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 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, by their original definition or with some adjustments. The scagnostics are shown to correctly order scatter plots with known interesting visual features. Following that the fine details of the package, including function explanations and tests are discussed. Finally the package is applied to data from the Australian Football League Women's (AFLW) and a simulated binary black hole (BBH) merger event to show its value as a tep in exploratory data analysis; macroeconomic and microeconomic data to show its ability to differentiate groups by shape; and finally the world bank indicators (WBI) to show how the package can be used to summarise the shape of an entire dataset.

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. Due to limitations in visualisation, it is often difficult to completely capture relationships that involve more than two dimensions. Therefore data sets that involve a large number of variables are difficult to visualise because the number of possible pairwise plots rises exponentially with the number of dimensions. Despite this difficulty, the visualisation process cannot be skipped. 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's suggestion to deal with the curse of dimensionality is to filter out uninteresting visualisations 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. 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 in feasible as the number of dimensions increases. 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 rather than a filtering of plots as suggested by Tukey. These methods reduce the dimensionality by creating a hierarchy of the variables and taking a subset,or 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. 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 of than those of the outskirts (Laa et al., 2020a). 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 et al., 2020). Unfortunately these methods of dimension reduction still involve some transformation of the data, and thus will still somewhat warp perception. 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. Visualising the pairwise plots in relation to their place in the data set's global scagnostic distribution, is one suggested solution (Dang and Wilkinson, 2014b), but ultimately the lack of context remains one of the limitations of using scagnostics alone as a high

dimensional visualisation technique.

Scagnostics have found a reasonably large number of applications since their initial introduction by Tukey. Laa and Cook (2020) used them in the `tourr` projection persuit to find interesting low level projections of linear combinations of variables. Dang and Wilkinson (2014a) showed scagnostics to be a valuable tool in finding hidden structures in biplots by combinging them with variable transformations (such as a log transform). Dang et al. (2013) used scagnostics to identify atypical sub-sequences from multivariate time series data for further analysis. These are only a small handful of examples of the ways in which scagnostics have been used to assist in th visualisation of high dimensional data.

Advancing Tukey's work, Wilkinson et al. (2005) defined computationally efficient measures that were later refined by Wilkinson and Wills (2008) which make up the foundations of the measures considered to be scagnostics. They were all constructed to range [0,1], and later scagnostics have maintained this scale. In addition to these foundational scagnostics, Grimm (2016) discussed the benefit of using two additional association scagnostics. 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 packages are based on the original C++ code written by 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 of the `cassowaryr` package; and finally several examples using sports, physics, time series, and world bank indicators data illustrate the usage of the package.

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`. 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 @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 using the tripack package (code by R. J. Renka. R functions by Albrecht Gebhardt. With contributions from Stephen Eglen <stephen@anc.ed.ac.uk> et al., 2020).
- **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 (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).

Before any of the scagnostics are calculated, outlying points are removed. Outerliers are defined as any point whose adjacent edges in the MST have edges larger than ω:

$$\omega = q_{75} + 1.5(q_{75} - q_{25})$$

Where q_i refers to the ith percentile of the sorted edge lengths of the MST.

The nine scagnostics defined by Wilkinson and Wills (2008) are detailed below with an explanation and a formula. To give further understanding in how these measures work, the scagnostics *skinny*, *outlying*, and *clumpy* are given an additional visual explanation in Figure @ref(fig:scagdrawn). We will let A = alpha Hull, C = convex hull, M = minimum spanning tree, and s = the scagnostic measure. Since some of the measures have a sample size dependence, w is a parameter used to adjust for the sample size.

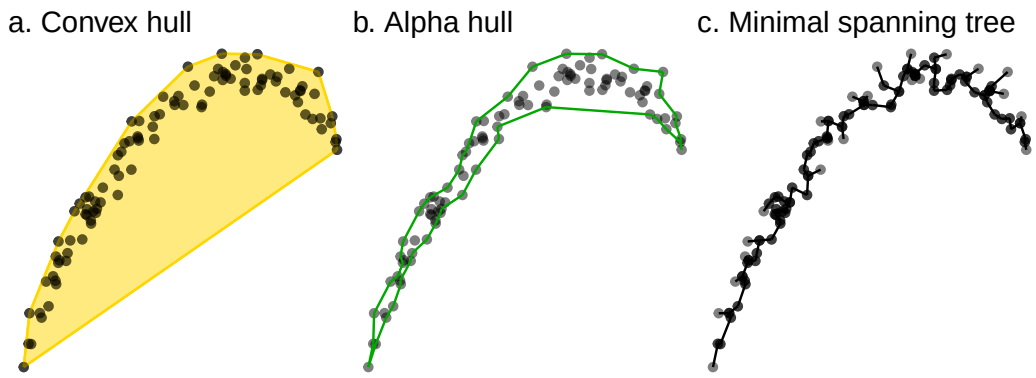

Figure 1: The building blocks for graph-based scagnostics: (a) convex hull, (b) alpha hull and (c) minimal spanning tree. The convex hull is a convex shell around all the data points. The alphahull contains all the points but allows concavities better capturing some shapes, but it needs tuning. The minimal spanning tree connects all points once, and has a single chain connecting central points. (#fig:building-blocks2)

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.

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

- **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 ($V^{(1)}$).

