster. This version of clumpy spreads the scatter plots more evenly between 0 and 1 and is more robust to outliers, however it does a poor job of ordering plots generally considered to be clumpy without the assistance of binning. Since this scagnostic cannot be used in large scale scagnostic calculations (such as those done on every pairwise combination of variables as is intended by the package) and it maintains the ordering issue from the original measure, it is not discussed here.

Therefore in order to fix the issues in the clumpy measure described above, we designed an adjusted clumpy measure, called clumpy2 in the package, and it is calculated as follows:

1. Sort the edges in the MST
2. Calculate the difference of adjacent vectors in this ordering, and find the index of the maximum. This maximum difference should indicate the jump from between cluster edges and inter-cluster edges.
3. Remove the between cluster edges from the MST and build clusters using the remaining edges
4. For each between cluster edge, take the smaller cluster (in number of observations) and take its median edge length. The clumpy value for that edge is the ratio between the large and small edge lengths $\frac{edge_{small}}{edge_{large}}$, with a two multiplicative penaltys, one for uneven clusters $\frac{2\times n_{small}}{n_{small}+n_{big}}$, and one for "stringy" scatter plots that is only applied if the stringy value is higher than 0.95, to reduce the arbitrarily large clumpy scores that come from striated plots $1 - s_{stringy}$.
5. Take the mean clumpy value for each between cluster edge, if it is below 1 it is beneath the threshold that is consdiered clumpy, and the value is adjusted to 1.
6. Clumpy 2 returns $1 - \frac{1}{mean(clumpy_i)}$

With this calculation, we generate the clumpy2 measure which is compared to the original clumpy measure in the figure below.

Here we can see the improvements made on the clumpy measure in both distribution from 0 to 1 and ordering. The measure is more spread out, and so values range more accurately from 0 to 1. More importantly the measures do a better job of ordering the scatter plots. On the original clumpy measure the "clusters" scatter plot was next to last, on the clumpy2 measure "clusters" is is identified as the most clumpy scatter plot. Clumpy 2 also has a penalty for uneven clusters (to avoid being large due

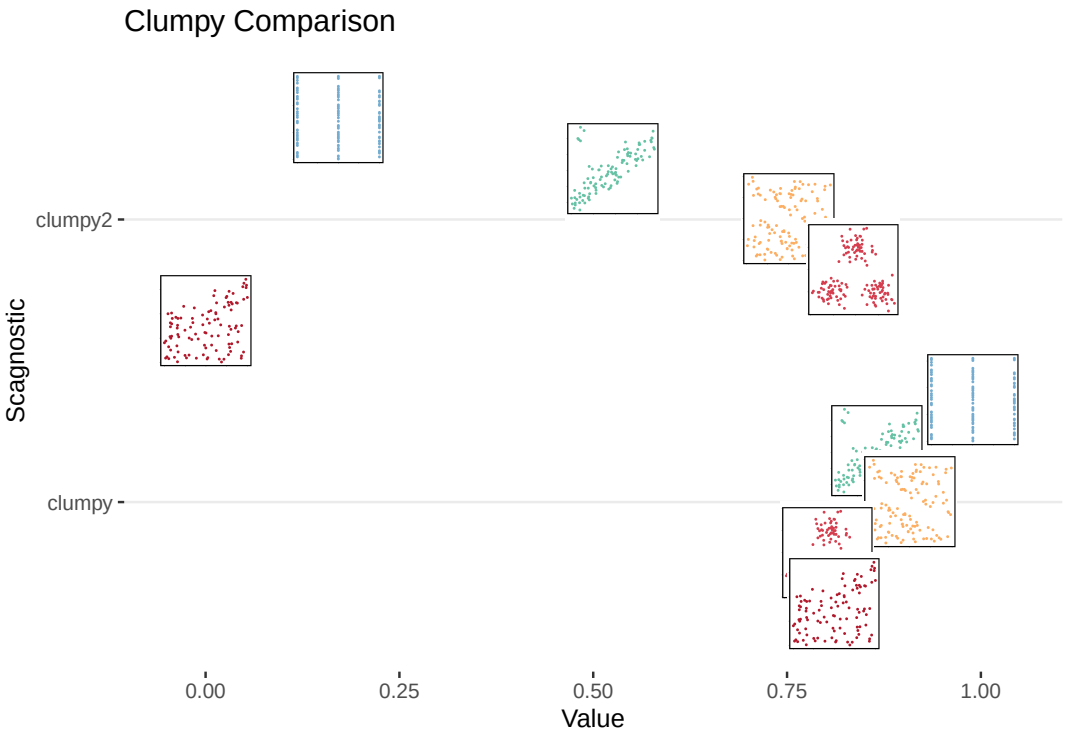


Figure 16: A Visual Table Comparison of Clumpy and Clumpy 2

to a small colelction of outliers) and clusters created arbitrarily due to discreteness (such as vlines) in order to better aling with the human interpretation of clumpy. With these changes, the stronger performance of clumpy2 is apparent in this visual table.

