Penguins Go Parallel: a grammar of graphics framework for generalized parallel coordinate plots

Susan Vander Plas

Department of Statistics, University of Nebraska-Lincoln

and

Yawei Ge

Department of Statistics, Iowa State University

and

Antony Unwin

Mathematisch-naturwissenschaftliche Fakultät, Universität Augsburg

and

Heike Hofmann

Department of Statistics, Iowa State University

November 23, 2022

Abstract

Parallel coordinate plots (PCP) are a valuable tool for exploratory data analysis of high-dimensional numerical data. The use of PCPs is limited when working with categorical variables or a mix of categorical and continuous variables. In this paper, we propose generalized parallel coordinate plots (GPCP) to extend the ability of PCPs from just numeric variables to dealing seamlessly with a mix of categorical and numeric variables in a single plot. In this process we find that existing solutions for categorical values only, such as hammock plots or parsets become edge cases in the new framework. By focusing on individual observations rather than a marginal frequency we gain additional flexibility. The resulting approach is implemented in the R package ggpcp.

1 Introduction

Few approaches in data visualization exist that are truly high-dimensional. Most visualizations are projections of data into two or three dimensions enhanced by facetting or additional mappings to plot aesthetics, such as point size and color. Parallel coordinate plots are one of the exceptions: in parallel coordinate plots we can actually visualize an arbitrary number of variables to get a visual summary of a high-dimensional data set. In a parallel coordinate plot, each variable takes the role of a vertical (or parallel) axis; giving the visualization its name. Multivariate observations are then plotted by connecting their respective values on each axis across all axes using poly-lines (cf. Figure 1). For just two variables this switch from orthogonal axes to parallel axes is equivalent to a switch from the familiar Euclidean geometry to the projective space. In the projective space, points take the role of lines, while lines are replaced by points, i.e. points falling on a line in the Euclidean space correspond to lines crossing in a single point in the projective space. This duality provides a good basis for interpreting geometric features observed in a parallel coordinate plots [Inselberg, 1985].

The origins of parallel coordinate plots date back to the 19th century and are, depending on the source, either attributed to d'Ocagne [1885] or Gannett [1880]. Modern era parallel coordinate plots go back to Inselberg [1985] and Wegman [1990]. Parallel coordinate plots are used in an exploratory setting as a way of getting a high-level overview of the marginal distributions involved, identifying outliers in the data, and finding potential clusters of points. In the absence of those, Parallel Coordinate Plots are often criticized for the amount of clutter they produce, resembling a game of mikado (also known as pickup-sticks – if you are not familiar with the game, imagine spilling a box of spaghetti) rather than organized data. This clutter is sometimes mitigated by the use of α-blending [Miller and Wegman, 1991], density estimation [Heinrich and Weiskopf, 2009], or edge-bundling parallel coordinate plots [McDonnell and Mueller, 2008]. For a detailed overview of these and other techniques see Heinrich and Weiskopf [2013].

While parallel coordinate plots are a powerful tool, using categorical variables alongside quantitative variables in PCPs is a great challenge. Modifications of parallel coordinate plots have been specifically developed to deal with categorical data: parallel set plots [Kosara et al., 2006], Hammock plots [Schonlau, 2003], and common angle plots [Hofmann and Vendettuoli, 2013]; unfortunately, these solutions do not accommodate quantitative variables. Instead, they are intended for use with tabular data and show bands of observations from one categorical variable to the next. Hammock plots and common angle plots mitigate effects of the sine-illusion

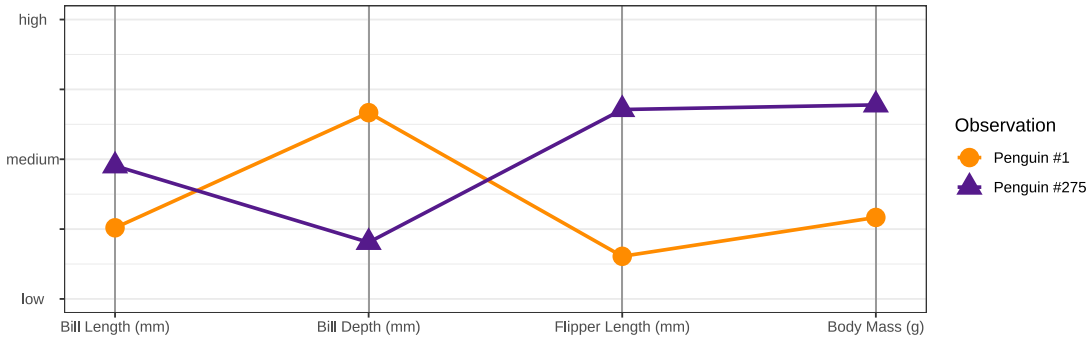


Figure 1: *Sketch of a parallel coordinate plot of two observations in four dimensions. Each dimension is shown as a vertical axis, observations are connected by poly-lines from one axis to the next. Two penguins from the Palmer Penguin data set (see section 5.1) were sampled for this example.*

[Day and Stecher, 1991, VanderPlas and Hofmann, 2015] on parallel sets plots.

An attempt to combine categorical and numeric variables in a parallel coordinate plot is introduced in the categorical parallel coordinate plots of Pilhöfer and Unwin [2013] by treating factor variables as numeric. Levels of categorical variables are transformed to numbers and variables are then used as if they were numeric. This introduces ties into the data, and the resulting parallel coordinate plot becomes uninformative, as it only shows a mesh of lines from each level of one variable to each level of the next variable. Unfortunately, the `extracat` package has not been updated recently and is no longer on CRAN. In this paper, we describe a generalization of parallel coordinate plots to accommodate both categorical and quantitative variables, developed using the grammar of graphics, implemented in the R package `ggpcp`. The resulting plots can be used to gain additional insights into multivariate data compared to plots created using other available software.

The remainder of the paper is organized as follows: Section 2 introduces the `ggpcp` syntax and explains the improvements in `ggpcp` over other parallel coordinate plot software packages. Section 3 describes the data processing for parallel coordinate plots and how this wrangling is separated from the plot rendering in `ggpcp`. Section 4 discusses the rendering of parallel coordinate plots and factors such as plotting order and tie-breaking which are important for the design of PCPs. Section 5 provides several examples which highlight different uses of generalized PCPs in exploratory settings.

2 Motivation and Package Usage

An important motivation for the `ggpcp` package is that other implementations of parallel coordinate plots for categorical variables make it difficult to follow a single observation across the chart. `ggpcp` alleviates this difficulty with two innovations: careful treatment of categorical variables to prevent line intersections at vertical axes, which maintains the visual ability to follow individual cases across the chart, and methods for ordering observations within categorical variables to reduce the amount of visual clutter. Together, these features allow for easier perception of lines in generalized parallel coordinate plots: by reducing the number of intersecting lines at pivot points along the vertical axes through case ordering, we allow our brains to leverage the gestalt principle of good continuation to follow one line across the plot. Reducing the number of line crossings at non-axis points simplifies the plot, reducing the overall cognitive load required to "untangle" (literally and metaphorically) the individual observations.