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

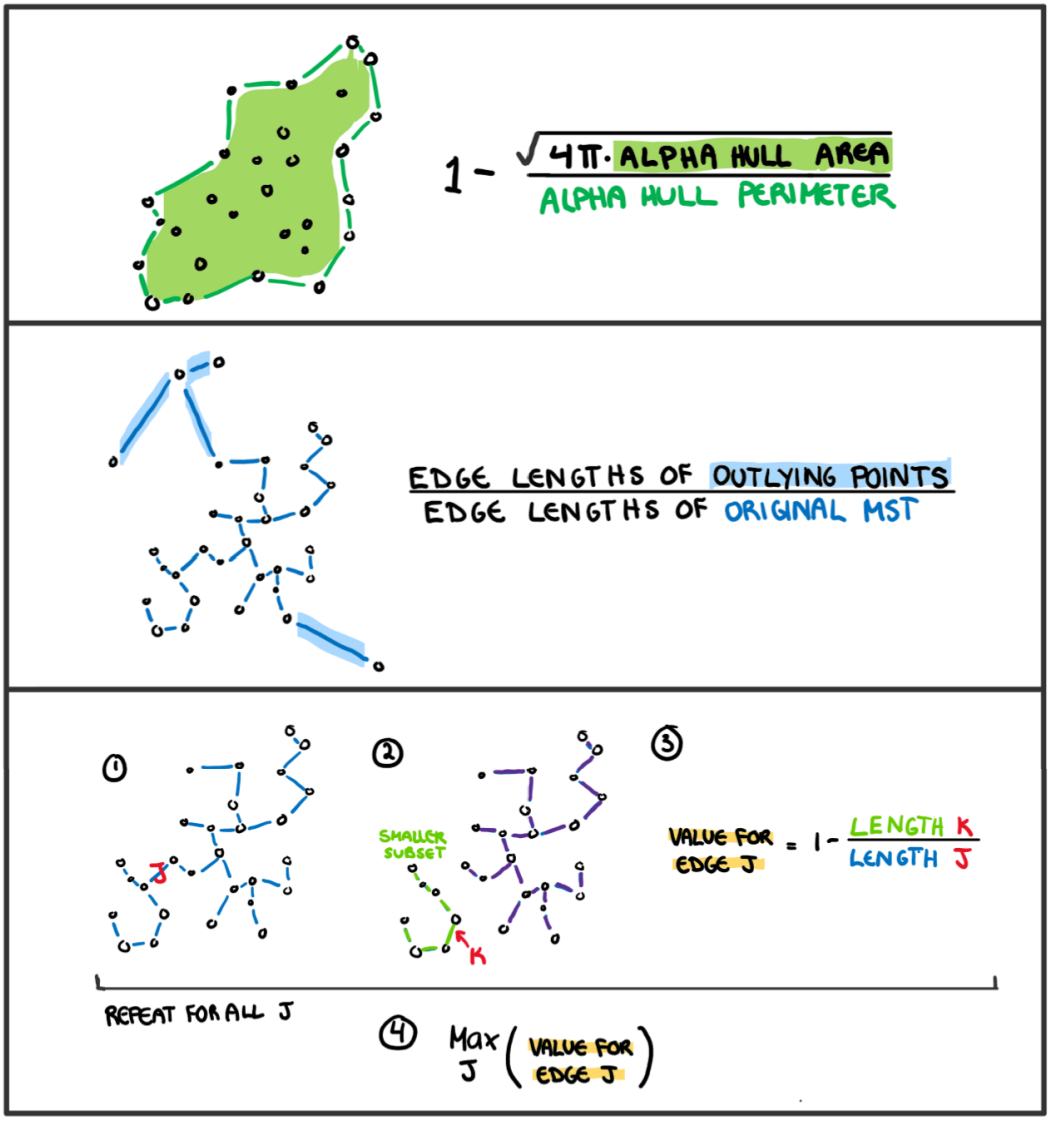

Figure 2: An visualisation of the calculation used compute the skinny (top), outlying (middle), and clumpy (bottom) scagnostics. The measure definitions are all distinct, and each illustrates a unique method of capturing a visual feature of a scatter plot.
(#fig:scagdrawn)

- **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 central 80th interpercentile range and the central 40th interpercentile range.

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

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

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

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

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

- **Striated:** This measure identifies features such as discreteness by finding parallel lines. 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 impleneted 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.

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

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

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

where

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

where

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

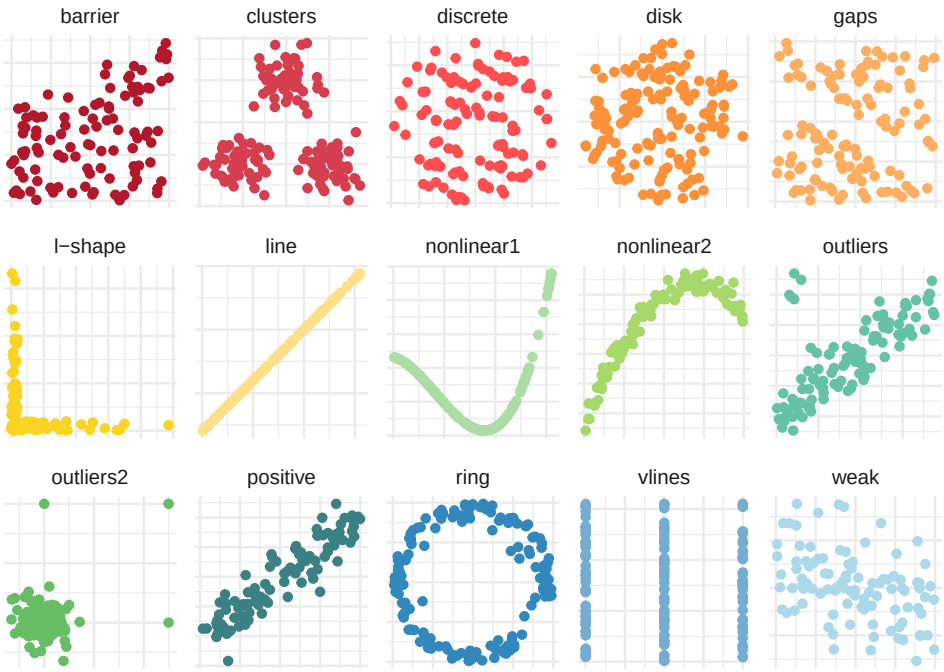

Figure 3: The scatter plots of the features dataset. These scatter plots were designed to each represent a distinct visual feature, for example the ring scatter plot is a hollow version of disk. The scagnostics need to be able to differentiate these plots.
(#fig:features-plot)

