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 data set increases, the process of visualising its structure and variable dependencies becomes more tedious. Scagnostics (scatterplot diagnostics) are a set of numerical measures describing visual features that can be used to identify interesting and abnormal scatterplots, and give a sense of priority to the variables we choose to visualise. A new set of scagnostics, containing previously defined measures and newly created measures, are implemented in the cassowaryr R package, which provides a user-friendly method to apply scagnostics to data in R. After implementing the scagnostics from the original defnitions we found that some measures did not work as expected, and variations were developed. The application of scagnostics is illustrated with four examples, sport statistics, astrophysics, collections of time series and economic indcators, that show they can be useful for exploratory data analysis, identifying shape differences between groups, and as a summary tool for the shape of multivariate data.

Introduction

Visualising our data is an important step in understanding the underlying structure and cannot be ignored. 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. These are designed to illustrate the pitfalls of numerical summaries and the importance of visualisation. Unfortunately 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. Due to limitations in visualisation, it is often difficult to completely capture relationships that involve more than two dimensions. Up to a moderate number of variables, a scatter plot matrix (SPLOM) can be used to create pairwise biplots of all variables, however, this solution quickly becomes infeasible as the number of dimensions increases and the number of plots in the SPLOM rises exponentially. There is a large amount of research into visualising high dimensional data, most of which focuses on some form of dimension reduction. Dimension reduction allows us to manage high dimensional data, usually by selecting a subset of important variables or performing a linear or non-linear transformation on the variables in a way that captures the relevant information. Unfortunately, these methods are not without their pitfalls. Linear transformations are subject to crowding, that is, projections concentrate data in the centre of the distribution making it difficult to differentiate data points (Diaconis and Freedman 1984). Non-linear transformations, such as t-distributed stochastic neighbour embedding (t-SNE) (Maaten and Hinton 2008) often have complex parameterisations, and can break the underlying global structure of the data, creating misleading visualisations. There are solutions within these methods that can somewhat mitigate these issues. To prevent crowding in a visualisation of a linear transformation, a burning sage tour from the `tourr` package proportionately zooms in more on the points in the centre of the visualisation than on those on the outskirts (Ursula Laa, Cook, and Lee 2022). To try and maintain a sense of global structure in a non-linear transformation, there are things like the `liminal` package which facilitates linked brushing between linear and non-linear transformations (Lee, Laa, and Cook 2020). These latter two methods of dimension reduction involve dynamic graphics, which is not always easy to manage or even publish.

Scagnostics are one possible alternative to variable transformations, and can be particularly useful for producing a hierarchy of variables, and capturing non-linear and unusual relationships between variables. The term scagnostics was introduced by John Tukey in 1982 (see pages 411, 427 and 433 of (ed) and Tukey (1992)). Tukey's suggestion to deal with the curse of dimensionality was to filter out uninteresting variables using a cognostic. A cognostic is a diagnostic that should be interpreted by a computer rather than a human and those specific to scatter plots are called scagnostics. Instead of trying to view every possible variable combination, the workload is reduced by calculating a series of visual features, and only presenting scatter plots that have interesting results on these feature combinations. As scagnostics filter the plots, rather than transform the data, they offer 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 select pairwise plots can leave our variable interpretations without context. Viewing pairwise plots of the data with another chart indicating their relative position in the scagnostic distributions (found by computing scagnostics on every possible pairwise plot in the data set), is one suggested solution (Dang and Wilkinson 2014a), but ultimately the lack of context remains a limitation in using scagnostics alone as a high dimensional visualisation

technique.

Scagnostics have found several applications since their initial introduction by Tukey. Dang and Wilkinson (2014b) showed scagnostics to be a valuable tool in finding hidden structures in biplots by combining them with variable transformations (such as a log transform). Dang, Anand, and Wilkinson (2013) used scagnostics to identify atypical sub-sequences from multivariate time series data for further analysis. Also, U. Laa and Cook (2020) used them in the `tourr` projection pursuit to find interesting low level projections of linear combinations of variables.

Advancing Tukey's work, L. Wilkinson, Anand, and Grossman (2005) defined computationally efficient measures that were later refined by Leland Wilkinson and Wills (2008) which make up the foundations of current scagnostics. In addition to these foundational scagnostics, Grimm (2016) introduced two additional association scagnostics. These two association measures are also used in the `tourr` projection pursuit (U. Laa and Cook 2020).

There are two scagnostics packages that compute the measures defined by L. Wilkinson, Anand, and Grossman (2005), `scagnostics` (Leland Wilkinson and Wills 2008) and the archived package `binostics` (Ursula Laa et al. 2020). Both packages are based on the original C++ code written by Leland Wilkinson and Wills (2008) which is difficult to read, debug, and extend upon. The two additional association measures discussed by Grimm (2016) were in the package `mbgraphic`, but that has also been archived by CRAN. These gaps in accessibility indicate that there is a need for a new implementation of scagnostics, one that enables a better diagnosis of results and provides graphical tools for examining these results.

The paper is organised as follows. The next section introduces the scagnostics, explains how they are calculated, and develops several new measures. This is followed by a summary of the implementation in the R package, `cassowaryr`. Finally, the example section applies scagnostics to four different data sets to illustrate their use in three different tasks. The Australian Football League Women's (AFLW) data and a simulated binary black hole (BBH) merger event show the scagnostics value as a tool in exploratory data analysis. The collection of time series example, using macroeconomic and microeconomic data, illustrates the scagnostics ability to differentiate groups by shape. The world bank indicators (WBI) shows how scagnostics can be used to summarise the shape of multivariate data.

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, Rodriguez-Casal, and. 2019) to construct an object called a `scree`. All the graph based scagnostics use the scree in their calculations, the association based scagnostics only use the raw data. The `scree` 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 1).

- **Convex hull:** The outside vertices of the graph, connected to make a convex polygon that contains all points. It is constructed using the interp package (Gebhardt, Bivand, and Sinclair 2022).
- **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 alpha hull package (Pateiro-Lopez, Rodriguez-Casal, and. 2019).
- **MST:** The minimum spanning tree (MST) is the shortest distance of branches that can be used to connect all the points. It is calculated from using the igraph package (Csardi and Nepusz 2006).

The nine scagnostics defined by Leland Wilkinson and Wills (2008) are detailed below with an explanation and a formula. They were all constructed to range [0,1], and later scagnostics have maintained this scale. To give further understanding in how these measures work, the scagnostics *skinny*, *outlying*, and *clumpy* are given an additional visual explanation in Figure 2. We will let $A =$ alpha hull, $C =$ convex hull, M = minimum spanning tree, and $s =$ the scagnostic measure.

Before any of the scagnostics are calculated, outlying points are removed. Outliers are defined as any point whose adjacent edges in the MST have edges larger $q_{75} + 1.5(q_{75} - q_{25})$, where q_i refers to the ith percentile of the sorted edge lengths of the MST.

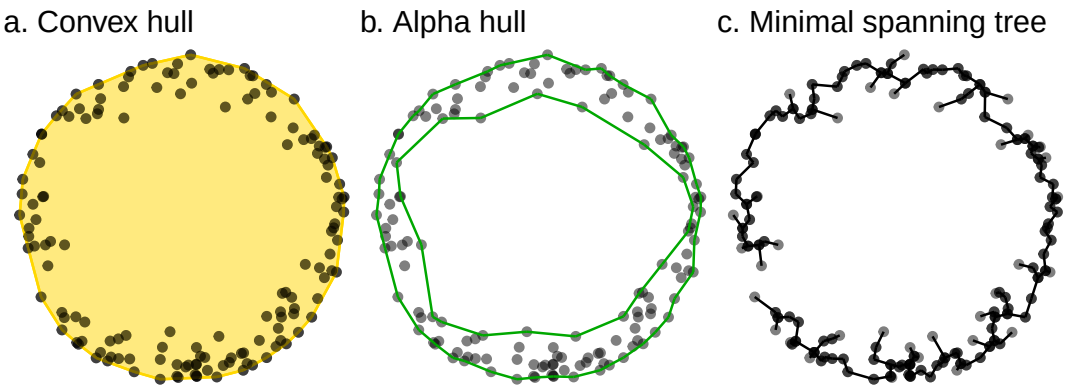

Figure 1: The building blocks for graph-based scagnostics: (a) convex hull, (b) alpha hull and (c) minimal spanning tree. The convex hull is a convex shell around all the data points. The alpha hull contains all the points but allows concavities better capturing some shapes, but it needs tuning. The minimal spanning tree connects all points once, and has a single chain connecting central points.