Listing 1: *A demonstration of ggpcp's data wrangling and plotting API.*

```
pcp <- penguins %>%                          # data management:
  filter(!is.na(sex)) %>%                    #   tidy workflow
  pcp_select(4,3,5:6, sex, species) %>%      #   variable selection (sec 3.1)
  pcp_scale(method="uniminmax") %>%          #   scale values (sec 3.2)
  pcp_arrange() %>%                          #   arrange categorical data
  ggplot(aes_pcp()) +                        # create chart layers:
    geom_pcp_axes() +                        #   vertical lines for axes
    geom_pcp(aes(colour = species),          #   line  segments
             alpha = 0.8, overplot="none") +
    geom_pcp_labels()                        #   label categories
```

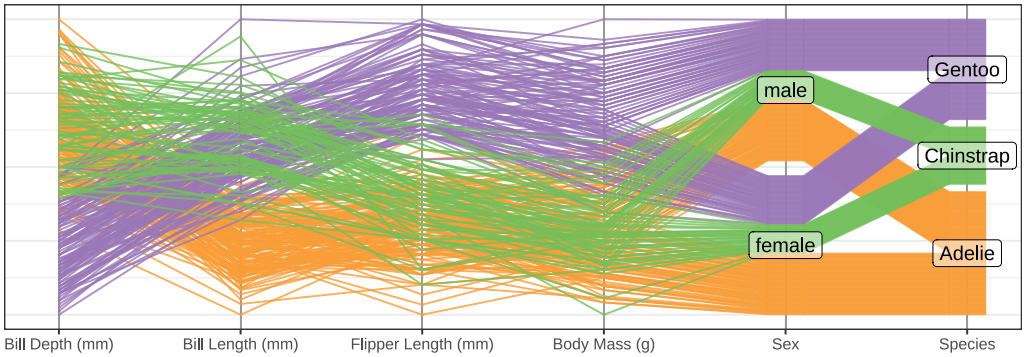


Figure 2: *The code in Listing 1 describes the data handling and basic structure of this parallel coordinate plot with both categorical and continuous data shown on vertical axes. Some minor modifications have been made to the plot for aesthetic purposes.*

In addition, `ggpcp` uses the full grammar of graphics philosophy instead of highly specific wrapper functions. This allows users to focus on the data, rather than the names of various parameters used for customization. `ggpcp` adopts tidy conventions for data wrangling, separating the necessary data manipulation to generate a parallel coordinate plot from the visual rendering, as shown in Listing 1. The arrangement of the parallel axes, ordering of cases, and scaling of variables are completed using `pcp_select`, `pcp_arrange`, and `pcp_scale`, respectively; the resulting data frame is then passed directly into the familiar `ggplot()` call. During the plotting stage, the only modification from default `ggplot2` syntax is the use of `aes_pcp()` in place of `aes()`; this is necessary to handle the multiple axes in a parallel coordinate plot while maintaining the ability to map all other variables of the original data frame to aesthetics such as linetype and color. The user has complete control over layers such as PCP lines (`geom_pcp`), labels (`geom_pcp_labels`),

and boxes around categorical variables (`geom_pcp_boxes`). The consequences of the choice to base `ggpcp` on the grammar of graphics framework, and the separation of the data wrangling and plotting, are discussed in Section 6.

An example parallel coordinate plot is shown in Figure 2, along with the `ggpcp` code to generate the plot in Listing 1. We can see that Gentoo penguins have smaller bill depth and larger flipper length and body mass than Chinstrap and Adelie penguins. Chinstrap penguins have longer bills than Adelie penguins, but are similar to Adelie penguins across most other measurements. Males tend to be larger than females across all three species. In addition, it is clear from Listing 1 that the data management process (lines 2-5) is entirely distinct from the plotting process in lines 6-9. This separation makes plots generated with `ggpcp` easy to prepare, use, and customize.

3 Data management

One of the ideas behind this re-implementation of parallel coordinate plots is to expose parallel coordinate plots at a functional level. Rather than using a single function with parameters controlling every aspect, we separate the data management from the visual rendering. In particular, we separate out the data management into three parts:

1. Variable selection and reshaping data,
2. Scaling of axes, both at the individual level and in the relationship of the axes to each other, and
3. Treatment of ties in categorical axes.

The code corresponding to each of these steps is shown in lines 3-5 of Listing 1.

The modularization of the data wrangling process has the additional advantage of laying out the necessary elements in successive steps. Some of these steps are optional: scaling variables might not be necessary if all variables are already on the same scale (i.e. method ‘raw‘ in GGally); similarly, using pcp_arrange to break ties is only necessary if categorical variables are present and we want to spread these observations out so that individual lines are visible. In addition, by exposing these elements of the pcp data wrangling process, we allow users to create additional functions for handling these tasks.

The treatment of ties is an aspect not generally addressed in the original parallel coordinate

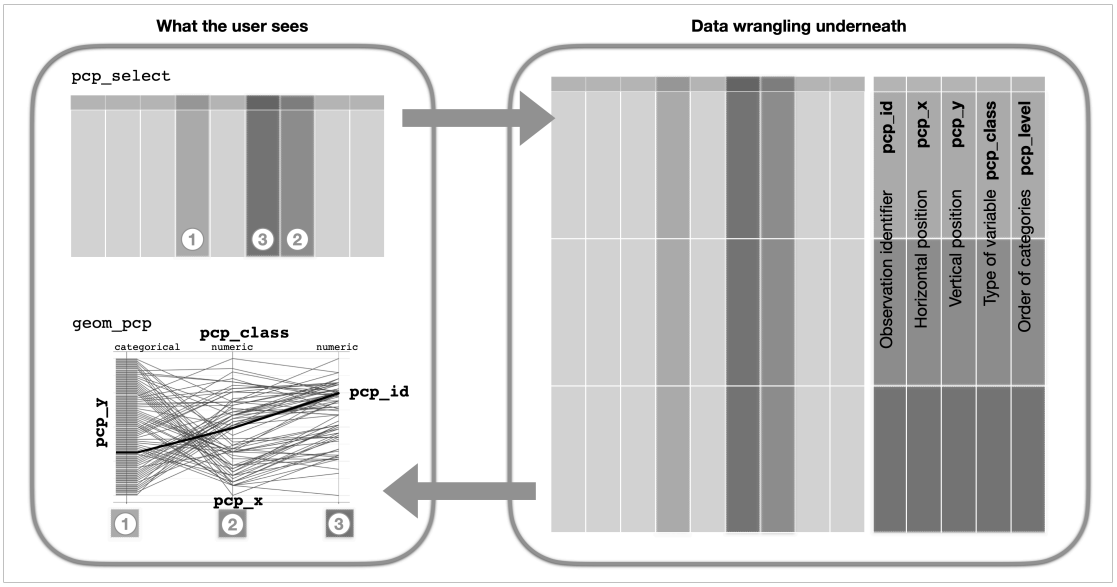


Figure 3: *The user selects a set of three variables (top left). On the right, an overview of the data wrangling step before a parallel coordinate plot can be drawn (bottom left). Note that the order in which variables are selected is reflected in the order in which variables are included in the parallel coordinate plot.*