Checking the scagnostics calculations

To test the packages ability to differentiate plots, we have creates a dataset called *features* (Figure @ref(fig:features-plot)), that contains a series of interesting and unique scatter plots. These scatter plots each typify a certain visual feature. Be it a deterministic relationship, discreteness in variables, or clustering, we should be able to use scagnostics to order these scatter plots based on the prevalence of these visual features.

Figure @ref(fig:visual-table) 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 montonic, 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 @ref(tab:scagissues-tb-pdf). We wanted the package to have binning as an optional method, considering choices in binning can lead to bias as noted in Wilkinson and Wills (2008) or unreproducible results as noted in Wang et al. (2020). Therefore the scagnostics were assessed without binning.

Several of the measures that do not have a uniform distrubution from 0 to 1 still correctly order the scatter plots. In order to truely assess the distribution of these functions, we would need to check the scagnostics on a large range of data from muliple diciplines before claiming that the distribution is truely warped by the removal of binning. Intuition on a small simulated features data set is not enough. Testing on the distribution and consistency of the binned scagnostics was done previously by Wilkinson and Wills (2008), however it was completed as a seperate research project to the creating of the original scagnostics. This task that is beyond the scope of this research, and so we will assume that

Table 1: (#tab:scagissues-tb-pdf)Summary of MST scagnostic issues.

Scagnostic	Issues
Striated	The striated measure can identify the specific case of one discrete variable and one continuous variable but cannot identify two discrete variables. Since by definition it is a subset of the stringy measure, they are highly correlated, and often plots that striated identifies as interesting have already been identified by stringy.
Sparse	While sparse does seem to identify spread out distributions, it rarely returns a value higher than 0.1. The removal of binning means the number of values that can cluster on one portion of the plane is infinite. Even if the rest of the scatter plot is sparse, this one cluster will arbitrarily keep the sparse value low. With a large number of observations on two continuous variables, this is unavoidable, which also means the measure does not have consistancy.
Skewed	This measure can identify skewed edge lengths, such as the L shape in the visual table, however its 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 struggles with distributions that have an increasing variance due to the removal of binning. 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 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. 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 does not correctly order the edges.

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.
(#fig:visual-table)

the scagnsotics range uniformly from 0 to 1 and only adjust those measures that provide an incorrect ordering. Therefore, in the section below we will discuss the adjustments made to the striated and clumpy scagnostics.

The adjusted scagnostics measures

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 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 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 @ref(fig:striated-vtable).

Figure @ref(fig:striated-vtable) shows that while these two measures may seem similar at a glace, 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, itis 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 @ref(fig:striated-vtable), striated2 can identify discreteness when it appears in both axis with a small number of observation, an additional version of discreteness that the original striated struggles to identify. Both version of

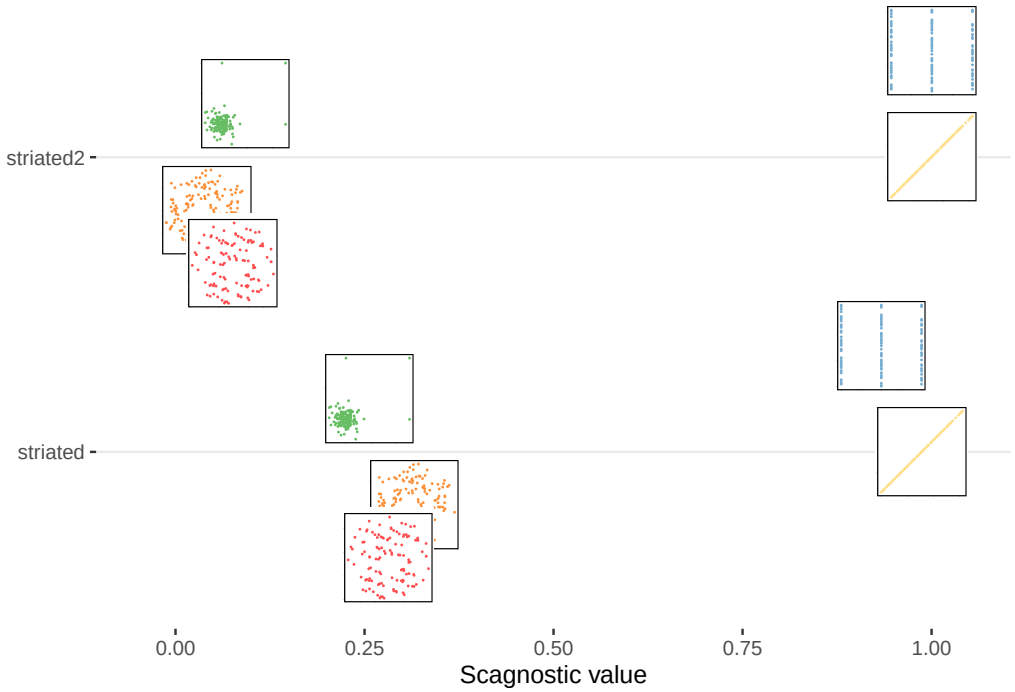

Figure 5: Using a visual table to compare the striated and it's adjusted counterpart, striated2 allows us to visualise the difference in the measures. While the functions may seem similar at a glance, striated2 has a stricter version of discreteness, hence why line and vlines have the same result and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

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, 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, 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

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. (#fig:clumpy-vtable)

edge lengths $\frac{edge_{small}}{edge_{large}}$, with a two multiplicative penaltys, 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 consdiered clumpy, and the value is adjusted to 1.
6. clumpy2 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 @ref(fig:clumpy-vtable). 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 colelction 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.