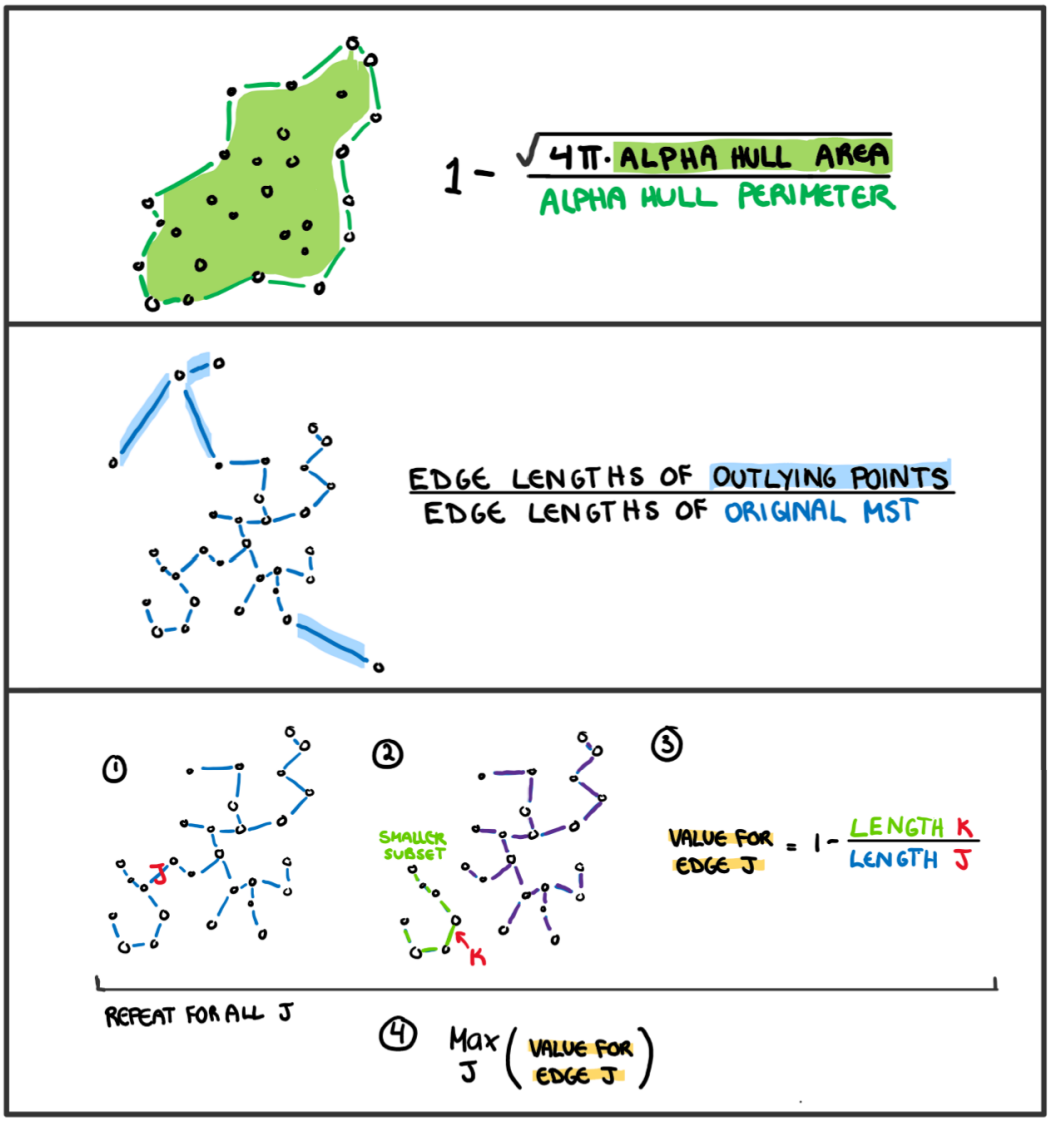

Figure 2: An visualisation of the calculation used to compute the skinny (top), outlying (middle), and clumpy (bottom) scagnostics. These three measure definitions are quite distinct, and each illustrates a unique method of capturing a visual feature in a scatter plot.

Graph-based scagnostics

- **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). Unlike the other scagnostic measures, a high value on convex does not correlate to an interesting scatter plot, rather it usually indicates a lack of relationship between the two variables.

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

- **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.

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

- **Outlying:** A measure of proportion and severity of outliers in dataset. Calculated by comparing the edge lengths of the outlying points in the MST ($M_{outliers}$) with the length of the entire MST (M).

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

- **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 90th to 50th percentile range and the central 80th interpercentile range.

$$s_{skewed} = \frac{q_{90} - q_{50}}{q_{90} - q_{10}}$$

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

$$s_{sparse} = q_{90}$$

- **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.

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

- **Striated:** This measure identifies features such as discreteness by finding parallel lines. Calculated by counting the proportion of vertices with only two edges that have an inner angle approximately between 135 and 220 degrees.

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

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 are two additional association scagnostics discussed by Grimm (2016) which are also implemented into the `cassowaryr` package.

- **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.

$$s_{splines} = \max_{i \in x,y} \left[1 - \frac{\text{Var}(\text{Residuals}_{model\ i=.})}{\text{Var}(i)} \right]$$

- **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

To test the packages ability to differentiate plots, we have created a dataset called *features* (Figure 3), that contains a series of interesting and unique scatter plots. These scatter plots each typify a certain visual feature. These include a deterministic relationship, discreteness in variables, or clustering, and scagnostics are used to order these scatter plots based on the prevalence of these various visual features.

Figure 4 shows scatter plots from the features data aligned on a 0 to 1 scale for each scagnostic. This visualisation displays a low, high, and moderate value for each scagnostic, and is useful to see how the scagnostics order data that typifies their visual feature. This plot gives us an idea of the issues some of the scagnostics face in their current state. 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 quite work as intended. The scagnostics based upon the convex hull (i.e. skinny and convex) work fine, as do the association measures such as monotonic, dcor and splines. The main issues come from the measures based on the MST. We can see in the figure that the sparse, stringy, skewed, and clumpy are each concentrated on a small portion of 0 to 1 number line. In addition to this, clumpy does not correctly order the scatter plots according to human intuition, and while it is not visible here, striated also struggles with a correct ordering. We suspect the reason for these warped distributions is the removal of binning as a preliminary step in calculating the scagnostics. The removal of binning allows for a large number of arbitrarily small edges, which upon testing was found to be the cause of a lot of these issues. A summary of how binning warped each MST scagnostic is provided in Table 1. We wanted the package to have binning as an optional method, considering choices in binning can lead to bias as noted in Leland Wilkinson and Wills (2008) or unreproducible results as noted in Wang et al. (2020). Therefore the scagnostics were assessed without binning.

While we mentioned some issues in the distributions of the scagnostics, in that they do not seem to range from 0 to 1, these measures will not be adjusted. Testing the distribution and consistency of the binned scagnostics was done previously by Leland Wilkinson and Wills (2008), however it was completed as a separate research project to the creating of the original scagnostics. Assessing the

Table 1: Summary of MST scagnostic issues.

Scagnostic	Issues
Striated	The striated measure can identify when one variable is discrete, and one is continuous, but is unable to identify two discrete variables. Additionally, since the edges used to calculate the striated measure are a subset of those used by the stringy measure, the two scagnostics are highly correlated, and often identify the same plots.
Sparse	While sparse does seem to identify spread out distributions, it rarely returns a value higher than 0.1. The removal of binning means that an infinitely large number of points can cluster in an infinsimally small space in the plot. This cluster can arbitrarily keep the sparse value low, even if the graph outside this cluster is very sparse. With a large number of observations on two continuous variables, this problem becomes unavoidable. Therefore, as the number of observations increase the sparse value approaches 0, independent of the shape of the scatterplot.
Skewed	This measure can identify skewed edge lengths, such as the L shape in the visual table, that typically require a log transformation. The skewed value rarely drop below 0.5 or rise above 0.8. Skewed seems to suffers from a similar issue to sparse reguarding binning.
Outlying	By definition an outlier must have all its adjacent edges in the MST above the outlying threshold. This means two or more observations that are close together but away from the main mass of data will not be identified as outliers, which does not align with human intuition. Even if we change the measure such that only one edge needs to be above the outlying threshold, it would only remove a single point. The measure also gives high outlying values to skewed scatter plots. If the number of points close to the centre of the cluster is large enough, outlying identifies the spread out points to be outliers and returns a large value, once again going against human intuition.
Stringy	This measure rarely drops below 0.5 even on data generated from an independent bivariate 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. That being said, it was not reported in the binned version of the scagnostics, and so is likely a result of binning.
Clumpy	With the removal of binning, clumpy does not identify a long edge connected to a short edge, but rather identifies any edge connected to an arbitrarily small edge. This means the clumpy measure rarely drops below 0.9, and also fails to order scatterplots similarly to human intuition.