plots of Inselberg [1985] and Wegman [1990]. We have found a need to deal with ties, because ties are visually the main obstacle to allowing the viewer to follow an observation from axis to axis through the high-dimensional space. If we can track a single observation through the high-dimensional space, we have the ability to look beyond the two-variable associations of adjacent axes. This allows users to more easily summarize main trends and identify observations which do not follow those trends. When ties cannot be separated and users cannot follow individual observations, higher-dimensional insights are next to impossible.

3.1 Variable Selection and Order of the Variables

One of the biggest strengths of the grammar of graphics is its mapping between data variables and visual aesthetics. In standard plots any mapping is a function between one data variable and one aesthetic. In a parallel coordinate plot, this one-to-one mapping between data and plot aesthetics is seemingly turned into a one-to-many mapping between arbitrarily many data variables to the x axis. By transforming the wide form of the data set into a long form [Wickham, 2014, 2021], we obtain a one-to-one mapping to a now discrete x axis consisting of the (names of the) original data variables.

From the user's perspective, this data reshaping is data selection; the data wrangling takes

place behind the scenes in `pcp_select(data, ...)`, which selects the variables to be included in the parallel coordinate plot. Variables can be specified by any combination of the following methods:

- position, e.g. `1:4, 7, 5, 4`,
- name, e.g. `class, age, sex, aede1:aede3` or
- using pattern selectors, e.g. `starts_with("aede")`, see `?tidyselect::select_helpers`

Variables can be selected multiple times and will then be included in the data and the resulting plot multiple times. Note that the order in which variables are selected determines the order in which the corresponding axis is drawn in the parallel coordinate plots. `pcp_select` transforms the selected variables to long form and embellishes the data set with a number of additional variables. All of the newly created and added variables start with the prefix `pcp_`:

- `pcp_id`: integer variable identifying each observation in the original dataset. This variable is used as the grouping variable to identify which values should be connected by a line segment in the parallel coordinate plot.
- `pcp_x`: discrete variable consisting of the names of the selected variables in the order that they were selected - this is the order in which the variables will be included in the plot.
- `pcp_y`: numeric variable containing the values of all of the selected variables. In case a selected variable is not numeric, it is converted to a factor variable and the (numeric) factor levels are saved in `pcp_y`.
- `pcp_class`: character variable containing the class information of a selected variable.
- `pcp_level`: character variable containing the factor levels of selected data variables. In case of numeric variables, the data values are stored (in textual form). The ordering of factor variables will be discussed below but it is implemented using this added variable.

As a consequence of these design decisions, users have several ways of performing different tasks within the flow of generating data for a parallel coordinate plot. For instance, users can reorder variables using `pcp_select` or after variable selection using the `pcp_x` variable. Motivations for reordering factors in parallel coordinate plots are discussed in more detail in Section 4.1.

Similarly to previous implementations of parallel coordinate plots which attempted to accommodate categorical variables, we initially treat factor variables as variables with labels and an associated (numerical) ordering of those labels. Whenever we assign a numeric value to the

ordering, we refer to the associated score, which is an integer value from one to the number of categories, if not specified explicitly otherwise. Ordered factors are plotted from the lowest level upwards. If a factor legend is included, it will need to be reversed to match this order by using `guides(color = guide_legend(reverse=TRUE))`, as shown in the example in Section 5.2. Where `ggpcp` differs from previous implementations of parallel coordinate plots is in the assignment of numerical values to individual observations within each factor level. This process is discussed further in Section 4.1.