Table 2: (#tab:datasets-tb-pdf)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.

Table 3: (#tab:scagfuncs-tb-pdf)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.

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, they are described in Table @ref(tab:datasets-tb-pdf).

Functions

Scagnostics functions

The scagnostics functions functions either directly calculate each scagnostic measure, or are involved in the process of calcuating a scanostic 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 calcuating scagnostics as they perform no outlier removal, however they are still an option for 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 @ref(tab:scagfuncs-tb-pdf) outlines these functions.

Table 4: (#tab:drawfuncs-tb-pdf)Cassowaryr drawing functions

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

Table 5: (#tab:calcfunc-tb-pdf)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.

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 @ref(tab:drawfuncs-tb-pdf) details these functions

Calculate functions

The summary functions are the perferred 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 @ref(tab:calcfunc-tb-pdf).

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 function computes all the scagnostics, and the usercan choose those of interest, using `select()`. The reason we need to use select is that the *scags* argument for `calc_scags()` is not currently recognised inside `summarise()`. The code below generates the summary data in Table @ref(tab:featuresscags-pdf).

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

Making summaries

There are two important summarise that should be made when calculating the scagnostics on a dataset, the top pair of variables for each scagnostic, and the top scagnostic for each pair of variables. The code for both are simple, but an example of how to calculate them on the long features data, alongside their output will be shown here.

To calculate the top pair of variables for each scagnostic, we would use the code below.

```
features_toppairs <- features_scags %>%
  pivot_longer(!feature, names_to = "scag", values_to = "value") %>%
  arrange(desc(value)) %>%
  group_by(scag) %>%
  slice_head(n=1)
```

To calculate the top scagnostic for each pair of variables, we would use the code below.

Table 6: (#tab:featuresscags-pdf)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	0.00	0.01
disk	0.02	0.40	0.09
gaps	0.00	0.75	0.06
l-shape	0.38	0.00	0.48
line	0.11	0.00	1.00
nonlinear1	0.27	0.00	0.17
nonlinear2	0.00	0.00	0.81
outliers	0.00	0.52	0.71
outliers2	0.59	0.00	0.06
positive	0.14	0.29	0.92
ring	0.02	0.45	0.04
vlines	0.00	0.17	0.08
weak	0.05	0.00	0.41

```
features_toppairs <- features_scags %>%
  pivot_longer(!feature, names_to = "scag", values_to = "value") %>%
  arrange(desc(value)) %>%
  group_by(scag) %>%
  slice_head(n=1)
```

While the code required to write them is simple and easily performed by the user, having them as ready functions in the package would help guide users to use the package most effectively. These functions would be called `top_scags()` and `top_pairs()` .

Tests

All the functions that calculate the scagnostic measures (all the measures that start with "sc") have tests written and implemented using the testthat package. They have all been compared to calculations completed by hand to ensure the difference in results from previous literature are due to other steps in the process, such as binning, and not a mistake in the write up of the code. These tests also illuminated some issues that allowed us to make meaningful changes to the definitions of the scagnostics and the implementation of the package. For example, several tests to check the outying scagnostic was working correctly illustrated some issues in the process of outlier removal, which is illustrated in Figure @ref(fig:outlying-test-plot).

Figure @ref(fig:outlying-test-plot) 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 red line, but instead the MST connected via the long black line. The difference between these choices is essentially random, they are the exact same length, but it has significant implications for the value returned by the outlying scagnostic. This test was designed to check the outlier removal process for internal outliers, point 1 should have been identified as an internal outlier which means both its edges were considered in the calculation of outlying, and points 2 and 3 are too close to each other for either to fulfill the outlying definition, so they are left alone. Using the `draw_mst()` function when the test failled showed the issue was an esentially random desicionin the MST construction. If the red line was used to construct the MST, both the red dashed line, and the line connecting point 1 to point 2 would be included in the outlying scagnostic calcuation, in the actual calculation it was only the edge between points 1 and 2, giving a significantly smaller value on the the outlying scagnostic. This shows that even the scagnostics that work reliably well and did not need significant adjustments are still succeptible to arbitrarily large changes resulting from seemingly small changes in the visual structure of the scatter plot.