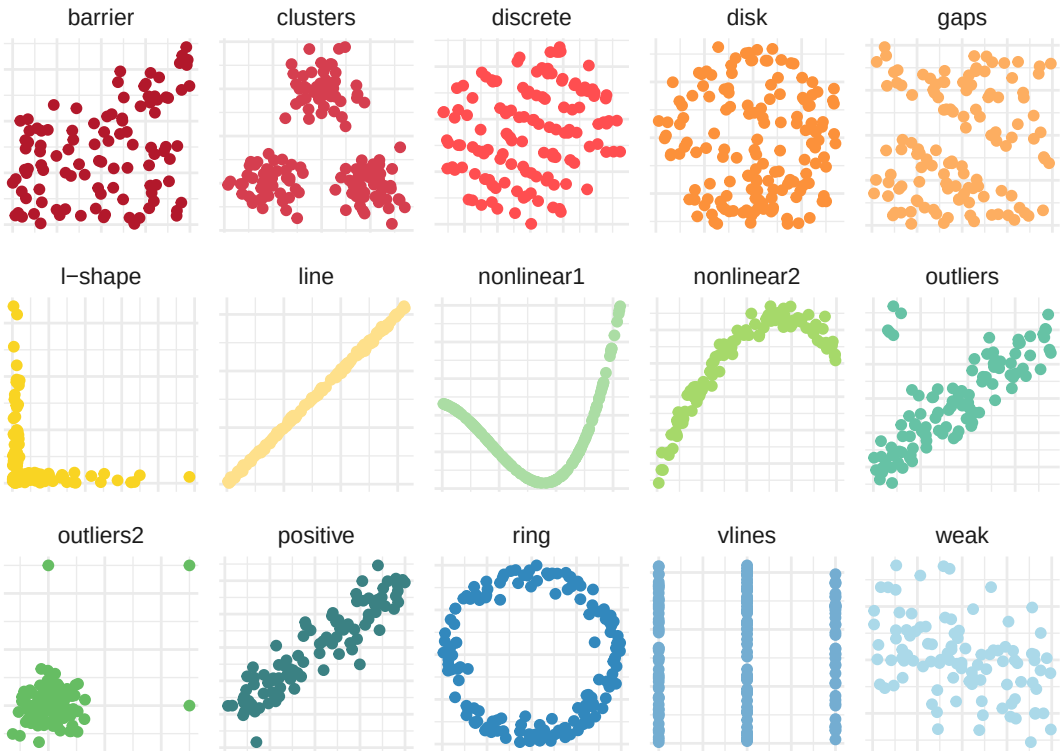

Figure 3: The scatter plots of the features dataset. These scatter plots were designed to each represent a distinct visual feature, for example the ring scatter plot is a hollow version of disk. The scagnostics need to be able to differentiate these plots.

scagnostic distributions is a considerable task, that would require checking the measures on a range of data from multiple disciplines to ensure we are finding the true "global" distributions. This task is beyond the scope of this research, and we will assume that the scagnostics range uniformly from 0 to 1 and only adjust those measures that provide an incorrect ordering.

The adjusted scagnostics measures

As the removal of binning causes the striated and clumpy measures to be in contention with human intuition, we provide adjusted versions of these scagnostics to correct this issue.

Striated adjusted The issues that need to be addressed with the new striated measure 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 measures 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 relaxed bounds around 180 degrees seems to have been to give an allowance for points moving due to binning. With the removal of binning this leeway is unnecessary and many plots that are not discrete are identified as such.

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 in Figure 5.

Figure 5 shows that while these two measures may seem similar at a glance, there are a few minor differences that make striated2 an improvement upon the original striated scagnostic. First of all, the perfect 1 value on striated goes to the *line* scatter plot. While this does fulfill the definition, it is not what the measure is supposed to be looking for, rather, it is supposed to be identifying the *vlines* scatter plot. Since striated does not count the right angles that go between the vertical lines, a truly striated plot will never get a full 1 on this measure, striated2 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. Additionally, while it is not visible in Figure 5, striated2 can identify discreteness when it appears in both axis with a small number of observation, an additional version of discreteness that the original

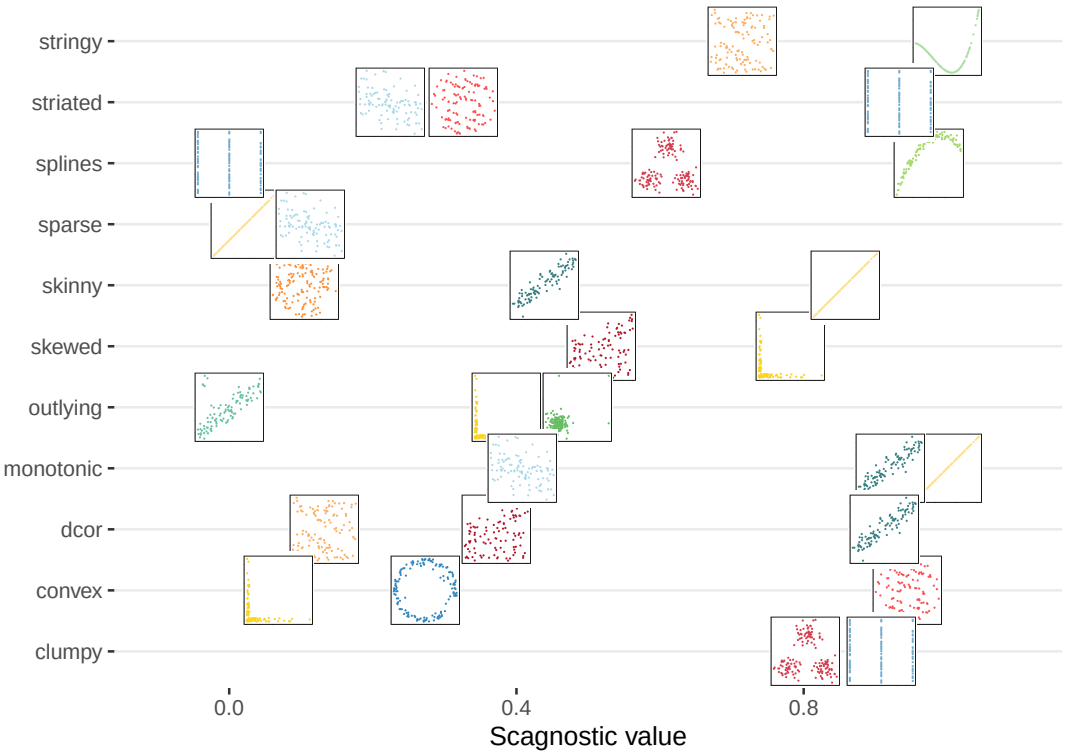

Figure 4: A visual table that displays a selection of scagnostics computed on the features data. The rows correspond to different scagnostics and the horizontal axis is the calculated value on a range of 0-1. Thumbnail plots of variable pairs are placed at their scagnostic value, and indicate the type of structure that would produce high or low or medium values. Some scagnostics, e.g. clumpy, need adjustment as they do not correctly order the scagnostics, or range from 0 to 1. Other measures, such as splines work without any changes to their definition.

Figure 5: Using a visual table to compare the striated and it's adjusted counterpart, striated2 allows us to visualise the difference in the measures. While the functions may seem similar at a glance, striated2 has a stricter version of discreteness, hence why line and vlines have the same result and plots with no discreteness score a 0.

striated struggles to identify. Both version of striated are unable to recognise the *discrete* plot, which is a noisy and rotated version of discreteness, so there is still room for improvement on this measure.

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 it 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 plots already identifies as interesting by striated)

Before creating a new clumpy measure, we looked into applying a different adjustment defined by Wang et al. (2020) that is a robust version of the original clumpy measure. 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. The robust clumpy measure builds 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 has a more uniform distribution between 0 and 1 and is more robust to outliers, however it still does a poor job of ordering plots 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, for our purposes it is not a good replacement for the clumpy scagnostic.

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

1. Sort the edges in the MST
2. Calculate the difference between adjacent edges 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 of the two clusters it is connected to and take its median edge length. The clumpy value for that edge is the ratio between the large and small edge lengths $\frac{\text{edge}_{small}}{\text{edge}_{large}}$, with a two multiplicative penalties, one for uneven clusters ($\sqrt{\frac{2\times n_{small}}{n_{small}+n_{big}}}$) and one for "stringy" scatter plots ($1 - s_{stringy}$) 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.
5. Take the mean clumpy value for each between cluster edge, if it is below 1 it is beneath the threshold that is considered clumpy, and the value is adjusted to 1.
6. clumpy2 returns $1 - \frac{1}{\text{mean}(clumpy_i)}$

With this calculation, we generate the clumpy2 measure which is compared to the original clumpy measure in the Figure 6. 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 measure does 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. Clumpy2 also has a penalty for uneven clusters (to avoid being large due to a small collection of outliers) and clusters created arbitrarily due to discreteness (such as *vlines*) in order to better align with the human interpretation of clumpy. With these changes, the stronger performance of clumpy2 is apparent in this visual table.

Implementation

Installation

The package can be installed from CRAN using

```
install.packages("cassowaryr")
```

and from GitHub using

```
remotes::install_github("numbats/cassowaryr")
```

to install the development version.

Figure 6: A visual table comparing the scagnostic values of clumpy and clumpy2. We can see the clusters plot is next to last in the ordering of the original clumpy measure, but first in clumpy2. It is clear that clumpy2 achieves a more ballanced distribution and more intuitive plot ordering.

Table 2: Cassowaryr data sets

data	explanation
features	Simulated data with special features.
anscombe_tidy	Data from Anscombes famous example in tidy format.
datasaurus_dozen	Datasaurus Dozen data in a long tidy format.
datasaurus_dozen_wide	Datasaurus Dozen Data in a wide tidy format.
numbat	A toy data set with a numbat shape hidden among noise variables.
pk	Parkinsons data from UCI machine learning archive.

Web site

More documentation of the package can be found at the web site `https://numbats.github.io/cassowaryr/`.

Data sets

The `cassowaryr` package comes with several data sets that load with the package. These are described in Table 2.