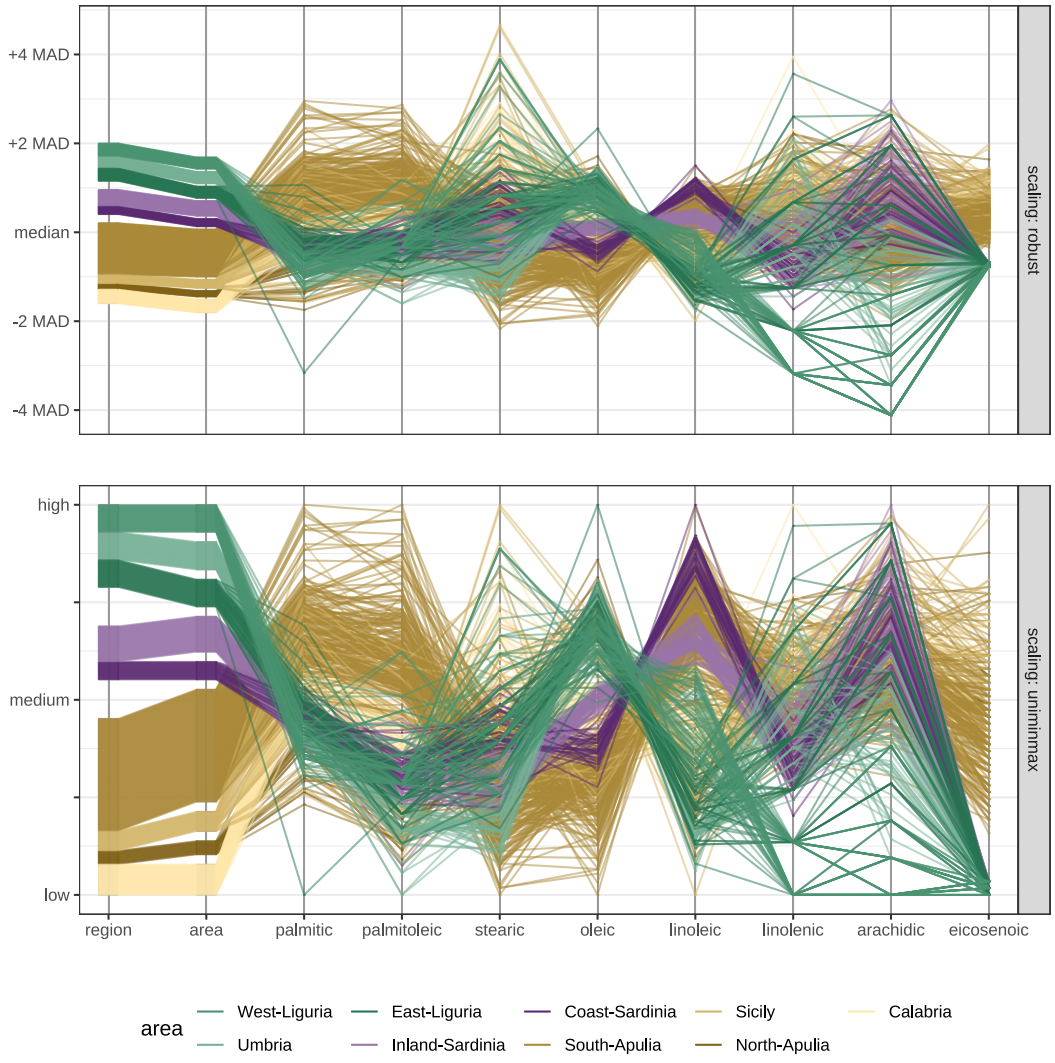


Figure 4: *Two scaling methods showing fatty acid compositions of olive oils from different regions in Italy, areas within each region are colored using similar hues within region (green for Northern Italy, purple for Sardinia, and tans for Southern Italy). The two scaling methods roughly allow the same conclusions. For readability, the y scale shows textual values rather numbers.*

3.2 Scaling

Figure 4 shows two of the scaling methods applied to the olive oil data [Forina et al., 1983, Wickham et al., 2011]: Measurements of fatty acids in 572 olive oils from three different regions in Italy are visualized as parallel coordinate plots. Similar to the findings by Cook and Swayne [2007], we see that eicosenoic acid is only found in increased quantities in olive oils from Southern Italy. Quantities of oleic and linoleic acids allow a separation between olive oils from Sardinia and Northern Italy. Both scaling methods enable us to come to these conclusions. While uniminmax scaling uses the space allotted to the chart most efficiently, the robust normalization method emphasizes the heavy tails and skewness of some of the measurements, such as the percentages of stearic and arachidic fatty acids. Both scaling methods are implemented as part of `pcp_scale`.

`pcp_scale(data, method)` scales the values on each axis and determines the relative relationship of the axes to each other. The `method` argument is a character string specifying the method to be used when transforming the values of each variable onto a common y axis. The default, `uniminmax`, univariately scales each variable onto a range of [0,1] with the minimum at zero and the maximum at one. `globalminmax` maps the values across all axes onto an interval of [0,1]. This method should only be used if the values across all variables are comparable. The method `robust` normalizes values univariately by mapping the median value to 0.5 and divide by four times the median absolute deviation. This corresponds to a mapping of a 95% confidence interval to an interval of 0 to 1. Values outside this range, as in the top plot in Figure 4, indicate a variability in the measurements larger than that of a normal distribution, as can be seen for several acid measurements.

4 Visual Rendering

4.1 Breaking ties on categorical axes

As discussed previously, one of the primary advantages of `ggpcp` over previous parallel coordinate plot software packages is that `ggpcp` handles categorical and continuous data in a way that allows users to trace a single observation through the projective space. This is accomplished through a tie-breaking algorithm: different categorical levels are grouped along the vertical axis in boxes proportional to the number of cases in each level. Within the box for a level, individual observations are arranged so that visual clutter is minimized and individual cases can be followed.

An interesting consequence of this treatment of case ordering with categorical variables is that

with large data sets, our approach looks extremely similar to existing solutions for categorical-only parallel set plots. This effect can be seen clearly in the first two vertical axes of Figure 4: similarly colored bands of lines are perceptually grouped using the Gestalt principle of common fate and, when observations are sufficiently dense, become perceptually similar to the parallelograms used in parallel set plots.

Figure 5 shows several approaches of dealing with categorical variables in parallel coordinate plots. The left-most panel shows two categorical variables and the typical net of lines that forms between them in an original parallel coordinate plot. The other three panels show three different approaches for breaking the ties resulting from the categorical variables, with our favored solution shown on the right: all observations are spaced out evenly. This results in a natural visualization of the marginal frequencies along each axis (additionally enhanced by the light gray boxes grouping observations in the same category). The ordering of the observations within the level is such that a minimal number of line crossings occurs between the axes. This method of dealing with categorical variables is the one we propose in the generalized parallel coordinate plot. While it is aesthetically pleasing, it also allows us, in the spirit of the original parallel coordinate plots, to follow an individual observation from left to right through the plot even for categorical variables. The other two solutions in the middle panels of Figure 5 show two intermediate solutions of breaking ties in categorical variables: jittering and equi-spaced (unordered) values.

When extended over multiple axes, the equispaced tie-breaking solution that reduces line crossings requires hierarchical sorting, which is implemented in the `ggpcp` function `pcp_arrange(data, method, space)`. The two implemented methods are `"from-left"` and `"from-right"`, meaning that tie breaks are determined hierarchically by variables' values from the left or the right, respectively. The parameter `space` specifies the amount of the y axis to use for space between levels of categorical variables. By default, 5% of the axis is used for spacing. While hierarchical sorting requires additional computations relative to the jittering or equally spaced solutions in Figure 5, this extra processing serves as "external cognition" [Scaife and Rogers, 1996] - the additional computer time reduces the cognitive load required to untangle the data as displayed in the chart.

Mart and Laguna [2003] discusses the NP hard problem of ordering categories to minimize line crossing. In PCPs, the category order is determined by the factor order (or the numerical scale in the continuous case); these line crossings are not avoidable through within-category sorting and are a function of the data itself. Hierarchical sorting of individual observations

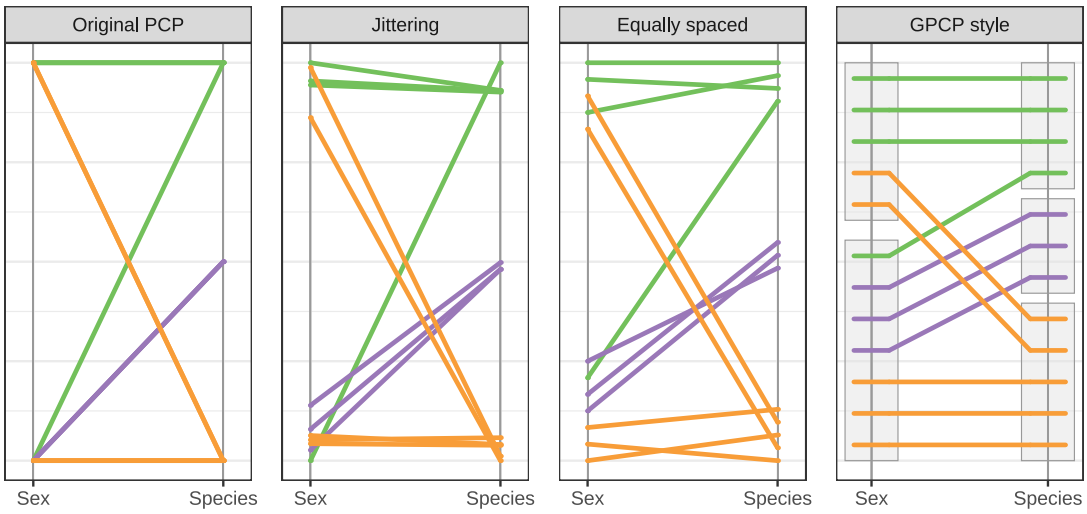


Figure 5: *Using 12 randomly sampled penguins from the Palmer penguin data, we show four different approaches of dealing with categorical variables: the panel on the left shows the typical net of lines resulting from categorical variables in regular parallel coordinate plots. In the other three panels, ties in categorical levels are broken using different approaches (from left to right): jittering, equi-spaced line segments and ordered equi-spaced line segments are shown.*

minimizes extraneous crossings within these categories in cases where there are multiple similar observations, contributing to the Gestalt of 'common fate' among individuals with similar values across a number of PCP axes.