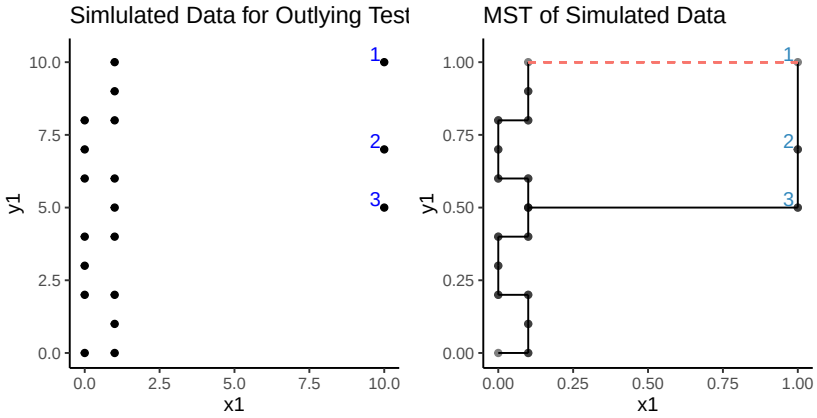

Figure 7: Plot of simulated data used for testing the 'outlying' scagnostic. The left plot shows the raw data, while the right plot presents the MST generated on that data. The edges that make it into the MST can be random and also have serious implications for outlier scagnostic. If the red edge is in the MST rather than the black edge that connects to 3, the outlying value on this plot is much higher. (#fig:outlying-test-plot)

Examples

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. 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 data set can be found in the Appendix. There are 68 variables, 33 of which are numeric, 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 @ref(fig:AFLW-scatters-static) displays 5 scatter plots (Plots 1 to 5 in the figure) 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 (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 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 early stages of exploratory data analysis. Extreme values on the scagnostic measurements identify atypical scatter plots.

The best way to identify interesting scatter plots using scagnostics is to construct a large interactive SPLOM. This is how Plots 1 to 5 were identified, 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.

Figure @ref(fig:threeplot-static) displays Plot 1, Plot2 and Plot 5 beneath the specific scatter plot of the scagnostics SPLOM that was used to identify the plot as interesting. Plot 1 was identifies 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, a visual feature that we can see 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 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

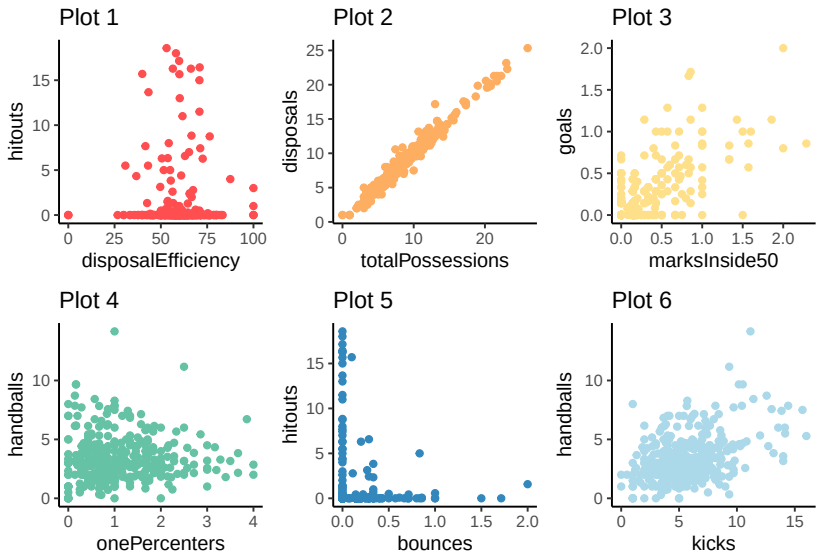

Figure 8: Six AFLW sport statistic scatter plots that were identified as identified as interesting by the scagnostics. Plots 1 to 5 had unique values an individual or pair of scagnostics, Plot 6 had middling values on all measures. There is a clear difference in structure between these plots that was identified by the scagnostics.
(#fig:AFLW-scatters-static)

the ball legally, so the strong linear relation indicates the level of play, i.e. few mistakes are made in a professional league. Plot 5 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. 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. These plots provide a clear example in 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 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 variables describe complicated physics phenomena, details of the variables will be left to the appendix and we will focus on the types of patterns observed. Since the size of the dataset is small enough, looking at the complete SPLOM 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.

The full data file contains 9998 posterior samples, and with such a large number of observations, the scagnostics cannot be computed within a reasonable timeframe without the assistance of binning. For our purpose a much smaller sample is sufficient, and we randomly sample 200 observations before computing the scagnostics. We will focus on the structures we know exist (i.e. non-linear and clustered relationships) by looking at which scatter plots have a significant difference in their splines and dcor values, as well as which plots stand out on the clumpy2 measure.

Figure @ref(fig:bbh-scags-static) shows scatterplots of the computed scagnostics measures. On the

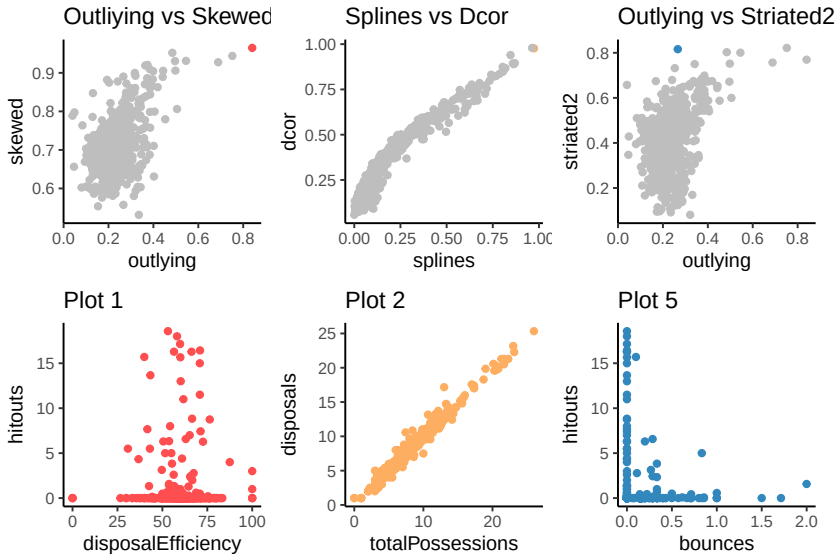

Figure 9: Three plots that were identified as interesting with the scagnostic scatter plot used to identify it. Each scatter plot of AFLW data is displayed below a plot of the two scagnostic measures it stood out on. One of the most useful ways to identify plots is through scatter plots of the scagnostics.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