Functions

Scagnostics functions

The scagnostics functions functions either directly calculate each scagnostic measure, or are involved in the process of calculating a scagnostic measure (such as making the hull objects). These functions are low level functions, and while they are exported by the package, they are not the intended method of calculating scagnostics as they perform no outlier removal, however they are still an option for

Table 3: Cassowaryr Scagnostic functions

Function	Explanation
scree	Generates a scree object that contains the Delaunay triangulation of the scater plot.
sc_clumpy	Compute the original clumpy scagnostic measure.
sc_clumpy2	Compute adjusted clumpy scagnositc measure.
sc_clumpy_r	Compute robust clumpy scagnostic measure.
sc_convex	Compute the original convex scagnostic measure
sc_dcor	Compute the distance correlation index.
sc_monotonic	Compute the Spearman correlation.
sc_outlying	Compute the original outlying scagnostic measure.
sc_skewed	Compute the original skewed scagnostic measure.
sc_skinny	Compute the original skinny scagnostic measure.
sc_sparse	Compute the original sparse scagnostic measure.
sc_sparse2	Compute adjusted sparse measure.
sc_splines	Compute the spline based index.
sc_striated	Compute the original stirated scagnostic measure.
sc_striated2	Compute angle adjusted striated measure.
sc_stringy	Compute stringy scagnostic measure.

Table 4: Cassowaryr drawing functions

Function	Explanation
draw_alphahull	Drawing the alpha hull.
draw_convexhull	Drawing the convex hull.
draw_mst	Drawing the MST.

users if they wish. In some cases, such as `sc_clumpy_r` for clumpy robust, they are the only method to calculate that scagnostic. Table 3 outlines these functions.

Drawing functions

The drawing functions are intended to be used to better understand the results of the scagnostic functions. The input is two numeric vectors and the output is a ggplot object that draws one of the graph based objects. Table 4 details these functions

Calculate functions

The summary functions are the preferred method for users to calculate scagnostics. The `calc_scags()` function is supposed to be used on long data and takes two numerical vectors as inputs. The `calc_scags_wide()` function is designed to take in a tibble of numerical variables and return the scagnostics on every possible pairwise scatter plot. Both functions return a tibble where each column is a scagnostics. These are the two main functions of the package.

The main arguments of the `calc_scags()` function are shown in Table 5.

Table 5: The main arguments for calc_scags().

Argument	Explanation
y	numeric vector of x values.
x	numeric vector of y values.
scags	collection of strings matching names of scagnostics to calculate: outlying, stringy, striated, striated2, striped, clumpy, clumpy2, sparse, skewed, convex, skinny, monotonic, splines, dcor. The default is to calculate all scagnostics.

Table 6: Summary of three scagnostics computed by calc_scags() on the long form of the features data.

feature	outlying	clumpy2	monotonic
barrier	0.00	0.00	0.35
clusters	0.06	0.83	0.03
discrete	0.00	0.49	0.21
disk	0.00	0.00	0.01
gaps	0.00	0.75	0.06
l-shape	0.38	0.00	0.48
line	0.05	0.88	1.00
nonlinear1	0.19	0.00	0.11
nonlinear2	0.00	0.00	0.81
outliers	0.00	0.52	0.71
outliers2	0.48	0.00	0.13
positive	0.14	0.29	0.92
ring	0.02	0.83	0.01
vlines	0.00	0.00	0.03
weak	0.05	0.00	0.41

While the `calc_scags()` function does not take in a tibble, it is designed to be seamlessly integrated into the tidy data work flow. Currently to specify the scagnostics on long form tidy data the function needs to be used in conjunction with `summarise()` and `group_by()`. The code below generates the summary data in Table 6.

```
features_scags <- features %>%
    group_by(feature) %>%
    summarise(calc_scags(x,y, c("outlying", "clumpy2", "monotonic")))
```

Making summaries

There are two important summarise that should be made when calculating the scagnostics on a data set, the top pair of variables for each scagnostic, and the top scagnostic for each pair of variables. While the code required to write them is simple and easily performed by the user, having them as ready functions in the package helps guide users to use the package most effectively. The functions that perform these summaries are called `top_scags()` and `top_pairs()` respectively. The code below uses `top_scags()` to find the top pair of variables for each scagnostic.

```
top_scags(features_scags)
```

```
#> # A tibble: 3 x 3
#>   feature   scag      value
#>   <chr>     <chr>     <dbl>
#> 1 line      clumpy2   0.884
#> 2 line      monotonic 0.999
#> 3 outliers2 outlying  0.484
```

Tests

All the scagnostic functions have tests written and implemented using the `testthat` (Wickham (2011)) package. They have all been compared to calculations completed by hand to ensure the difference in results from previous literature is due to pre-processing steps such as binning, and not mistakes in the code. These tests illuminated the issues that allowed us to make meaningful changes to the definitions of clumpy and striated and understand some pitfalls of the package. For example, several tests to check the outlying scagnostic was working correctly informed us of some issues in the process of outlier removal, which is illustrated in Figure 7.

Figure 7 shows the an example of a simulated test set, combined with the associated MST. When creating this test data set, we assumed the MST would connect via the dashed red line between points

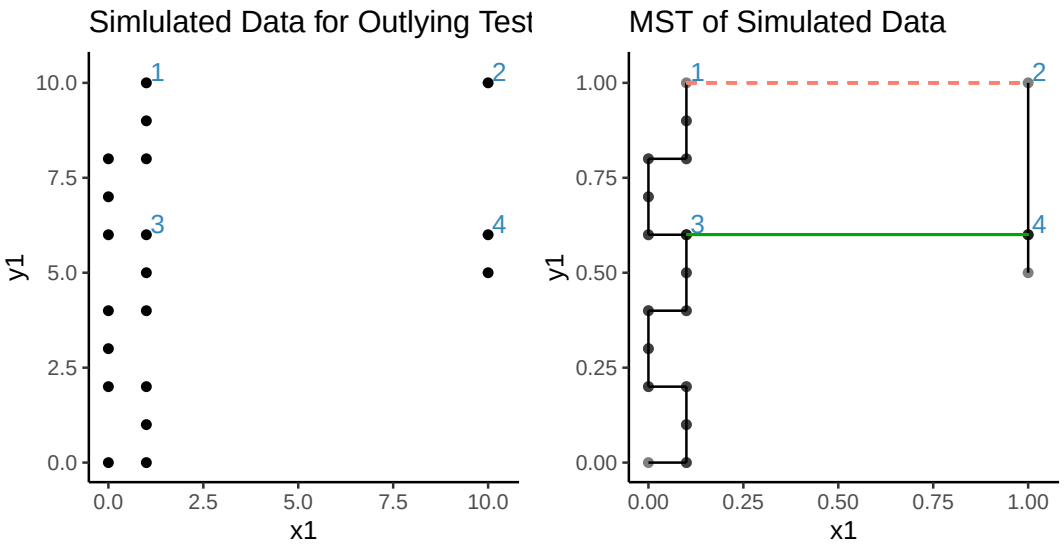

Figure 7: Plot of simulated data used for testing the outlying scagnostic. The left plot shows the raw data, while the right plot presents the MST generated on that data. When we created the test we expected the red dashed line to be in the MST, but instead the green line that connects points 3 and 4 is. If the red edge is in the MST rather than the black edge, the outlying value on this plot is much higher.

1 and 2, but instead the MST connected via the solid green line between points 3 and 4. The difference between these choices is essentially random, both edges are the exact same length, but it has significant implications for the value returned by the outlying scagnostic. When the MST contains the red dashed line between points 1 and 2, point 2 is identified as an internal outlier. This means both the red dashed edge, and the edge between points 2 and 4 are included in the outlying scagnostic calculation which evalutates to be $\frac{13}{29} = 0.45$. When we use the actual MST (with the green edge) then point 2 is still identified as an outlier, but the edge between points 2 and 4 is the only edge that goes into the outlying calculation. This results in a significantly smaller value on the the outlying scagnostic, which evaluates to be $\frac{4}{29} = 0.14$. This shows that the scagnostics are susceptible to surprisingly large changes due to arbitrary decisions made by the algorithms that construct the graph objects.

Examples

AFLW player statistics

The Australian Football League Women's (AFLW) is the national semi-professional 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. These variables are recorded per player per game, so the stats are averaged for each player over the course of the season. The description of each statistic in the data set can be found in the appendix A1. There are 68 variables, 33 of which are numeric. The the others are categorical, e.g. players names or match ids, and they would not be used in scagnostic calculations. This means there are 528 possible scatterplots, significantly more than a single person could view and analyse themselves and so we use scagnostics to identify which pairwise plots might be interesting to examine.

Figure 8 displays five scatter plots (Plots 1 to 5 in the figure) that were identified as interesting due to having a particularly high or low value on a scagnostic, or some unusual combination of two or more scagnostics. In addition to these five, there is a 6th plot (Plot 6 in the figure) that is included to display what a middling value on almost all of the scagnostics looks like. Most scatter plots score middling values on the scagnostics, so Plot 6 is also a good indication of what we would look at if we picked variables to plot ourselves with no intuition. The visual structure that changes significantly between Plots 1 to 5, and the lack of interesting visual features in Plot 6, shows the benefit of using scagnostics in the early stages of exploratory data analysis.

The best way to identify interesting scatter plots is to construct a large interactive SPLOM of the scagnostic values,each point representing a scatter plot in the data set. This is how Plots 1 to 6 were identified, but for the sake of space, we are only going to show the scatter plots of the SPLOM that led to the selection of Plots 1, 2, and 5.