Software implementation

Installation

Data sets

Functions

Scagnostics functions

Drawing functions

Summary functions

Tests

Examples

Collections of time series

GOAL: Use scagnostics to find difference in shapes between groups. Here we want to first use features to describe a time series, and then secondly choosing pairs of features where there is the biggest difference between groups according to a scagnostic.

A paragraph describing the compenginets data

Analysis notes:

- A big collection of time series. How do we understand the range of types of time series?
- Calculate time series features. Using tsfeatures this will give 13 values for each time series.

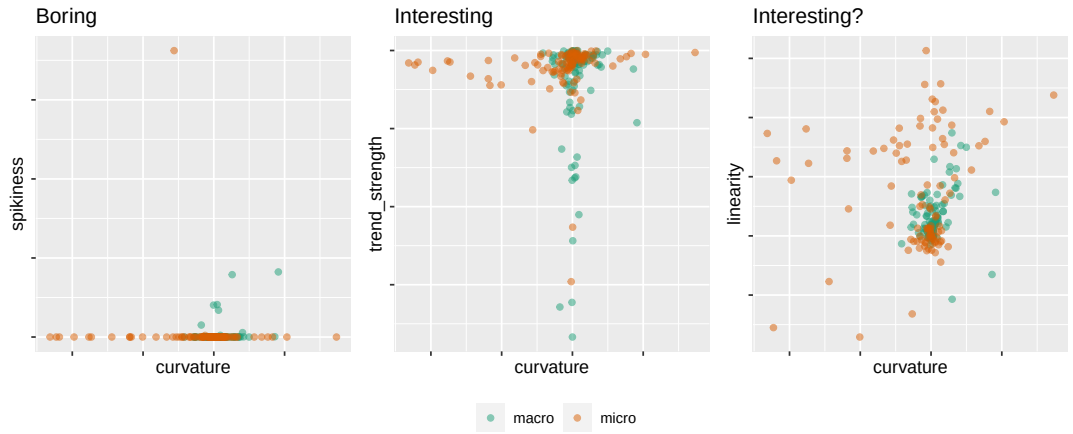


Figure 17: Interesting differences between two groups of time series detected by scagnostics. The time series are described by time series features, in order to handle different length series. Scagnostics are computed on these features separately for each set to explore for shape differences.

- Use scagnostics to find pairs of tsfeatures that are interesting, eg high on splines but low on monotonic
- Plot the pair of tsfeatures, what's unusual about the two?
- Select unusual series from the tsfeatures, and plot

Compare two sets of time series

This analysis compares the features of macroeconomic and microeconomic series, using scagnostics. The goal of the comparison is to compare shapes, not necessarily centres of groups as might be done in LDA or other machine learning methods.

Here, just a small set of features is examined (because code fragile) but what emerges as interesting is the difference between curvature and trend strength (Figure @ref(fig:timeseries)). Microeconomic series tend to have high values on trend strength, and a range of values on curvature. In comparison macroeconomic series tend to have near constant average values on curvature, and highly varied on trend strength.

Plotting a few series actually suggests that the microeconomic series contain lots of micro structure, which might be what we should expect (Figure @ref(fig:tsplots)). Interestingly the trend strength seems to pick up the jaggies!

Black hole mergers

This is a simulated dataset that contains posterior samples for describing an observed gravitational waves signal from a black hole merger in terms of position in the sky (ra, dec, distance), time of the event (time) and the black hole properties (masses m1 and m2; spin related properties alpha, theta_jn, chi_tot, chi_eff, chi_p) and additional nuisance parameters psi (polarisation angle) and phi_jl (orbital phase). There are thus 13 variables and it is still feasible to look at a complete SPLOM, providing a good cross check of the scagnostics.

The data contains 9998 posterior samples, without binning it is too long to compute the scagnostics on such a large number of observations. For our purpose a much smaller sample is sufficient, and we randomly sample 200 observations before computing the scagnostics.

Figure @ref(fig:bbh-scags-static) shows combinations that stand out: time-ra, dec-ra, dec-time (low convex, high skinny) dec-ra and time-ra also have higher splines than dcor (both high, non-linear functional relation), while m1-m2, dec-time and chi_p-chi_tot have higher dcor than splines (still both high, m1-m2 and chi_p-chi_tot are linear relations with noise, dec-time is strong association but not function). (Figure @ref(fig:blackholes) shows the pairs of variables with the detected features.) The final plot is showing clumpy vs skewed, shows that clumpy isn't really doing what we expect since we would expect much more structure in clumpy (in particular plots with time break up into two well separated groups, plots with ra in three separated groups, some other variables introduce less pronounced separation between groups). Included time-alpha as one example, this has clumpy of 0.9 and skewed of 0.7.