4.2 Variable Ordering and Transformations

There are many different goals one might have when drawing a PCP; these goals shape any effort the designer might put into optimization of visual appearance. For instance, the order of factor levels is an important consideration if the goal is to minimize line crossings and thus the visual complexity of the parallel coordinate plot. As previously discussed, some line crossings can be removed by sorting, however, others can only be removed through reordering of factor levels. While *automatic* sorting of factor levels is computationally difficult and statistically undesirable given that many factors have some implicit or explicit ordering that should not be automatically optimized, *manually* reordering factors can reduce the number of line crossings to produce a simpler and more comprehensible PCP. For example, in the last panel of Figure 5, a reordering of the second factor so that the dark (purple) lines are on the bottom could reduce the overall number of line crossings to just two crossings, once the hierarchical sorting is updated to accommodate the new factor order.

As briefly discussed in Section 3.1, users can transform individual variables, reordering factors or reversing an axis, using a mutate statement before variable selection. Univariate transformations like these may be useful to reduce the overall visual complexity of a parallel coordinate plot by reducing the number of negatively correlated axes and crossing lines which are hard to follow. An example showing the benefits of reordering and transforming variables for visual clarity is provided in Figure 8.

4.3 Line Segment Plotting Order

While `ggpcp` allows us to follow a single observation through a plot, as the number of observations increases, this becomes more difficult due to overplotting. As more observations and line segments are drawn, more lines cross each other, increasing the effort required to follow a poly-line from one side of the plot to the other. Coloring by groups and utilizing α-blending improves the readability of plots. However, the order of drawing the cases affects what can be seen due to overplotting.

As a countermeasure, the order in which line segments are plotted should be carefully chosen. The parameter `overplot` defaults to option "small-on-top", where groups are plotted in order of size from largest to smallest so that the smallest group is plotted last – effectively putting the small group on top.

An alternative setting, "none", is very flexible, but requires the user to specify the order in which observations are drawn in the data processing step. The order in which observations are listed in the original data is preserved throughout the data wrangling process and directly informs the order in which lines are rendered in the layers. The use and effect of `overplot` are demonstrated in Listing 2 and Figure 6, respectively.

```
1 pcp_df <- penguins %>%
2   arrange(sex) %>% # NA last = top of PCP axis
3   pcp_select(sex, species) %>%
4   pcp_scale(method="uniminmax") %>%
5   pcp_arrange()
```

```
6 ggplot(pcp_df, aes_pcp()) + # draw lines in the provided order
7   geom_pcp_axes() +
8   geom_pcp(aes(colour = species), overplot = "none") +
```

```
  geom_pcp_labels() +
  theme_pcp()
```

```
ggplot(pcp_df, aes_pcp()) + # draw the smallest category last
  geom_pcp_axes() +
  geom_pcp(aes(colour = species), overplot = "small-on-top") +
  geom_pcp_labels() +
  theme_pcp()
```

Listing 2: *The `overplot` parameter can be used to control the order in which lines are plotted, affecting the visual appearance and emphasis of parallel coordinate plots. Line 2 specifies the order of the dataset; lines 6-8 plot the data using the user-specified ordering while lines 9-11 plot the data using the default 'small-on-top' ordering.*

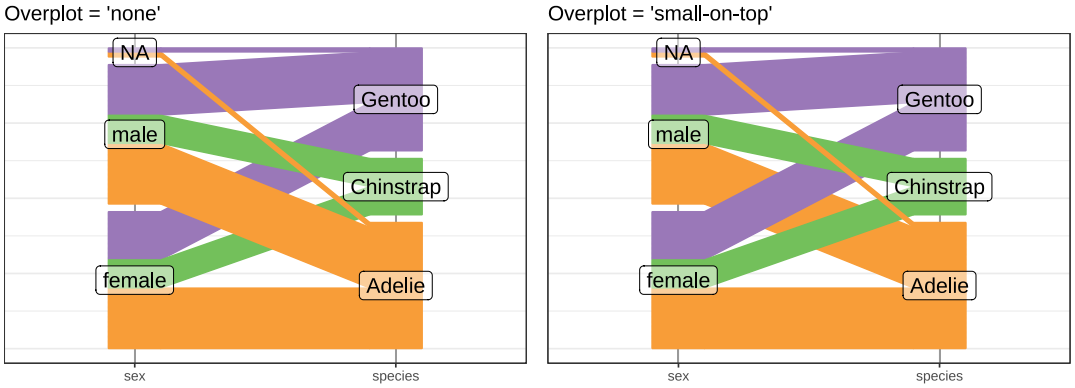


Figure 6: *The code in Listing 2 generates two parallel coordinate plots. The plot on the left uses the ordering of the dataset to determine line plotting order; as a result, Adelie penguins are plotted last (on top). On the right, we use the "small-on-top" default; this ensures that the smallest categories, sex = NA and Chinstrap, are plotted last.*

5 Examples

5.1 Palmers Penguins

Several aspects of Parallel Coordinate Plots depend on orderings: the order of variables along the x axis, the order of levels in a categorical variable, the orderings of cases within categorical variable levels, and the order in which lines are drawn. Orderings should therefore (a) have good defaults, and (b) be easily changeable.