Figure 9 displays Plot 1, Plot2 and Plot 5 beneath the scatter plot of the scagnostics SPLOM that

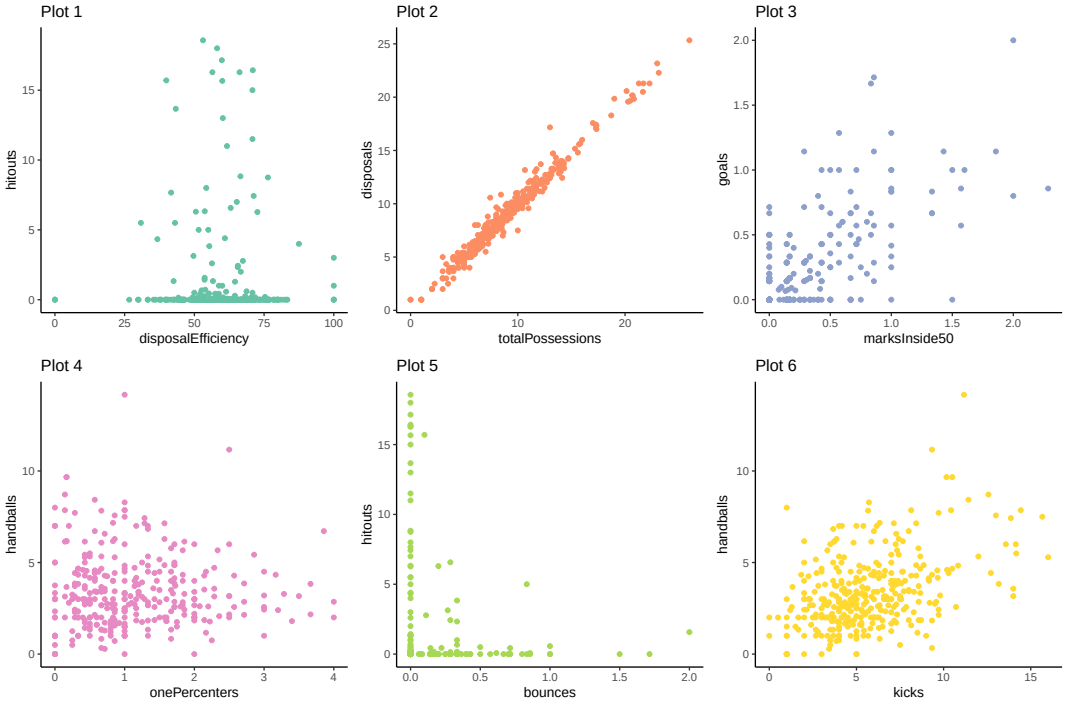

Figure 8: Six AFLW sport statistics scatter plots that were identified as identified as interesting by the scagnostics. Plots 1 to 5 had unique values on some combination of the scagnostics, Plot 6 had middling values on all measures. There is a clear difference in structure between these plots indicating the scagnostics ability to identify interesting visual features.

was used to identify it as interesting. Plot 1 was identified as interesting as it returned high values on both outlying and skewed. Intuitively, this would indicate that even after removing outliers, the data was still disproportionately spread out. This visual feature can be seen very clearly in Plot 1. 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, those that don't often have a non-linear relationship. The splines vs dcor plot tells us that there is a strong linear relationship between total possessions 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 strong linear relation indicates the level of play, and tells us few mistakes are made in a professional league. Plot 5 is an excellent example in what new information we can learn from a unique plot identified with scagnostics. This plot is high on striated2 and moderate to low on outlying, telling us most of the points will be at straight or right angles and a little spread out. Hitouts measure the number of times the player punches the ball after the referee throws it back into play, and are typically done by tall players. Bounces have to be done while running, and are typically done by fast players. The L-shape of the plot tells us that players who do one very rarely perform the other, so the tallest player in AFL is rarely the fastest. The skewed spread along both of the statistics tells us these are both specialised skills. These plots provide a clear example of the unique information gained using scagnostics as a tool in exploratory data analysis.

Non-linear shapes in black hole mergers

Physics data often contains multiple variables with highly non-linear or clustered pairwise relationships, which makes this type of data ideal for displaying the applications of splines and clumpy2; two scagnostics that's uses were not particularly visible in the AFLW example. Here we will use scagnostics to explore data that comes from a simulation of a model describing a binary black hole (BBH) merger event. The data contains 13 variables that describe the BBH event, and each point is a posterior sample that could describe the event. As the data describes a complicated physics phenomena, brief details of the variables will be left to appendix A2 as a deeper understanding requires a non-trivial amount of knowledge in physics. Therefore, we will focus on the types of patterns observed rather than an interpretation of these patterns.

The full data file contains 9998 posterior samples, and with such a large number of observations, the scagnostics cannot be computed within a reasonable time frame without the assistance of binning.

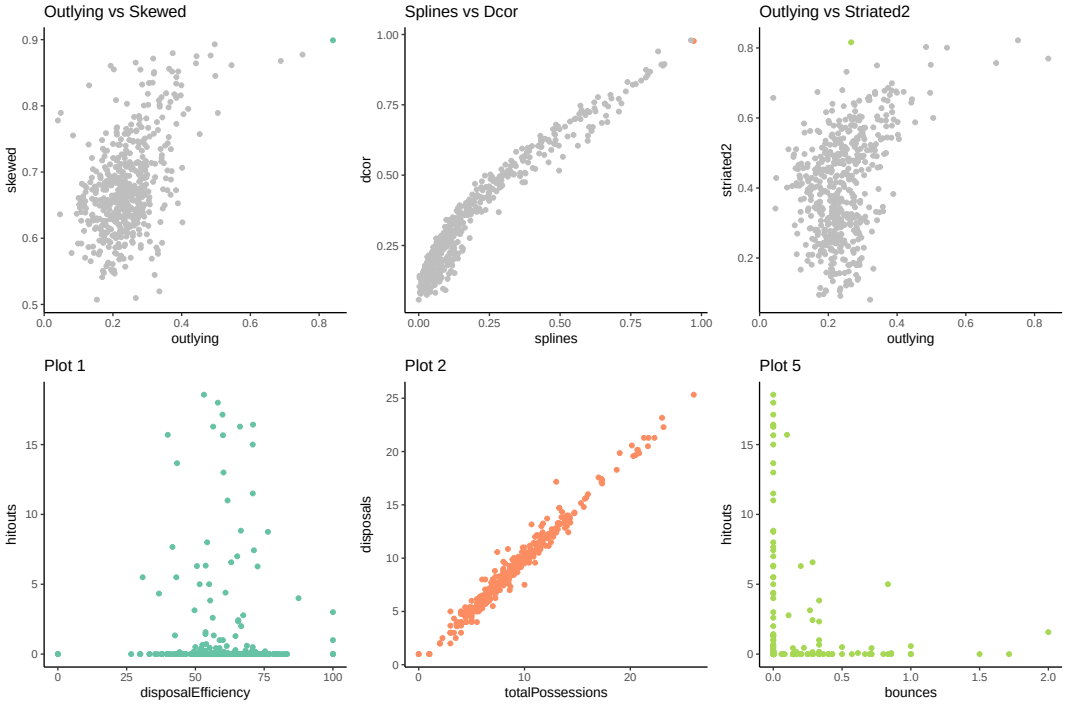

Figure 9: Three plots that were identified as interesting with the scagnostic scatter plot used to identify it. Each scatter plot of AFLW data is displayed below a plot of the two scagnostic measures it stood out on. One of the most useful ways to identify plots is through scatter plots of the scagnostics.

For our purpose a much smaller sample is sufficient, and we randomly sample 200 observations before computing the scagnostics. Since the size of the dataset is small enough, looking at the complete SPLOM of the data is still feasible, and could be used to identify several interesting scatterplots. We will omit the SPLOM here, however it allows us to see the presence of non-linear and non-functional relations between pairs of variables that we except the scagnostics to pick out. For this reason, we will focus on scatter plots that have low convex values and high skinny values, a significant difference in their splines and dcor values, or stand out on the clumpy2 measure, as these three qualities indicate the presence of non-linear or clustered relationship.

Figure 10 shows scatter plots of the relevant scagnostics measures; outlying, skinny, splines, dcor, clumpy, and clumpy2. On the left plot we see three points with very low values of the convex measure and high values of skinny. This combination of variables indicates some narrow shape and turns out to correspond to all the possible combinations containing the variables time, ra and dec. The corresponding scatterplots are shown in the upper row of Figure 11. This pattern arises because the location of an event observed from gravitational waves can only be localized when using a network

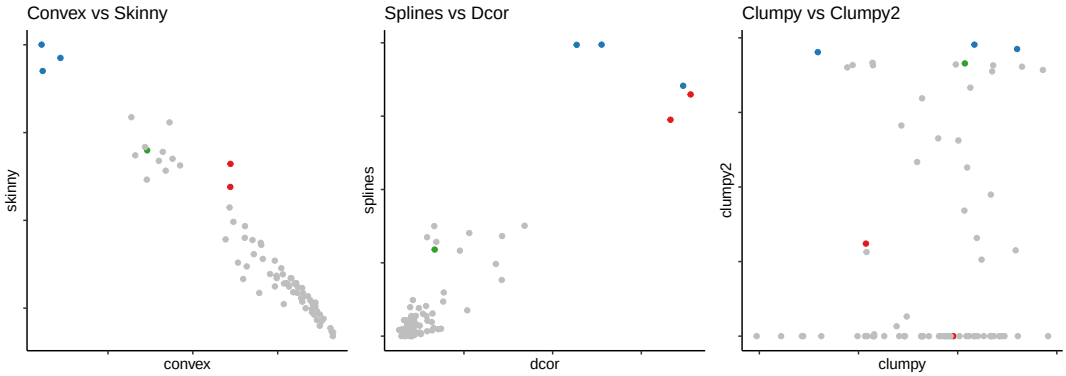

Figure 10: Selected pairs of scagnostics computed for the black hole mergers data. The coloured points in these plots align with the plot colours of Figure 11. Groups of parameter combinations can be seen to stand out in the left plot (high on skinny and low on convex) and in the middle plot (high on both dcor and splines). The plot on the right shows clumpy vs clumpy2, where we can see a large impact of the correction.