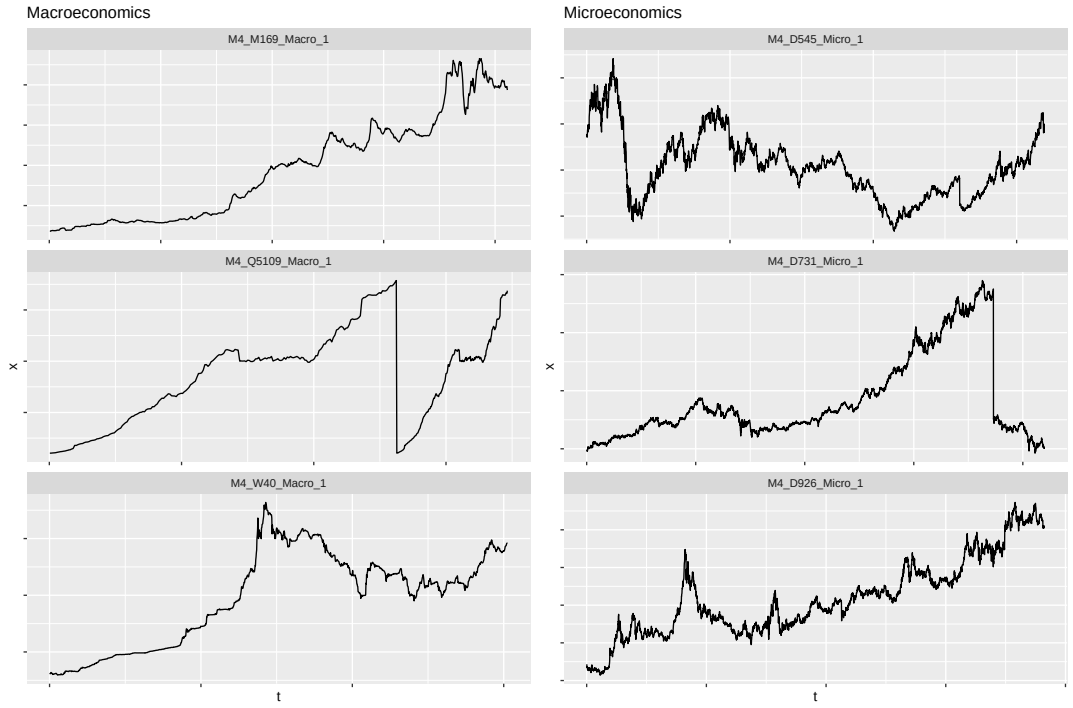


Figure 18: Selection of series from the two groups, macroeconomics and microeconomics. The difference is in the jagginess of the two series.