The top of Figure 7 shows a generalized parallel coordinate plot of the Palmer penguins data [Horst et al., 2020]. The numeric data consists of body measurements of three species of penguins:

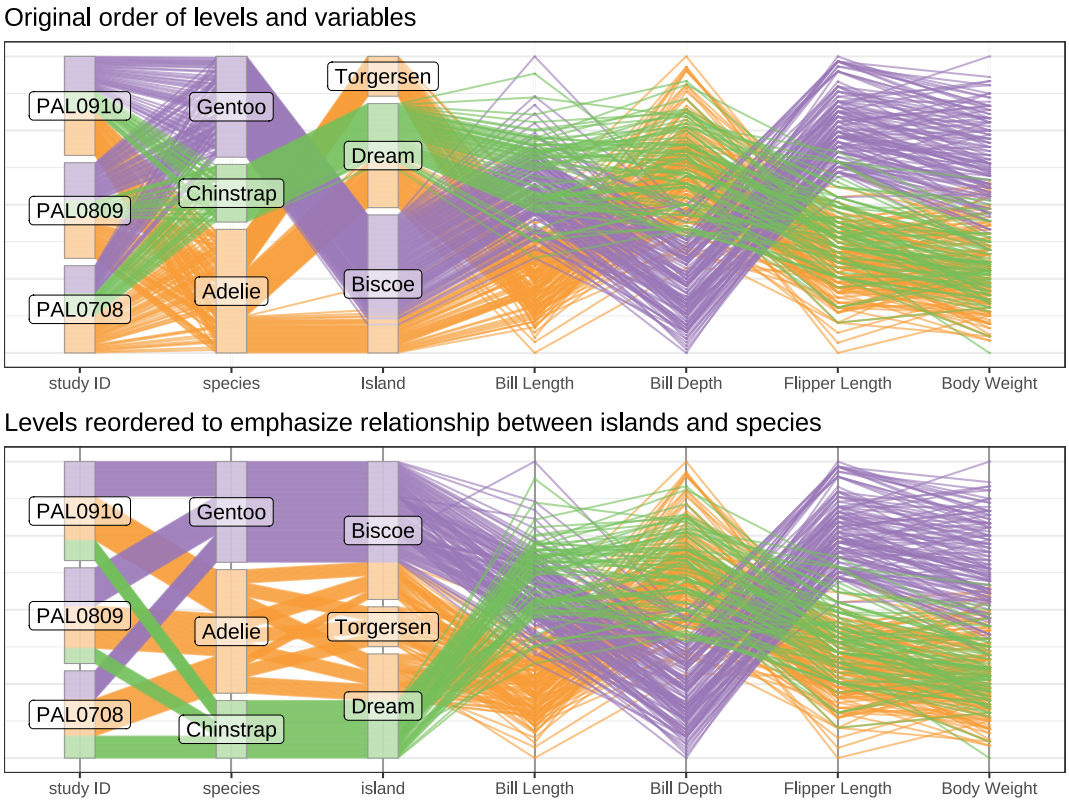


Figure 7: *Both of the levels of the island and the species variable re-ordered to reflect that two of the species are each only found on one island.*

bill length, bill depth, flipper length, body weight. Adelie penguins generally have smaller bill lengths than the other two species, while Gentoo penguins can be distinguished by their relatively large flipper lengths. The bottom of Figure 7 shows the effect of re-ordering the levels of both the 'species' and the 'island' variables in the generalized parallel coordinate plots. This re-ordering of factor levels has the effect of emphasizing that Gentoo penguins and Chinstrap penguins are each found on only one island, while Adelie penguins are found on all three islands. In addition, only after levels of 'island' and 'species' are re-ordered can we see that for each species the numbers of penguins in the three years of the study (the study ID variable) were roughly the same.

Distinguishing species

Factor level ordering is but one consideration when constructing parallel coordinate plots. It is also important to carefully order the variables on the x-axis, as shown in Figure 8, where the variables have been re-ordered from Figure 7 to allow the viewer to identify which body measurements distinguish the species. In addition to the re-ordering, the axis for bill depth has

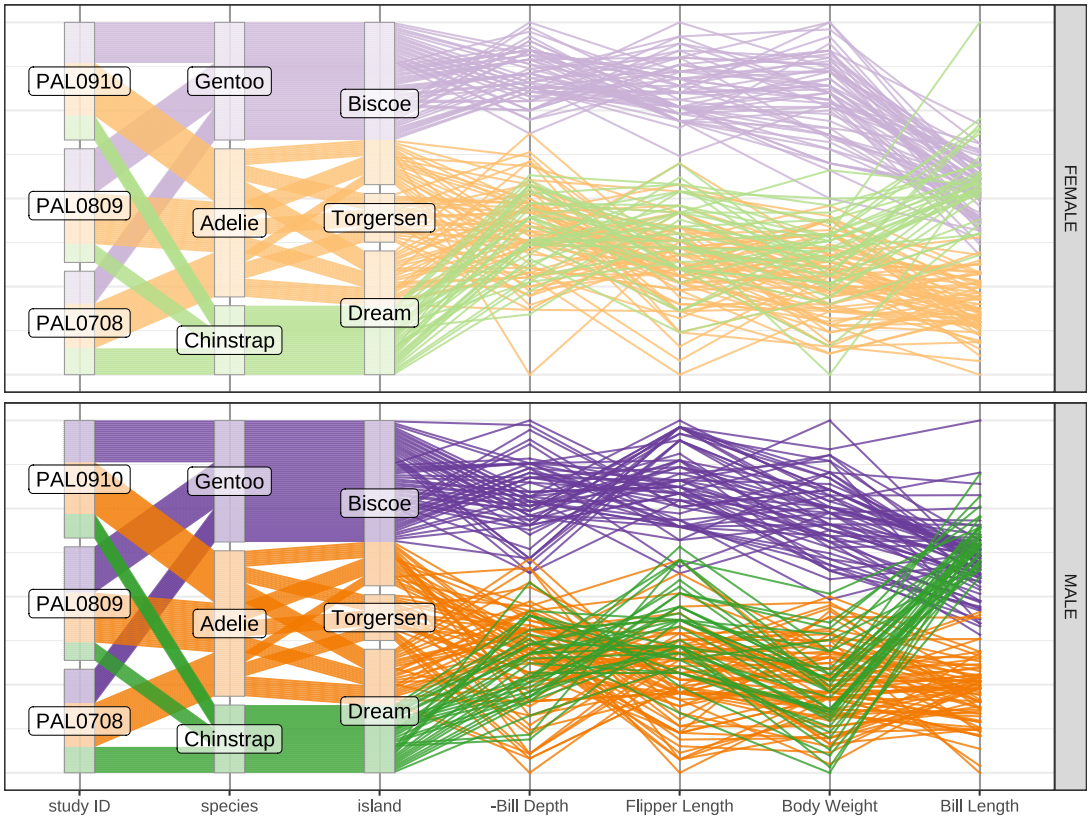


Figure 8: *Changing the order of the variables along the x-axis emphasizes the differences in body measurements between the species.*

been reversed. Both changes help to separate the species. Gentoo penguins have the lowest bill depth, while generally having the longest flippers and largest mass. Reversing the axis for bill depths aligns the smallest bill depths with the longest flippers, moving Gentoo penguins closer together as a group. The plot shows that the Gentoo penguins are bigger, that Gentoo and Chinstrap are both only found on single islands, and, finally, that Adelie and Chinstrap are distinguished by the lengths of their bills.