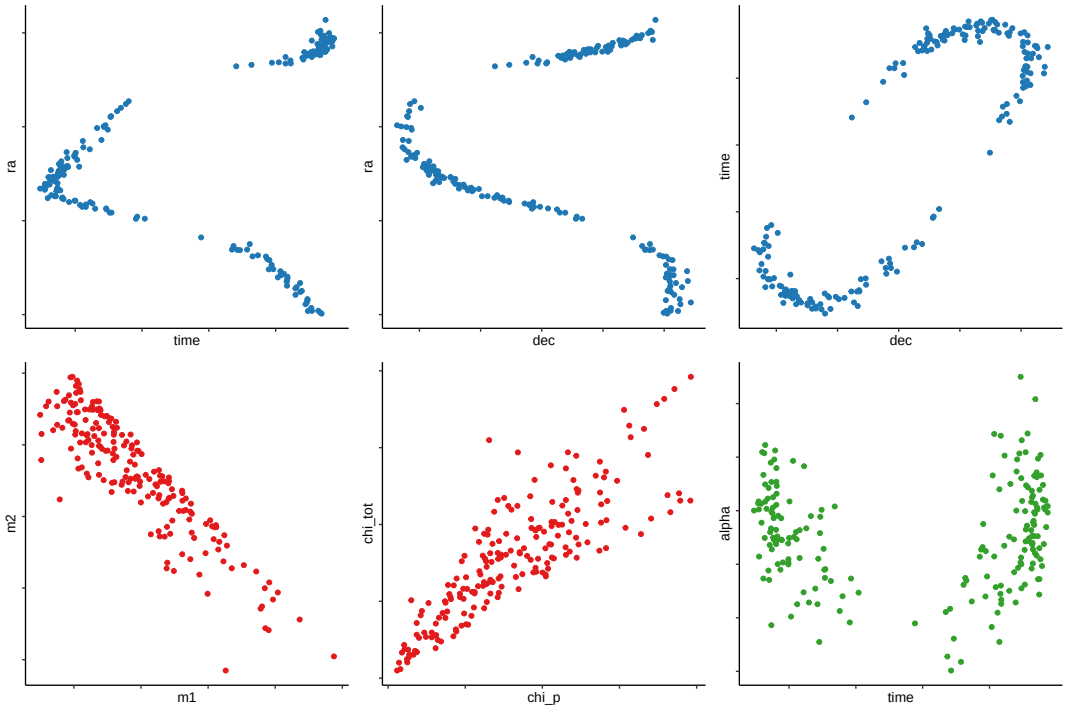

Figure 11: Features in the BBH data that stand out on several of the scagnostics measures (convey, skinny, splines and dcor), showing strong relations between variables including non-linear and non-functional dependencies. The colour of these point align with the plots position in the plots of Figure 10. The final example (time vs alpha) is expected to take high values in clumpy, but only stands out on the corrected clumpy2.

of three detectors (as described in Fairhurst (2009)), and observations with one or two detectors will result in a degeneracy between the location in the sky (parametrized by ra and dec) and the time of the event (time). It will lead to the observed pattern of a broken ring in this three dimensional space, thus inducing both non-linear dependence and clustering in the posterior sample.

These three variables also stand out in the middle plot of Figure 10, where the two plots involving ra that have a non-linear but functional relation have a somewhat higher values in the splines measure compared to dcor. On the other hand dec vs time does not exhibit a functional relation, and consequently gets a higher dcor score compared to splines (with both measures still taking large values). This also happens for two other combinations: m1 vs m2 and chi_p vs chi_tot, which are shown in the bottom row (left and middle) of Figure 11. We see that both these combinations show noisy linear relations.

Another interesting aspect of this dataset is that there are several combinations that lead to visible separations between groups of points. This makes it an ideal test case for our new implementation of clumpy2. The right plot in Figure 10 shows clumpy vs clumpy2, and reveals large differences between the two measures. In particular there are many combinations without visible clustering, that still score high on clumpy, but where clumpy2 is zero. On the other hand we can see that there are several combinations that do lead to visible separation between groups that stand out in terms of clumpy2, but not the original clumpy. One example is time vs alpha, shown in the bottom right plot of Figure 11.

Shape differences between groups

A potential application of scagnostics is to detect shape differences between groups, in contrast to the more common classification by differences in means or gaps. A difference in shape corresponds to different covariance patterns between groups. A way to think about this is contrasting linear discriminant analysis (LDA) with quadratic discriminant analysis (QDA). LDA assumes the distribution of each group is multivariate normal with the same means and the same variance. The classification boundary is planar at a maximal distance from each groups mean, with orientation of the plane determined by the pooled variance-covariance matrix. QDA similarly assumes the distribution of each group is normal, but relaxes the assumption of equal variance-covariance. The boundary between groups has a quadratic form. So from this, it can be seen that QDA is more flexible than LDA as it accommodates shape differences. However shape differences are constrained to be elliptical based on the assumption

Table 7: Pairs of time series features that have large differences between macroeconomic and microeconomic series on a range of scagnostics.

Variable 1	Variable 2	Scagnostic	Macro Value	Micro Value	Difference
acf1	trend strength	clumpy2	0.83	0.00	0.83
longest flat spot	trend strength	convex	0.12	0.62	0.50
pacf5	diff1 acf1	outlying	0.32	0.71	0.39
curvature	trend strength	skewed	0.65	0.84	0.19
longest flat spot	trend strength	skinny	0.64	0.37	0.27
acf1	trend strength	sparse	0.04	0.11	0.07
pacf5	acf1	splines	0.88	0.00	0.88
longest flat spot	diff1 acf1	striated2	0.13	0.06	0.06
diff1 acf1	trend strength	stringy	0.84	0.73	0.11

of normality. Scagnostics could be utilised to broadly identify irregular shape differences between groups.

To illustrate this analysis, we consider comparing two sets of time series, described by their features, such as trend, spikiness, acf, 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. The two groups chosen for comparison are macroeconomic and microeconomic series. The data for 100 series of each type is pulled from the self-organizing database of time-series data (Fulcher et al. 2020), using the `compenginets` R package (R. Hyndman and Yang 2021). Since the time series are different lengths, each is described by a set of time series features (chapter 4 of R. J. Hyndman and Athanasopoulos 2021) using the `feasts` R package (O'Hara-Wild, Hyndman, and Wang 2021).

To keep the illustration simple, a small set of seven features is examined. Table 7 tabulates pairs of features that have the biggest difference between groups across a range of scagnostics. Plotting a handful of these in (Figure 12), we can see the differences in shape that the scagnostics have identified. For example, the scagnostic *skewed* selects the pair of features, curvature and trend strength, and reveals a shape difference in these features between macroeconomic and microeconomic time series (middle plot). Macroeconomic series tend to have moderate curvature and varied trend, while microeconomic series tend to have strong trends and varied curvature. These feature differences that are identified in the scagnostic can be seen in the time series themselves in Figure 13. We can see from this example, and the other comparisons in the plot, that the scagnostics have identified a differences in shape that would not have been apparent from only examining mean differences between groups. Other shape differences can be read from the other pairs of features (left and right plots).

While we have shown that the scagnostics succeed in identifying difference in shapes between groups, this does not automatically transfer to a classification technique. Utilising the scagnostics ability to identify between group shape differences is an early step in using them for classification. It is not uncommon for supervised learning methods to be born from unsupervised learning methods. For example, principal component analysis (PCA), an unsupervised learning method, transforms a data set by making linear combinations of the variables in the direction of most variance. Performing this linear transformation wile also trying to explain the variance in a response variable is a supervised version of PCA known as principal component regression. However, despite its promise, developing a classification technique based on scagnostics is beyond the scope of this paper.

Processing and describing data with many variables

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

The downloaded data uses indicators from 2018 and after some pre-processing to remove variables or countries which have mostly missing values, there are 20 indicators (variables) and 79 countries in our data set. The variable names are somewhat cryptic, and their descriptions are left to the appendix A3. The scagnostics will be calculated on this pairwise complete data, allowing for a few sporadic missings.