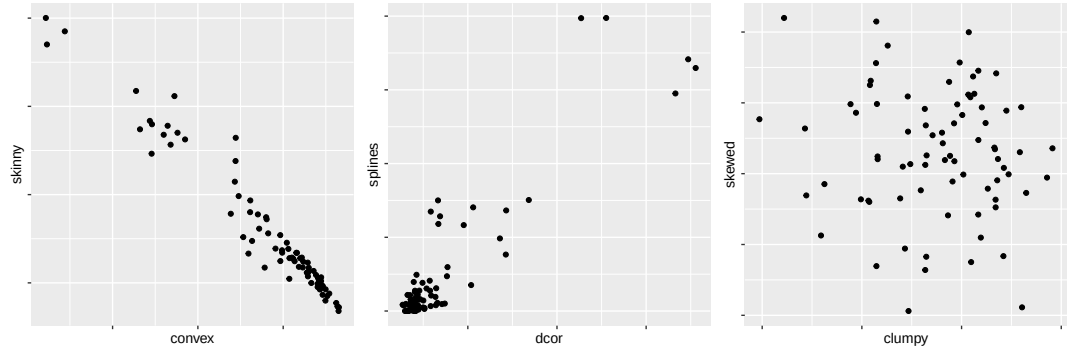


Figure 19: Pairs of scagnostics computed for the black hole mergers data. *XXX*

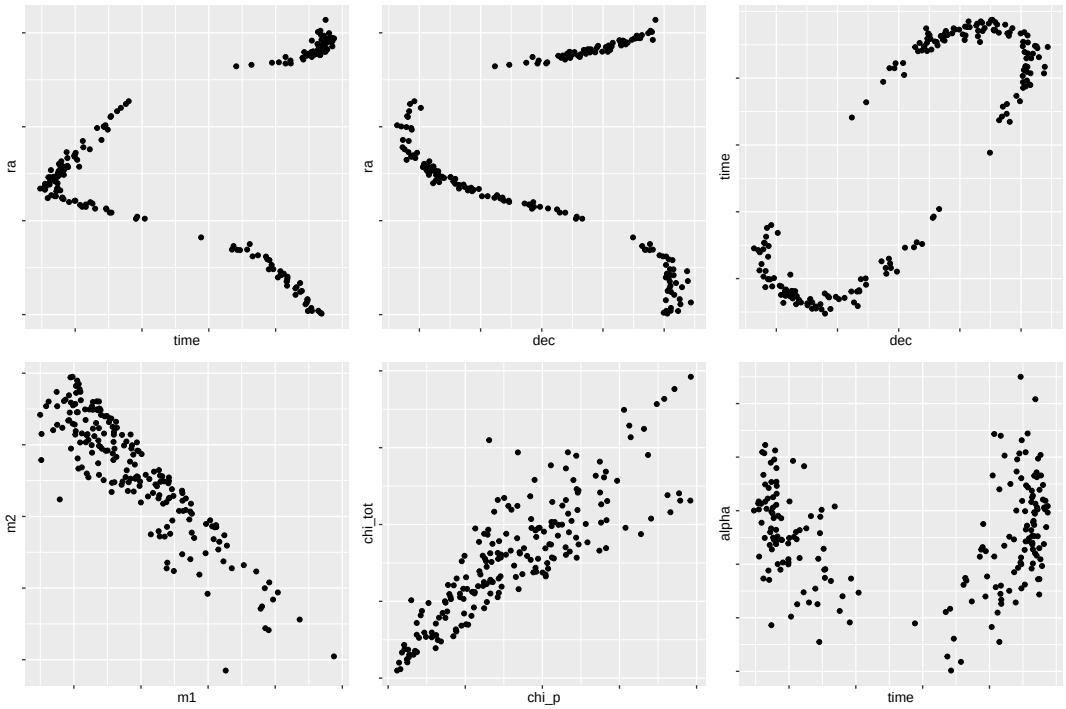
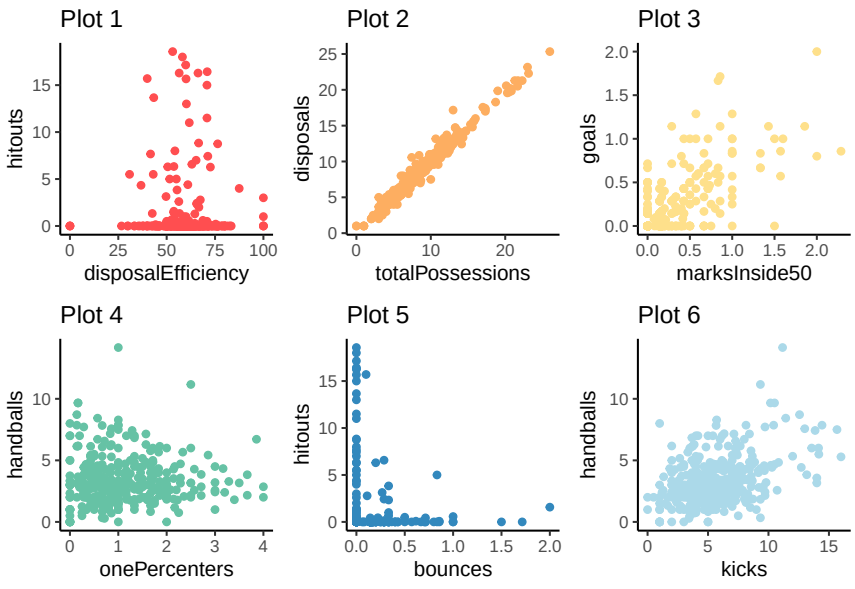