As `ggpcp` uses the ggplot2 API, faceting is fully supported. Figure 8 is faceted by gender: while the results are the same for the two sexes, any variability of body measures due to sex is removed from the plot by facetting. This makes the results stand out more. Interestingly, some potential outliers that were not visible previously now become visible. Note for example the two Gentoo males with particularly short flippers, and the Chinstrap female with an exceptionally long bill.

Determining sex

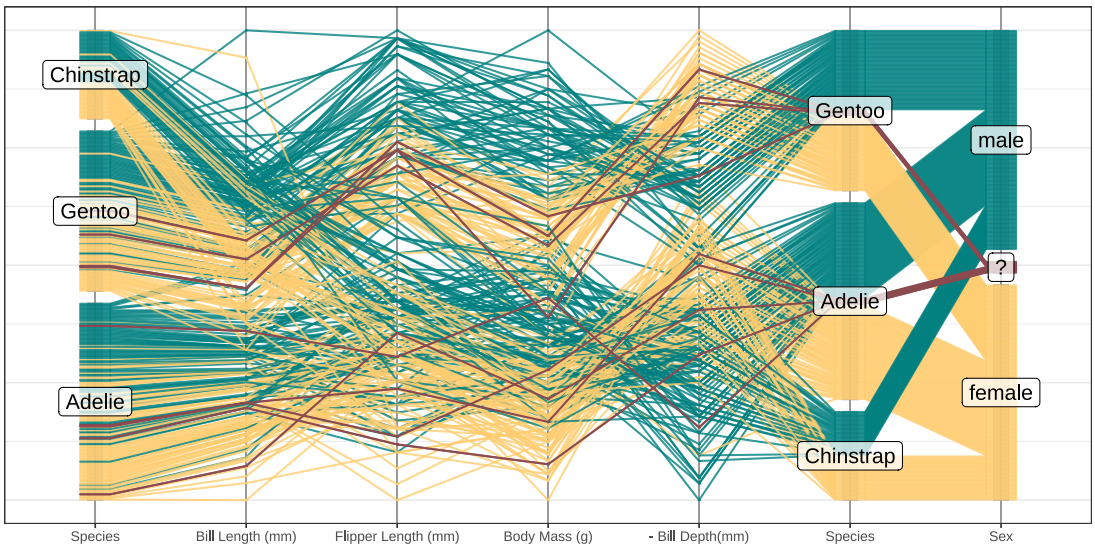


Figure 9: *Generalized Parallel Coordinate Plot of the Palmer penguins data with sex of penguin mapped to color. Dark lines represent penguins for which sex could not be determined. We see that researchers were able to sex all of the Chinstrap penguins. Note that species is included twice (with different order of the levels).*

Figure 9 shows that within each species, the males tend to be larger in size and heavier than the females. For several of the penguins, sex could not be determined because either the sexing primer did not amplify or no blood sample was obtained [Gorman et al., 2014]. These penguins are represented by dark lines. Comparing these penguins' body measurements to those of the other penguins, we can make suggestions regarding their sex.

In Figure 10 we explore this idea a bit further. This figure is based on the same data as Figure 9, however, we exclude Chinstrap penguins as researchers were able to sex all of those penguins. The body measurements of all sexed penguins are summarised by two ribbons for each sex and species. The inner ribbons are bounded by the 25% and the 75% percentile values on each axis. The lighter ribbon covers 95% of observations on each variable. We use these ribbons to reduce the noise introduced by individual lines. Body measurements of the unsexed animals are represented as line segments on top of the ribbons. This helps us to evaluate and assess the lines drawn for individual, unsexed penguins within the context of the marginal distributions (in this case their putative sex and species).

While we facet both by species and sex, note that the axes are re-scaled within each species to make use of the full range in y. However, we use the same scale between the two sexes of each species. This different treatment of faceting variables is achieved by the use of a `group_by` statement before `pcp_scale`. Listing 3 shows the code for prepping the data shown in Figure 10.

By grouping on species but not on sex (line 9), data is being rescaled within species but the same scaling is used across males and females. Measurements for unsexed animals are shown as line segments on top of inter-quantile ribbons of both sexes. Viewers are encouraged to draw a conclusion about an animal's sex based on their values within the (2d density) context of their species and putative sex. Statistically, this comparison relates to a likelihood ratio test: the viewer is asked to make an assessment of the likelihood to observe the measurements of an animal under each of the two competing hypotheses of sex.

Listing 3: *Code to prepare data for Figure 9 by relabling penguins with NA sex as '?' and ordering sex so that penguins of unknown sex are between the male and female labels.*

```
1  penguins_pcp <- penguins %>%
2    filter(species != "Chinstrap") %>%
   # no unsexed animals in Chinstrap
3    mutate(
4      sex = ifelse(is.na(sex),"?", as.character(sex)),
   # make assignment more readable
5      sex = factor(sex, levels = c("female", "?", "male"))
6    ) %>%
7    filter(!is.na(body_mass_g)) %>%
8    pcp_select(6:5, 3:4) %>%
9    group_by(species) %>%                          # re-scale by species
10   pcp_scale() %>%
11   pcp_arrange()
```

Chinstrap penguins are excluded (line 2) because all of their individuals in the data have a sex assigned. The general pattern of measurements of the Gentoo penguins suggests that three of the four individuals with missing sex information are female (the three with the lowest bill depth). The fourth animal has an exceptionally deep bill, however, all other measurements suggest that this animal, too, is female. For further evidence, we find from the original data that their nest partners are all sexed as male; this additional information is shown in Figure 10. While assuming that nest partners are male and female is not a perfect method, in particular, for penguins, which have been shown to live in same-sex partnerships, in all three of the studies considered for this data only nests with breeding successes have been considered. More details can be found in Gorman et al. [2014]. For Adelie penguins determining sex is not quite as clear-cut, but based