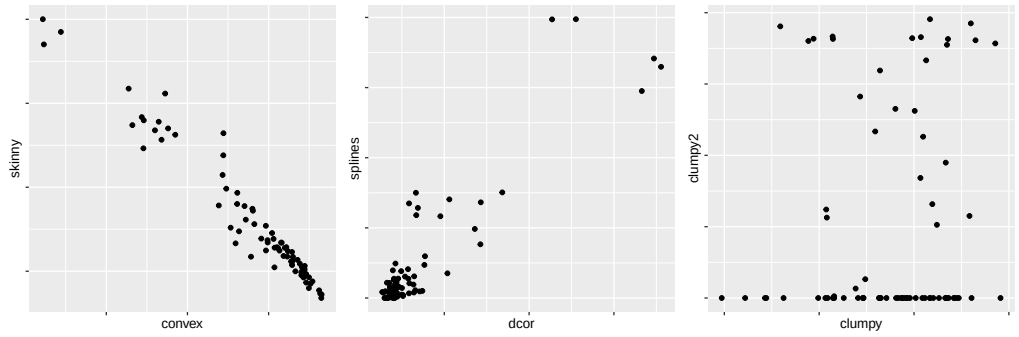

Figure 10: Pairs of scagnostics computed for the black hole mergers data. XXX (#fig:bbh-scags-static)

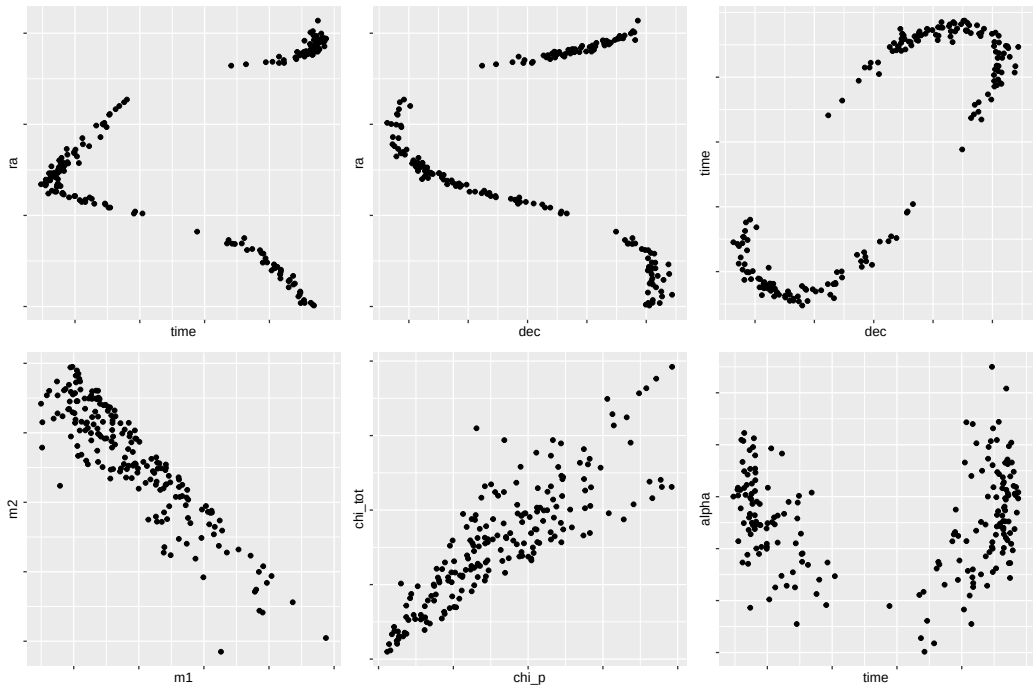

Figure 11: Features in the BBH data that stand out on several of the scagnostics measures (convey, skinny, splines and dcor), showing strong relations between variables including non-linear and non-functional dependencies. The final example (time vs alpha) is expected to take high values in clumpy, but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

left plot we see three points with very low values of the convex measure and high values of skinny. These are all possible combinations containing the variables time, ra and dec, and the corresponding scatterplots are shown in the upper row of Figure @ref(fig:blackholes). This pattern arises because the location of an event observed from gravitational waves can only be localized when using a network 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 @ref(fig:bbh-scags-static), where it is interesting to note that the combinations with non-linear but functional relation (time vs ra and dec vs ra) have 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 @ref(fig:blackholes). We see that both these combinations show noisy linear relations.

Another interesting aspect with this dataset is that there are several combinations that lead to visible separations between groups of points. It is thus an ideal testcase for our new implementation of clumpy2. The right plot in Figure @ref(fig:bbh-scags-static) 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 @ref(fig:blackholes).