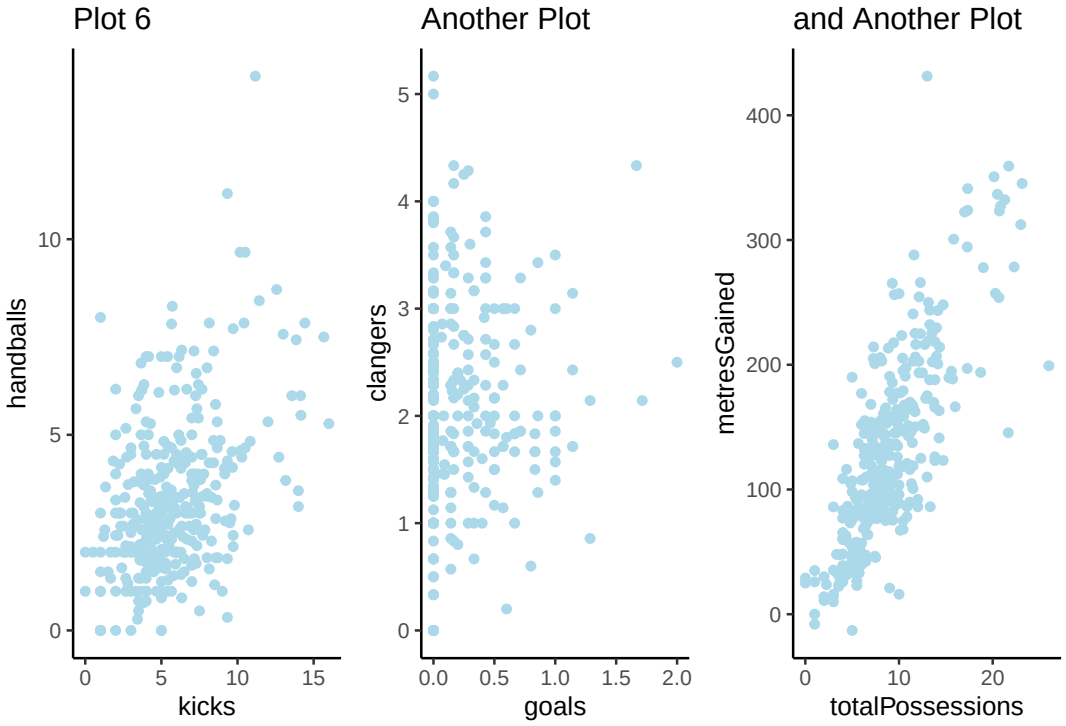
Figure 20: Features in XXX showing XXX

AFL player statistics

The Australian Football League Women's (AFLW) is the national semi-profesisonal Australia Rules football league for female players. Here we will analyse data sourced from the official AFL website with information on the 2020 season, in which the league had 14 teams and 1932 players. There are 68 variables, 38 of which are numeric. The others are categorical, like the players names or match ids, which would not be used in scagnostic calculations. These numeric variables are recorded per player per game, and a description of each variable in this data set can be found in the appendix. With 33 numeric variables, there are 528 possible scatterplots to make. This is much more than we could possibly plot ourselves, and so we can use the scagnostics to identify which might be interesting to examine ourselves. The figure below displays 5 scatter plots that were identified as having a particularly high or low value on a scagnostic, or an unusual combination of two or more scagnostics. In addition to these 5, there is a 6th plot that is included to display what a middling value on almost all of the scagnostics looks like. You may like to test your scagnostics knowledge by guessing which plot is the middling value on all the scagnsotics.

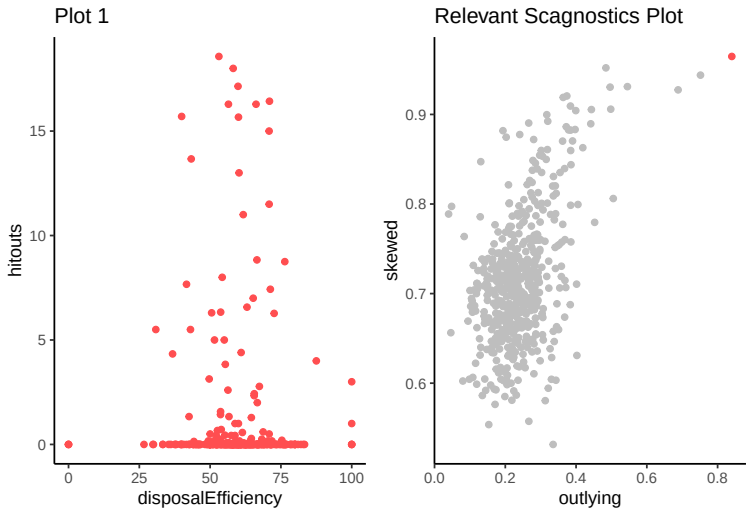


Plots 1 to 5 are examples of unusual combinations of scagnostics, Plot 6 is an example of a scatter plot that was had moderate values across all the scagnostics and was mostly picked at random. We can present Plot 6 alongside two other scatter plots that were selected arbitrarily (the same way we would if we were going to try and do EDA ourselves) to give an idea of what we would get if we arbitrarily selected variables to plot.