Figure 14 (left plot) shows a summary of the top scagnostic value for each pair of variables, that is, only the highest scagnostic value for each pair of variables is saved. This is displayed as a

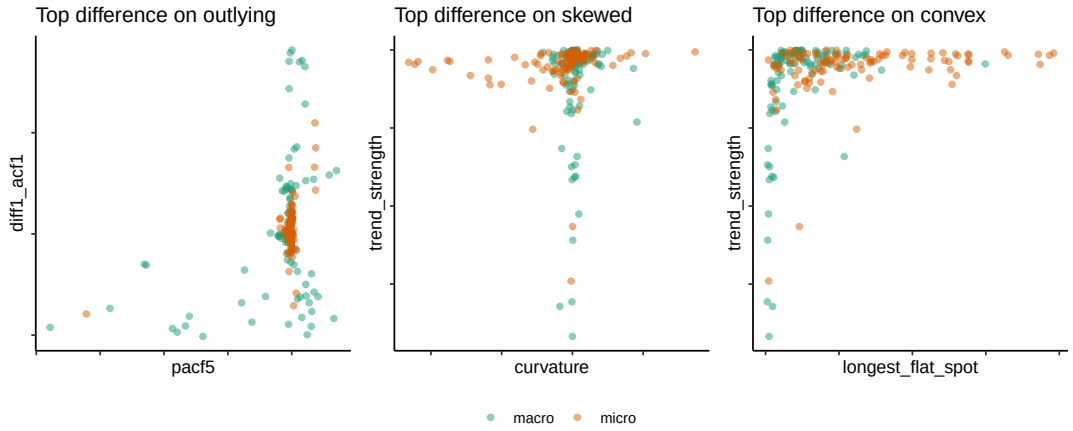

Figure 12: Interesting differences in the visual features of two groups of time series can be detected by scagnostics. The three plots represent the variable pairs that maximised the differences between outlying (left), skewed (middle) and convex (right). While these distributions may have overlapping means, their differencs in shape have been identified by the scagnostics.

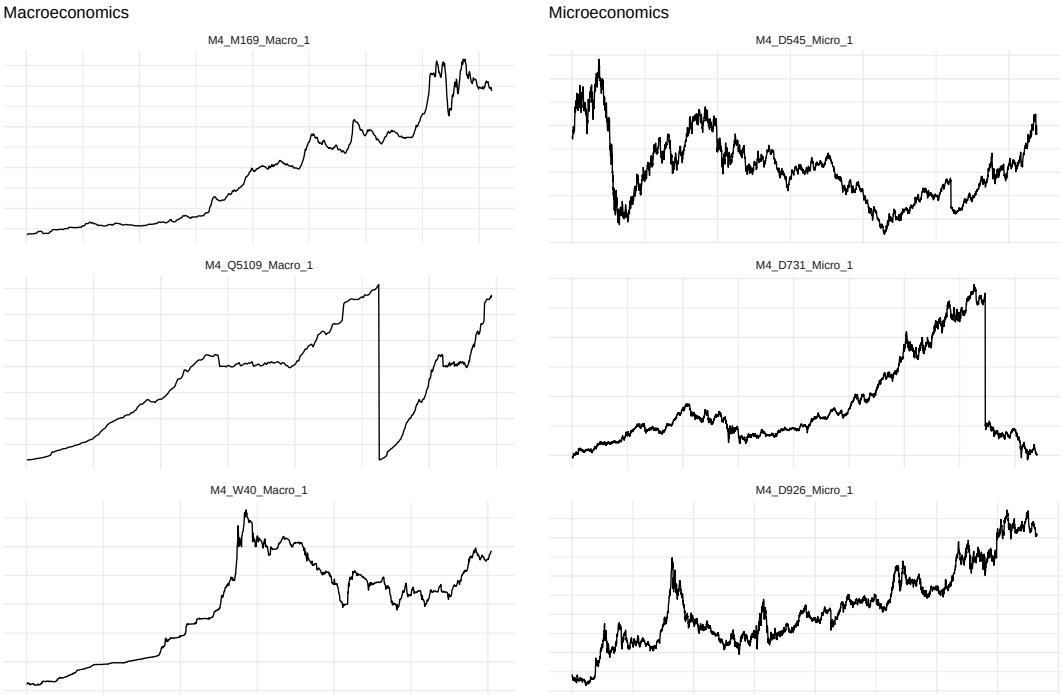

Figure 13: Selection of three series from the 100 of each of two groups, macroeconomics and microeconomics. The difference between the groups appears to be primarily in the jagginess of the two series, which a little surprisingly, is captured by the time series features curvature and trend strength. The way that trend strength is calculated, on closer inspection, could lead to describing jagginess.

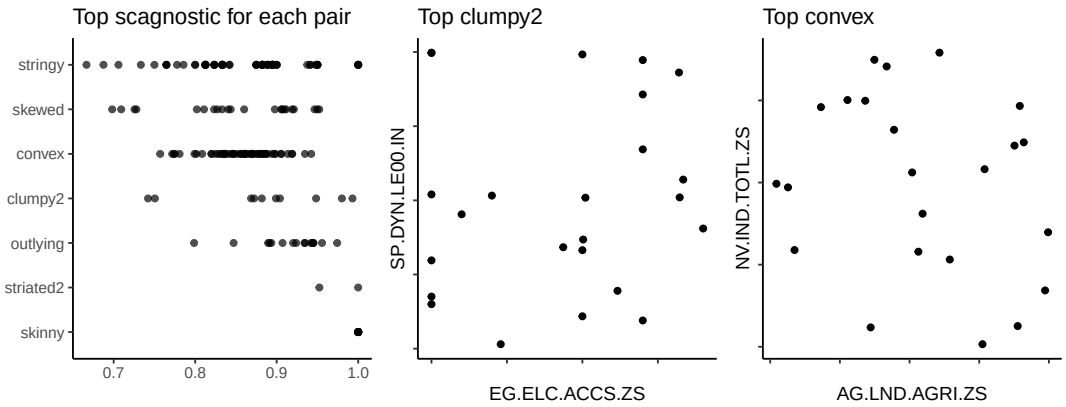

Figure 14: Using scagnostics to explore the variety of relationships present in the WBI data. The side-by-side dotplot (left) shows one point for each pair of variables, with its highest scagnostic value among all scagnostics calculated. Most of the pairs of indicators exhibit outliers, skewed, stringy or convex. There is one pair that has clumpy as the highest value. The plots, middle and right, show the pair of variables with highest value on clumpy2 and convex, respectively.

vertically-oriented side-by-side dotplot. For the WBI data, the pairs of variables are producing mostly high values on stringy, skewed, convex outlying, and clumpy2, while the scagnostics striated2, and skinny are only the highest for a single pair of variables. In addition, missing from the plot but not the calculations is that splines was not the highest for any pair of variables in the data set.

This tells us that in the WBI data, the relationships between variables is dominated by outliers (as noted by the high values on the outlying scagnostic, and to some extent also skewed and stringy), and no relationship (given bu the high values on convex). Scagnostics might be useful for obtaining alternative descriptive summaries of data with many variables. We can visualise some individual scatter plots to check this shape summary. The middle plot of Figure 14 show the pair of variables where clumpy2 has the highest value, a plot that does not stand out as "clumpy". The right plot in the figure is one of many in which convex has the highest value (a plot with no relationship). This helps us confirm what the shape summary tells us that the data, that most variable pairs have no relationships or clustering.

Conclusion

Scagnostics are a useful tool to identify the visual features in scatter plots. Building upon previous work, a new set of scagnostics, that do not require binning, have been implemented and made available in the *cassowaryr* R package. Explanations of the scagnostics and the details of the package were provided, along with four examples of use; AFLW, black hole mergers, time series and WBI. AFLW shows the general use of the *cassowaryr* package, and how to use scagnostics to find unique scatter plots. In this example, we also showed that looking at specific pairwise scatter plots can give us valuable information about a data set by interpreting the *hitouts vs bounces* scatter plot in Figure 9. Using simulated data of a black hole merger event, we then displayed the packages ability to identify pairwise non-linear relationships with convex, skinny, splines, and dcor, as well as the improvement clumpy2 made in identifying clustering. The time series example displayed how the scagnostics could be used for classification, illustrating how macro and micro economic time series can be distinguished by different shapes in feature sets. In the last example, the scagnostics applied to the WBI show that they can be used to produce an overall *shape summary* of a data set.

There are a number of different directions to build upon this research. The original scagnostics used binning as a pre-processing step. This will be useful to include to improve speed of calculations with larger sample sizes. The scale of the scagnostics is the same, ranging from 0 through 1, which allows comparison across scagnostics. However, the distribution may be different. In the WBI example, describing the overall shape of the data requires the assumption that the scagnostics are all uniformly distributed from 0 to 1. However, as some of our results suggest, and also as seen in the visual table of the features scatter plots (Figure 4), this may not be the case. It would be a substantial task to identify the distributions of the scagnostics, and make adjustments to transform to uniform distributions. The scagnostics would need to be studied using large volumes of data to capture their performance over all possible data shapes. Scagnostics could be developed further into a classification technique that identifies shape differences between groups. While the time series example shows that scagnostics can identify shape differences between groups, extending this observation to a stand alone classification

technique is outside the scope of this paper but is still a possible area of future research. Dang and Wilkinson (2014b) showed that scagnostics can be used to identify hidden structure in pairwise plots after transformations such as a log of the original data. This offers a natural extension for the scagnostics that could be added to the `cassowaryr` package. Finally, scagnostics could be used to examine scatterplots generated by a 2D projection of multiple variables, as in the projection pursuit guided tour in the `tourr` package. The primary barrier in past use has been in optimisation, because the scagnostic values over smooth sequences of projections can be noisy. This would need to be checked for the new scagnostic measures is order to use them for projection pursuit.