Shape differences between groups

A potential application of scagnostics is to detect shape differences between groups. Commonly, classification focuses on differences in the means, or separations between groups, there are few techniques that focus on difference in shape. This difference in shape occurs when the variance patters between groups are different, and quadratic discriminant analysis (QDA), is a classical example of a method that takes this difference in variance into consideration. QDA assumes the distribution of

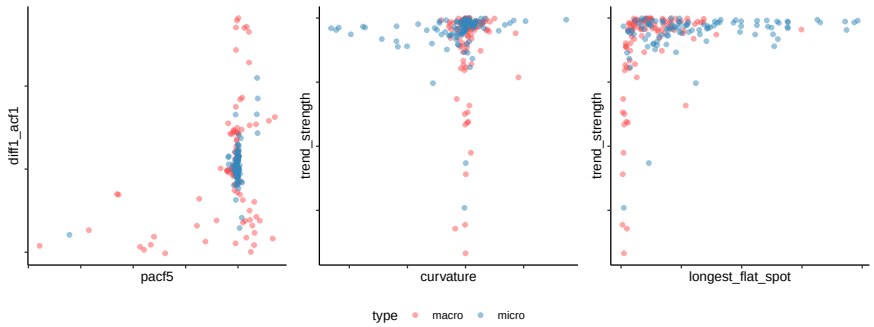

Figure 12: Interesting differences between two groups of time series detected by scagnostics. The time series are described by time series features, in order to handle different length series. Scagnostics are computed on these features separately for each set to explore for shape differences.
(#fig:timeseries)

each group is normal, and then draws a curved boundary between them that is furthest from each groups mean, but also respects that one group might have a larger elliptical variance-covariance than another group. While this method is useful when groups have different shapes, the technique is still limited by an assumption of normality. Scagnostics could be utilised in a similar fashion to QDA to identify irregular shape differences between.

This analysis compares the features of two large collections time series, and then tries to differentiate them 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 is pulled from the self-organizing database of time-series data (Fulcher et al., 2020), using the compenginets R package (Hyndman and Yang, 2021). Since the time series are different lengths, each is described by a set of time series features (chapter 4 of fpp, 2021) using the feasts R package (O'Hara-Wild et al., 2021).

For illustration, just a small set of features is examined, but still enough that the list of scatter plots identified by the scagnostics is significantly smaller than the list of all possible scatter plots. **Table** shows the pair of features that maximises the difference between groups for each scagnostic. Plotting a handful of these in (Figure @ref(fig:timeseries)), we can see the difference in shape that the scagnostics have identified. For example, the difference between the curvature and trend strength features shows both types of time series have, on average, strong trends and moderate curvature, however the former varies more in the macroeconomic series and the later in the microeconomic series. We can see from this example, and the other comparisons in the plot, that the scagnostics have identified a difference in shape that is not apparent in the mean of the data. Similar comments can be made about the other two plots in Figure @ref(fig:timeseries).

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 surpervised learning methods to be born from unsupervised learning methods. For example, principal component analysis transforms a dataset by making linear combinations of the old variables in the direction of most variance, and using these transformed variables in a linear regression can improve results. However, despite its promise, developing a classification technique is beyond the scope of this research.

Var1	Var2	scags	macro_value	micro_value	scag_dif
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	0.66	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

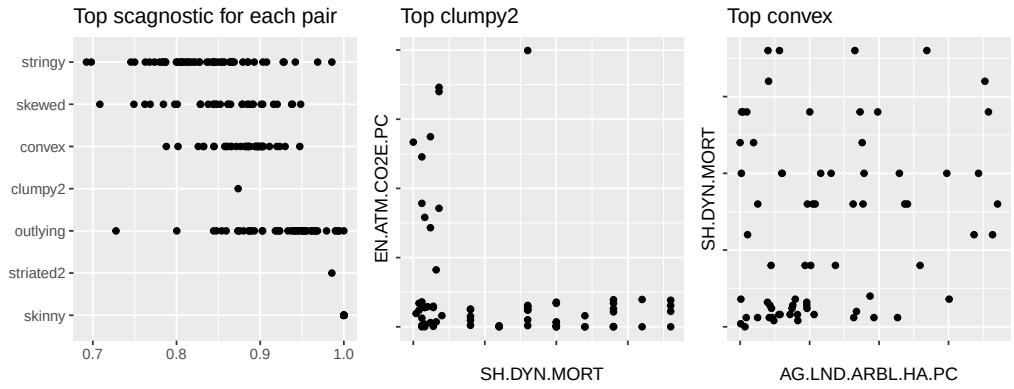

Figure 13: The range of scagnostics calculated on data with a large number of variables can help to inform the analyst about the types of relationships present. The side-by-side dotplot (left) shows one point for each pair of variables, with its highest scagnostic value among all scagnostics calculated. 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.
(#fig:wbistatic)

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 to where scagnostics can be used to identify pairs of indicators with interesting relationships, and efficiently handle missing values on a pairwise basis.

The example uses indicators from 2018 for a number of countries. The downloaded 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, allowing for a few sporadic missings. After pre-processing, there are 20 indicators (variables) and 79 countries.

Figure @ref(fig:wbistatic) (left plot) shows a summary of the top scagnostic value for each pair of variables. That is, only the highest value on any scagnostic for each pair of variables is saved. This is displayed as a vertically-oriented side-by-side dotplot. For the WBI data, the pairs of variables are producing mostly high values on stringy, skewed, convex and outlying. The scagnostics clumpy2, 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 and striped were not 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 (outlying scagnostic, and to some extent also reflected by skewned and stringy), and no relationship (convex). Scagnostics might be useful for obtaining alternative descriptive summaries of data with many variables.

The middle and right plots of Figure @ref(fig:wbistatic) show the pair of variables where clumpy2 has the highest value, and where convex has the highest value, respectively. This tells us that the data is not very clumpy.

Conclusion

Scagnostics are a useful tool to identify the visual features in scatter plots. By building upon the earlier work, we have successfully implemented previously defined scagnostics into the *cassowaryr* package, and adjusted some of them so they continue to work without a pre-processing binning step. The package is shown to work, with details on its functions, the testing process and its possible applications. We displayed these applications by giving four examples. The first, AFLW was designed to show the general use of the *cassowaryr* package, and show how to best 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 our dataset, namely by interpreting the *hitouts vs bounces* scatter plot in Figure @ref(fig:threeplot-static) . Using simulated data of a black hole merger, we then displayed the packages ability to identify pairwise non-linear relationship with splines and dcor, as well as

the improvement clumpy2 made in identifying clustering. The time series example displayed the packages value as a method of classification where it successfully identified scatter plots where the macro and micro economic time series had different shape. Finally the scagnostics were applied to the World Bank indicators to show how the scagnostics can be used to product an overall "shape summary" of a data set.