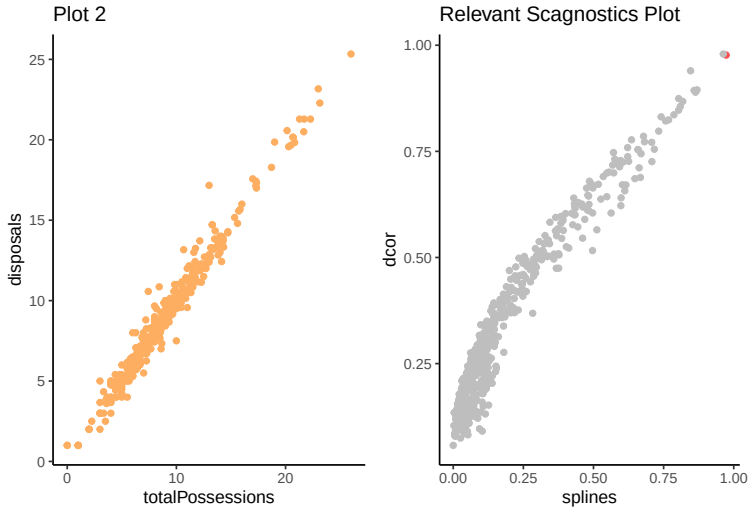


We could plot scatter plots like this all day, but most of the scatter plots in this data set look something like this, and when compared to Plots 1 to 5 we can see that the extreme values on the scagnostic measurements identify atypical scatter plots. While it is interesting to know that scagnostics can pick out interesting scatter plots, we still need to know how to use them. Typically the plots with strange scagnostic combinations are identified using an interactive SPLOM, but for the sake of space, we are only going to show the specific scatter plots of the SPLOM that led to the selection of Plots 1, 2, and 5. Lets start with Plot 1.

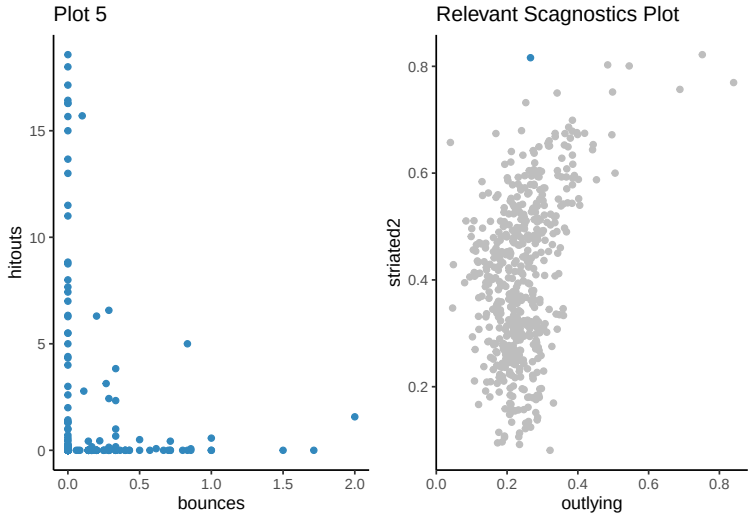
Plot 1 We identified Plot 1 as interesting as it returned high values on both Outlying and Skewed. This indicates that even after removing outliers, the data was still disproportionately spread out, a trend we can see very clearly in the identified scatter plot.



Plot 2 Plot 2 scored very highly on all the association measures, which indicates a strong relationship between the two variables. The three association measures typically have strong correlation, and scatter plots that stay within the large mass in the center have a linear relation, scatter plots that deviate from this large correlation typically have some strong non-linear relationship. Unfortunately that does not appear here, and so none of these variable pairs have strong non-linear relationships, rather our highest scagnostic on the association measures indicates the linear relationship between total posessions and disposals. Total possessions is the number of times the player has the ball and disposals is the number of times the player gets rid of the ball legally, so the high correlation makes sense, this is a professional league so most of the players succeed in getting rid of the ball legally.



Plot 5 This plot is an excellent example in what new information we can learn from a unique pairwise relationship. This scatter plot is separate from the mass of pairwise relationships because it was high on striated_2 and low on outlying, which tells us most of the points are at right angles and a little spread out (but not enough for a high outlying value). This plot tells us something interesting about the physicality of the players. If a specific sports statistic is related to position, we would see a relationship have a lower triangular structure similar to that of Plot 4, however this plot does not have a lower triangular structure, is has an L-shape. This means these statistics are not about position, but rather the physical abilities of the players. Hitouts measure the number of times the player punches the ball after the referee throws it back into play, bounces have to be done while running, and are typically done by fast players. The l-shape tells us that players who do one very rarely perform the other. The moderate spread along both of the statistics tells us these are both somewhat specialised skills, and the players who specialise in one do not specialise in the other, i.e. in AFL the tallest player in the team is rarely the fastest.



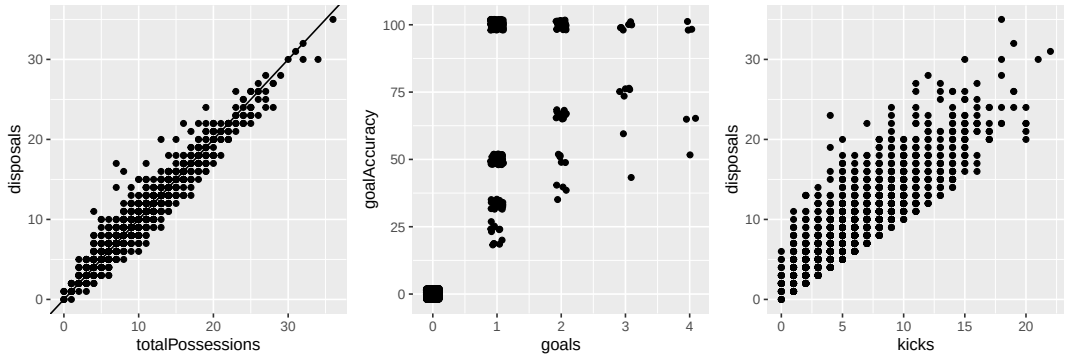


Figure 21: Scatterplots with high values on the splines scagnostic.