The implementation in R also makes it easy to expand the scagnostics collection and develop new measures. Other researchers can easily expand the package with new measures or enhance the current code because the software is openly available on GitHub, encouraging contributions from the broader community.

Acknowledgements

This article is created using **knitr** (Xie 2015) and **rmarkdown** (Xie, Allaire, and Grolemund 2018) in R. The source code for reproducing this paper can be found at: https://github.com/harriet-mason/paper-cassowaryr.

Appendix

The data used in this paper is available at https://github.com/harriet-mason/paper-cassowaryr/tree/main/data

A1: AFLW Data Dictionary

- *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).
- *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 opponents goals.
- *marksInside50*: the number of marks a player gets within the 50m arc around the opponents 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, e.g. 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

A2: Black Hole Merger Data Dictionary

- *position* in the sky is characterized by three variables: ra (right ascension), dec (declination) and distance
- *time* of the event: time
- *mass* of the two black holes: m1, m2 (with m1 > m2)
- *spin* related properties: angles alpha, theta_jn, chi_tot, chi_eff, chi_p
- *polarisation angle*: psi
- *orbital phase*: phi_jl

For a detailed description including a diagram explaining the different angles describing the spin we refer to Smith et al. (2016).

A3: World Bank Indicators Data Dictionary

- NV.AGR.TOTL.ZS Agriculture, value added (% of GDP)
- EN.ATM.CO2E.PC CO2 emissions (metric tons per capita)
- NE.EXP.GNFS.ZS Exports of goods and services (% of GDP)
- DT.DOD.DECT.CD External debt stocks, total (DOD, current US$)
- SP.DYN.TFRT.IN Fertility rate, total (births per woman)
- NY.GDP.MKTP.CD GDP (current US$)
- NY.GNP.MKTP.PP.CD GNI, PPP (current international $) TX.VAL.TECH.MF.ZS High-technology exports (% of manufactured exports)
- SH.IMM.MEAS Immunization, measles (% of children ages 12-23 months)
- NE.IMP.GNFS.ZS Imports of goods and services (% of GDP)
- NV.IND.TOTL.ZS Industry, value added (% of GDP)
- SP.DYN.LE00.IN Life expectancy at birth, total (years)
- TG.VAL.TOTL.GD.ZS Merchandise trade (% of GDP)
- MS.MIL.XPND.GD.ZS Use and distribution of these data are subject to Stockholm International Peace Research Institute (SIPRI) terms and conditions. Military expenditure (% of GDP)
- IT.CEL.SETS.P2 Mobile cellular subscriptions (per 100 people)
- SH.DYN.MORT Mortality rate, under-5 (per 1,000 live births)
- SP.POP.TOTL Population, total
- SH.DYN.AIDS.ZS Prevalence of HIV, total (% of population ages 15-49)
- GC.TAX.TOTL.GD.ZS Tax revenue (% of GDP)
- EG.ELC.ACCS.ZS Access to electricity (% of population)
- NY.ADJ.NNTY.CD Adjusted net national income (current US$)
- SH.HIV.INCD.TL Adults (ages 15+) and children (ages 0-14) newly infected with HIV
- AG.LND.AGRI.ZS Agricultural land (% of land area)
- ER.FSH.AQUA.MT Aquaculture production (metric tons)
- AG.LND.ARBL.HA.PC Arable land (hectares per person)
- MS.MIL.TOTL.TF.ZS Armed forces personnel (% of total labor force)
- FB.BNK.CAPA.ZS Bank capital to assets ratio (%)
- SE.COM.DURS Compulsory education, duration (years)

References

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. https://www.tandfonline.com/doi/abs/10.1080/00031305.1973.10478966.

Csardi, Gabor, and Tamas Nepusz. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal* Complex Systems: 1695. https://igraph.org.

Dang, Tuan Nhon, Anushka Anand, and Leland Wilkinson. 2013. "TimeSeer: Scagnostics for High-Dimensional Time Series." *IEEE Transactions on Visualization and Computer Graphics* 19 (3): 470–83. https://doi.org/10.1109/TVCG.2012.128.

Dang, Tuan Nhon, and Leland Wilkinson. 2014a. "ScagExplorer: Exploring Scatterplots by Their Scagnostics." In *2014 IEEE Pacific Visualization Symposium*, 73–80. https://doi.org/10.1109/PacificVis.2014.42.

———. 2014b. "Transforming Scagnostics to Reveal Hidden Features." *IEEE Transactions on Visualization and Computer Graphics* 20 (12): 1624–32.

Diaconis, Persi, and David Freedman. 1984. "Asymptotics of Graphical Projection Pursuit." *The Annals of Statistics* 12 (3): 793–815. http://www.jstor.org/stable/2240961.

(ed), D. R. Cox, and John Wilder Tukey. 1992. *The Collected Works of John w. Tukey*. Chapman; Hall/CRC.

Fairhurst, Stephen. 2009. "Triangulation of Gravitational Wave Sources with a Network of Detectors." *New Journal of Physics* 11 (12): 123006. https://doi.org/10.1088/1367-2630/11/12/123006.

Fulcher, Ben D, Carl H Lubba, Sarab S Sethi, and Nick S Jones. 2020. "A Self-Organizing, Living Library of Time-Series Data." *Scientific Data* 7 (1): 213–13. https://www.comp-engine.org.

Gebhardt, Albrecht, Roger Bivand, and David Sinclair. 2022. *Interp: Interpolation Methods*. https://CRAN.R-project.org/package=interp.

Grimm, Katrin. 2016. "Kennzahlenbasierte Grafikauswahl." Doctoral thesis, Universität Augsburg.

Hyndman, R. J., and G. Athanasopoulos. 2021. *Forecasting: Principles and Practice, 3rd Edition*. OTexts: Melbourne, Australia, OTexts.com/fpp3.

Hyndman, Rob, and Yangzhuoran Yang. 2021. *Compenginets: Time Series from Http://Www.comp-Engine.org/Timeseries/*. https://github.com/robjhyndman/compenginets.

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

Laa, Ursula, Dianne Cook, and Stuart Lee. 2022. "Burning Sage: Reversing the Curse of Dimensionality in the Visualization of High-Dimensional Data." *Journal of Computational and Graphical Statistics* 31 (1): 40–49. https://doi.org/10.1080/10618600.2021.1963264.

Laa, Ursula, Hadley Wickham, Dianne Cook, and Heike Hofmann. 2020. "Binostics: Computing Scagnostics Measures in r and c++." https://github.com/uschiLaa/paper-binostics.

Lee, Stuart, Ursula Laa, and Dianne Cook. 2020. "Casting Multiple Shadows: High-Dimensional Interactive Data Visualisation with Tours and Embeddings." https://arxiv.org/abs/2012.06077.

Locke, Steph, and Lucy D'Agostino McGowan. 2018. *datasauRus: Datasets from the Datasaurus Dozen*. https://CRAN.R-project.org/package=datasauRus.

Maaten, L. J. P van der, and G. E Hinton. 2008. "Visualizing High-Dimensional Data Using t-SNE." *Journal of Machine Learning Research* 9 (nov): 2579–2605.

O'Hara-Wild, Mitchell, Rob Hyndman, and Earo Wang. 2021. *Feasts: Feature Extraction and Statistics for Time Series*. https://CRAN.R-project.org/package=feasts.

Pateiro-Lopez, Beatriz, Alberto Rodriguez-Casal, and. 2019. *Alphahull: Generalization of the Convex Hull of a Sample of Points in the Plane*. https://CRAN.R-project.org/package=alphahull.

Smith, Rory, Scott E. Field, Kent Blackburn, Carl-Johan Haster, Michael Pürrer, Vivien Raymond, and Patricia Schmidt. 2016. "Fast and Accurate Inference on Gravitational Waves from Precessing Compact Binaries." *Phys. Rev. D* 94 (August): 044031. https://doi.org/10.1103/PhysRevD.94.044031.

Wang, Yunhai, Zeyu Wang, Tingting Liu, Michael Correll, Zhanglin Cheng, Oliver Deussen, and Michael Sedlmair. 2020. "Improving the Robustness of Scagnostics." *IEEE Transactions on Visualisations and Computer Graphics* 26 (1): 759–69.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

Wilkinson, L., A. Anand, and R. Grossman. 2005. "Graph-Theoretic Scagnostics." In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, 157–64.

Wilkinson, Leland, and Graham Wills. 2008. "Scagnostics Distributions." *Journal of Computational and Graphical Statistics* 17 (2): 473–91.

World Bank. 2021. "World Development Indicators. The World Bank Group." https://databank.worldbank.org/source/world-development-indicators.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.name/knitr/.

Xie, Yihui, Joseph J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Chapman;

Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Harriet Mason
Monash University
Department of Econometrics and Business Statistics
Melbourne, Australia
https://harrietmason.netlify.app/
ORCiD: 0009-0007-4568-8215
hmas0003@student.monash.edu

Stuart Lee
University of Melbourne
Melbourne Data Analytics Platform
Melbourne, Australia
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: 0000-0002-3813-7155
dicook@monash.edu