The greatest limitation in this project was the 1 year time limit on the research. When the presenting the research proposal, I thought there would be time to code up all the original scagnositcs, fix any issues with them, implement binning as an option, and have time left over to create scagnostics completely of my own design. This clearly did not occur. Ultimately coding up the previous scagnostics was more time consuming than I originally thought it would be, expecially because there was a fair bit of room for interpretation on some of the scagnostics. For example, skinny and convex do no specify what to do in the event the alphahull has no area (this occurs when the data is on a perfect straight line), and so we used intuition to set them to 1 and 0 respectively. Additionally, a large portion of the project was invisible to me at the outset, such as the testing to ensure the scagnostics were working according to definition, rounds of debugging, and meeting CRAN requirements. These elements produced very little output but were required for the rigour of the project. On top of this, to reduce dependencies most of the functions were written only using base R, which made the project more challenging. For these reasons, the software aspect of this thesis narrowed in scope throughtout the year, however the number of examples and applications increased, as we recognised new ways the scagnostics could be used throughout the year. Ultimately the final project is significantly different to its original goals, but it contains an equal amount of work.

There is a large amount of future work that could build upon this research. To start with, the distribution of the scagnostics have been largely warped by the removal of binning as a preprocessing step. In the World Bank indicators example, we made the assumption that the scagnostics are all uniformly distributed from 0 to 1, however, looking at some of our results as well as well as the visual table of the features scatter plots (Figure @ref(fig:visual-table)), shows this may not be the case. It would be a substantial task to identify the distributions of the scagnostics, and make adjustments to rectify any irregularities. The scagnostics would need to be reassessed using large volumes of data to ensure the measures are not simply identifying the structures of that data set. Looking only at the outlying scagnostic on the features data may lead us to beleive that the maximum outlier value is 0.5 and its distribution is warped, however other data sets in this paper had scatter plots that measured a perfect 1 on outlying. Without this wide variety of data it is difficult for us to make comments on the spread of the scagnostics. Scagnostics could also be expanded to be the basis of a classification technique that identifies shape differences in groups. While we showed in the time series example that scagnostics can identify shape differences between groups, expanding that observation to stand alone classification technique is outside the scope of this research. *Transforming Scagnostics to Reveal Hidden Features* (Dang and Wilkinson (2014a)) showed that scagnostics can be used to identify useful structure in pairwise plots after transformations such as log or logit transformations. This paper showed a possible natural extension for the scagnostics in the original code, but could also be considered a natural extension of the cassowaryr package. Finally, there are a handful of scagnostics that are used in the projection pursuit in the *tourr* package to find the best view of a large number of variables. The scagnostics described here could also be implemented into the package to improve the projection pursuit. Previously scagnostics have has a large amount of noise (i.e. the tourr struggles to reliably move in the direction that will have the greatest increase in the scagnostic) and checking if these scagnostics maintain that issue, and developing them to negate this issue, is another possible area of future research. It is clear from this list that there is a significant amount of research that could be built upon this scagnostics.

Bibliography

Forecasting: principles and practice, 3rd edition. OTexts: Melbourne, Australia, OTexts.com/fpp3, 2021. Accessed on October 18, 2021. [p18]

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]

F. code by R. J. Renka. R functions by Albrecht Gebhardt. With contributions from Stephen Eglen <stephen@anc.ed.ac.uk>, S. Zuyev, and D. White. *tripack: Triangulation of Irregularly Spaced Data*, 2020. URL https://CRAN.R-project.org/package=tripack. R package version 1.3-9.1. [p2]

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. Transforming scagnostics to reveal hidden features. *IEEE transactions on visualization and computer graphics*, 20(12):1624–1632, 2014a. ISSN 1077-2626. [p2, 20]

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

T. N. Dang, A. Anand, and L. Wilkinson. Timeseer: Scagnostics for high-dimensional time series. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):470–483, 2013. doi: 10.1109/TVCG.2012.128. [p2]

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]

S. Fairhurst. Triangulation of gravitational wave sources with a network of detectors. 11(12):123006, dec 2009. doi: 10.1088/1367-2630/11/12/123006. URL https://doi.org/10.1088/1367-2630/11/12/123006. [p17]

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. URL https://www.comp-engine.org. [p18]

K. Grimm. *Kennzahlenbasierte Grafikauswahl*. doctoral thesis, Universität Augsburg, 2016. [p2, 5]

R. Hyndman and Y. Yang. *compenginets: Time series from http://www.comp-engine.org/timeseries/*, 2021. URL https://github.com/robjhyndman/compenginets. R package version 0.1. [p18]

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. [p2]

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]

M. O'Hara-Wild, R. Hyndman, and E. Wang. *feasts: Feature Extraction and Statistics for Time Series*, 2021. URL https://CRAN.R-project.org/package=feasts. R package version 0.2.2. [p18]

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. [p6, 9]

L. Wilkinson and G. Wills. Scagnostics distributions. *Journal of Computational and Graphical Statistics*, 17 (2):473–491, 2008. [p2, 6]

L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 157–164, Oct 2005. [p2]

World Bank. World development indicators. the world bank group. https://databank.worldbank.org/source/world-development-indicators, 2021. Accessed 14 October 2021. [p19]

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