Splines Work

|Var1 |Var2 | splines| |:——————|:——————|——-:| |totalPossessions |disposals | 0.94| |clearances.totalClearances |clearances.stoppageClearances | 0.88| |goalAccuracy |goals | 0.83| |metresGained |kicks | 0.77| |dreamTeamPoints |disposals | 0.74| |disposals |kicks | 0.72| |dreamTeamPoints |totalPossessions | 0.72| |totalPossessions |uncontestedPossessions | 0.68| |dreamTeamPoints |kicks | 0.67| |uncontestedPossessions |disposals | 0.66|

Figure @ref(fig:aflwstatic) shows three scatterplots that score highly on the splines scagnostic. Each of these shows a relatively strong monotonic relationship between the two variables. In the interactive version of the plot, mouse over reveals some high-performing players, e.g. Anne Hatchard has a lot of possessions, disposals and kicks, and Kaitlyn Ashmore kicked 4 goals in a match with 100% accuracy.

NOTE: Each player is represented multiple times here, I think. The stats are per game. Maybe it is better to aggregate for each player and re-do the statistics?

World Bank Development Indicators

The World Bank delivers a lot of development indicators (World Bank, 2021), for many countries and multiple years. The sheer volume of indicators, in addition to substantial missing values, makes a barrier to analysis. This is a good example to where scagnostics can be used to identify pairs of indicators with interesting relationships.

Here we have downloaded indicators from 2018 for a number of countries. First, the data needs some pre-processing, to remove variables which have mostly missing values, and countries which have mostly missing values. The scagnostics will be calculated on the pairwise complete data, so it is ok to leave a few sporadic missings. At the end of the pre-processing, there are 20 indicators for 79 countries.

Summary

Appendix

AFLW Data Variable Descriptions

- *timeOnGroundPercentage*: percentage of the game the player was on the field.
- *goals*: the 6 points a team gets when the kick the ball between the two big posts.
- *behinds*: the 1 point a team gets when they kick the ball between the big post and small post.
- *kicks*: number of kicks done by the player in this game.
- *handballs*: number of handballs does by the player in the game.
- *disposals*: the number kicks and handballs a player has.
- *marks*: total number of marks in the game (the ball travels more than 15m and the player catches it without another player touching it or it hitting the ground).

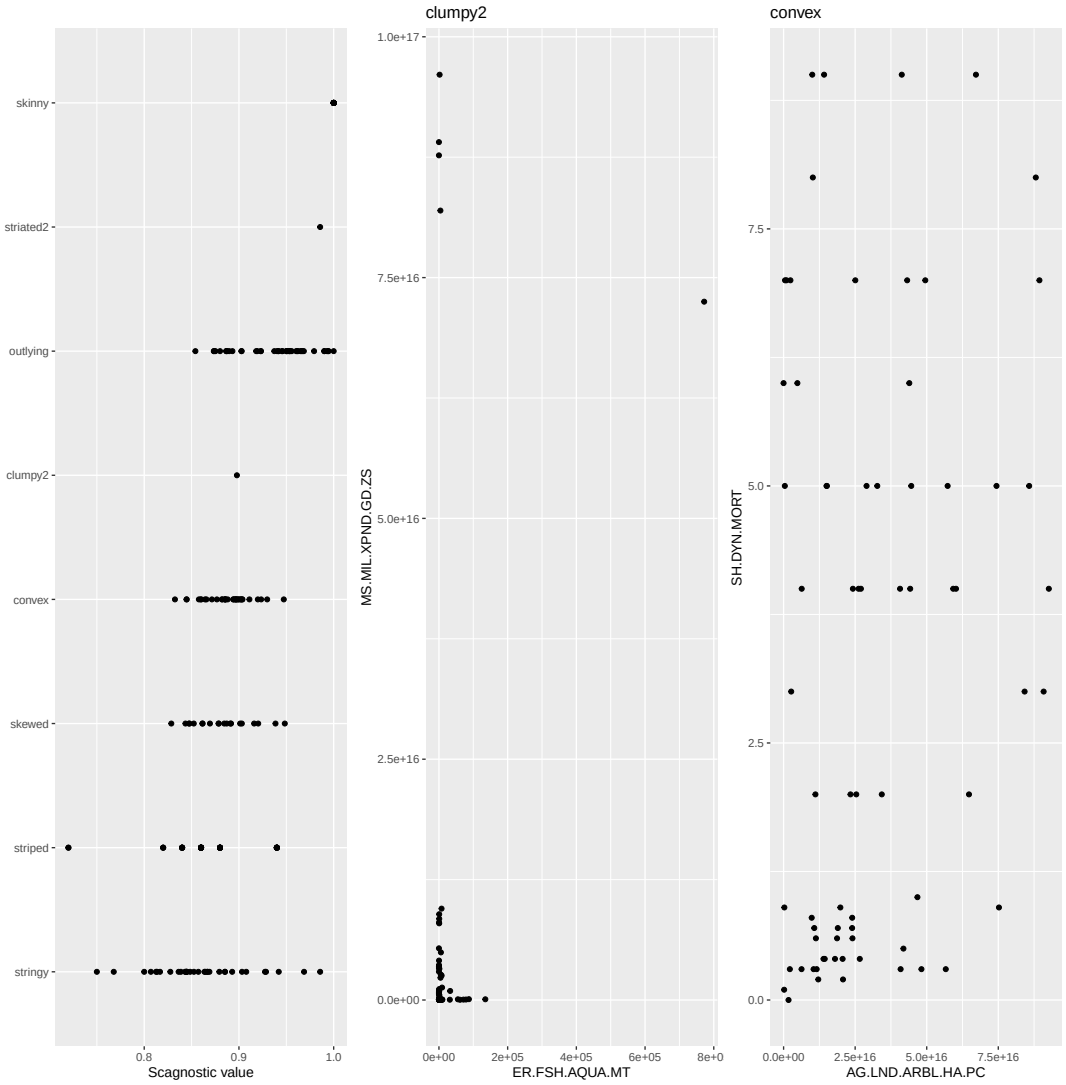


Figure 22: Most of the pairs of indicators exhibit outliersor are stringy. There is one pair that has clumpy as the highest value. There are numerous pairs that have a highest value on convex.

- *bounces*: the number of times a player bounced the ball in a game. A player must bounce the ball if they travel more than 15m and they can only bounce the ball once.

- *tackles*: Number of tackles performed by the player.

- *contestedPossessions*: the number of disposals a player has under pressure, i.e if a player is getting tackled and the get a handball or kick out of the scuffle.

- *uncontestedPossessions*: the number of disposals a player has under no pressure where they have space and time to get rid of the ball.

- *totalPossessions*: The total number of time the player has the ball.
- *inside50s*: the number of times the player has the ball within the 50m arc around the oponents goals.

- *marksInside50*: the number of marks a player gets within the 50m arc around the oponents goals.

- *contestedMarks*: the number of marks a player has under pressure.
- *hitouts*: this is how many times a player or team taps or punching the ball from a stoppage.
- *onePercenters*: all the things a player can do without registering a disposal. Eg. Spoils (punching the ball to stop someone from marking it), Shepparding (blocking for a teammate), smothering. *disposalEfficiency*: a measure of how well a player disposes of the ball. E.g. if a player kicks or handballs to the opposition a lot, they will have a low disposal efficiency percentage.

- *clangers*: this is how many times a player or team dispose of the ball and it results in a turnover to the other team.

- *freesFor*: this player was awarded a free kick.

- *freesAgainst*: this player caused a free kick to be awarded to the other team.

- *dreamTeamPoints*: this is fantasy football scoring points.

- *rebound50s*: how many times the player exits the ball out of their defence 50m arc.

- *goalAssists*: number of times the player gave the pass immediately before the player that scored a goal.
- *goalAccuracy*: percentage ratio of the number of goals kicked to the number of goal attempts.

- *turnovers*: this players disposal caused a turnover (the ball touches the ground and the other team get it).

- *intercepts*: number of times this player intercepts the disposal of the other team.
- *tacklesInside50*: number of tackles performed by this player within their defence 50m arc.

- *shotsAtGoal*: number of total shots at goal for this player (sum of goals, behinds and misses)
- *scoreInvolvements*: number of times the player was involved in a passage of play leading up to a goal.
- *metresGained*: how far a player has been able to advance the ball without turning it over.

- *clearances.centreClearances*: this is the clearance from the centre bounce after a goal or at the start of a quarter
- *clearances.stoppageClearances*: all the clearance from stoppages around the ground
- *clearances.totalClearances*: how many time a player or team clears the ball from a stoppage or from the centre

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]

Y. Wang, Z. Wang, T. Liu, M. Correll, Z. Cheng, O. Deussen, and M. Sedlmair. Improving the robustness of scagnostics. *IEEE Transactions on Visualisations and Computer Graphics*, 26(1):759–769, 2020. [p8, 10]

L. Wilkinson and G. Wills. Scagnostics distributions. *Journal of Computational and Graphical Statistics*, 17 (2):473–491, 2008. [p2, 7]

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]

World Bank. World development indicators. the world bank group. https://databank.worldbank.org/source/world-development-indicators, 2021. Accessed 14 October 2021. [p17]

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