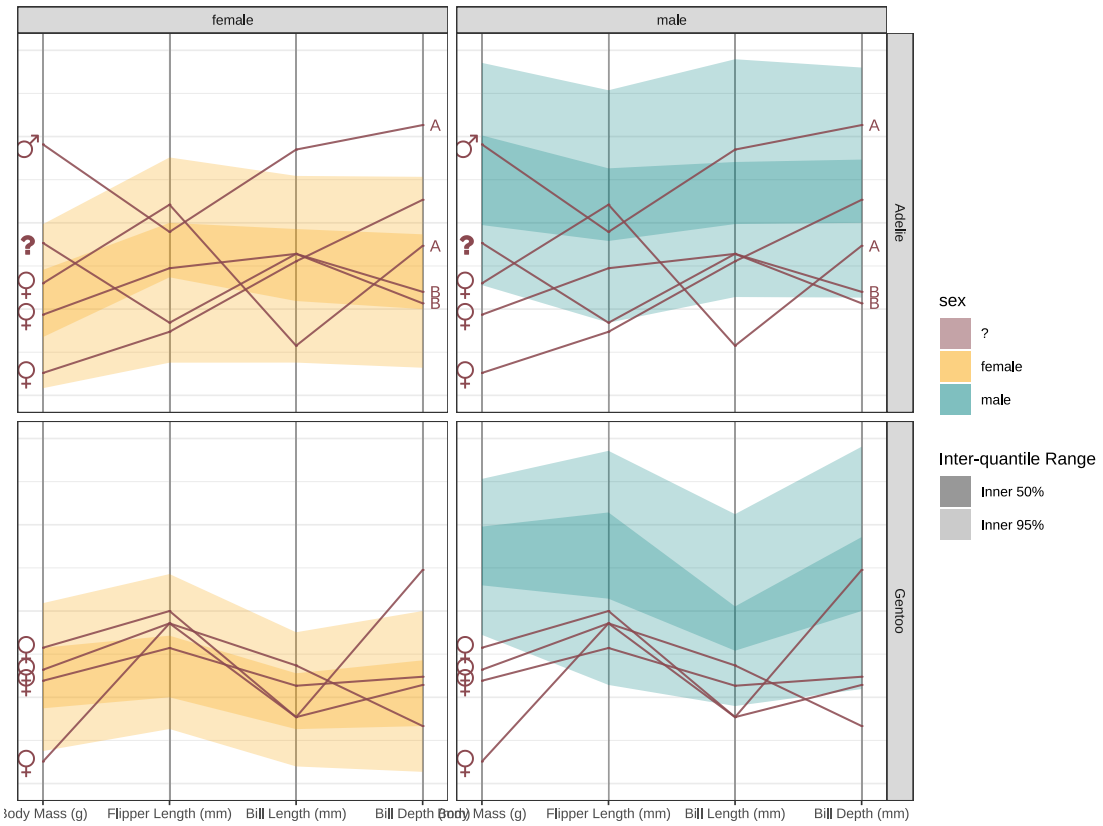


Figure 10: *Closer investigation of non-sexed Adelie and Gentoo penguins. The group_by call before pcp_scale is responsible for scaling by species while the same scale is kept across sex within species. Penguins without assigned sex (based on blood markers) are drawn on top of both sexes. The labels to the left of the ribbons are our best guess at a penguin's sex based on body measurements of other penguins of the same species. The letters on the right indicate nests – two penguins with the same letter share the same nest.*

on body mass and bill length measurements the three lightest penguins might be female, while the heaviest one could be male. The fifth penguin, marked '?', exhibits measurements that are neither typically male not typically female. We can assess these inferences using the additional information that four of the five unsexed Adelie penguins are nest partners. The un-partnered penguin is the lightest and has measurements which are more consistent with female penguins. The Adelie penguin indicated by ? is the partner of a female penguin (pair B) and might be assumed to be male. The remaining pair of unsexed Adelie penguins (pair A) consists of a putative male and female; this is consistent with the breeding pair assumption in the study.

5.2 Getting a second, third, ... and seventh opinion

Figure 11 shows data from Agresti [2002] published as part of the `poLCA` package [Linzer and Lewis, 2011]. Seven pathologists were asked to assess the same 118 slides for the presence or

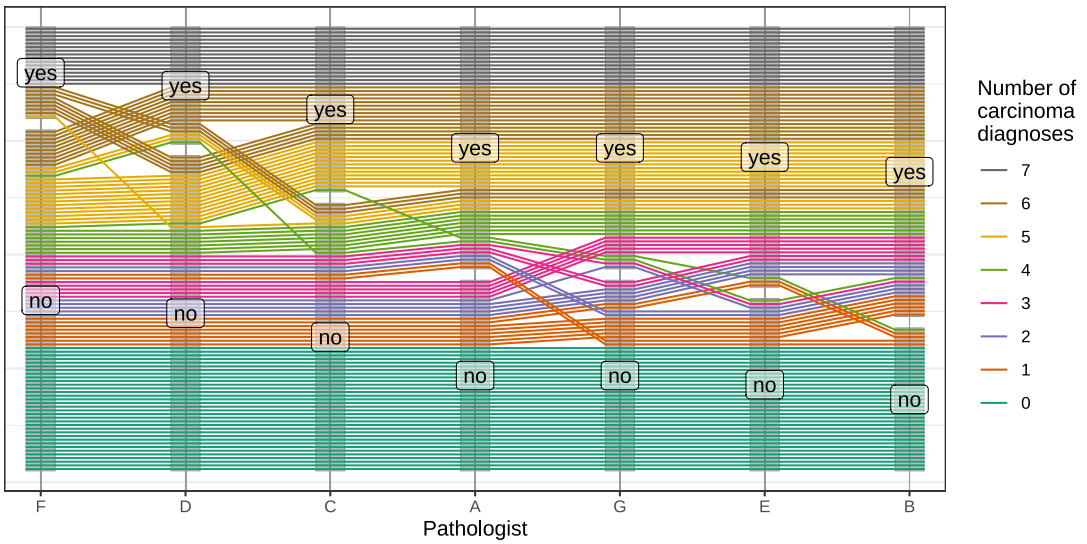


Figure 11: *Pathologists' diagnoses of absence (no) or presence (yes) of carcinoma in the uterine cervix based on 118 slides. Each slide is shown by a poly-line.*

absence of carcinoma in the uterine cervix. Binary responses for each slide were recorded (yes/no). Pathologists all agreed on about 25% of slides, which they considered to be carcinoma free, and a further 12.5% of slides, which were considered to show carcinoma by all pathologists. For the remaining 62.5% of slides there was some disagreement and it is clear that this disagreement is not random. The pathologists have been ordered from left to right from the fewest number of overall carcinoma diagnoses made to the highest number. This shows a strong level of agreement between adjacent axes. Note, in this example we do not need to scale the variables. Aside from the actual scale the values are ordered in the same way.

Landis and Koch [1977, Table 1] allow us a closer look at this data. The pathologists evaluated the slides using five levels from 1 to 5, given as: (1) Negative, (2) Atypical Squamous Hyperplasia, (3) Carcinoma in Situ, (4) Squamous Carcinoma with Early Stromal Invasion, and (5) Invasive Carcinoma. Agresti [2002] classified levels 1 and 2 as "no" and levels 3 to 5 as "yes". Figure 12 gives an overview of this more detailed data. The different pathologists are drawn in the same order as in Figure 11. The results for each scan are colored by the overall average score (rounded to the closest integer). Compared to the previous figure, Figure 12 shows more variability between pathologists' evaluations, but only few scans have vastly different scores assigned to them. Pathologist C, in particular, rates two scans as negative, that all other pathologists rate as quite advanced cancer. Mostly, the variability between pathologists' assessments stems from a difference in applying the categories rather than from an actual difference of opinions. The

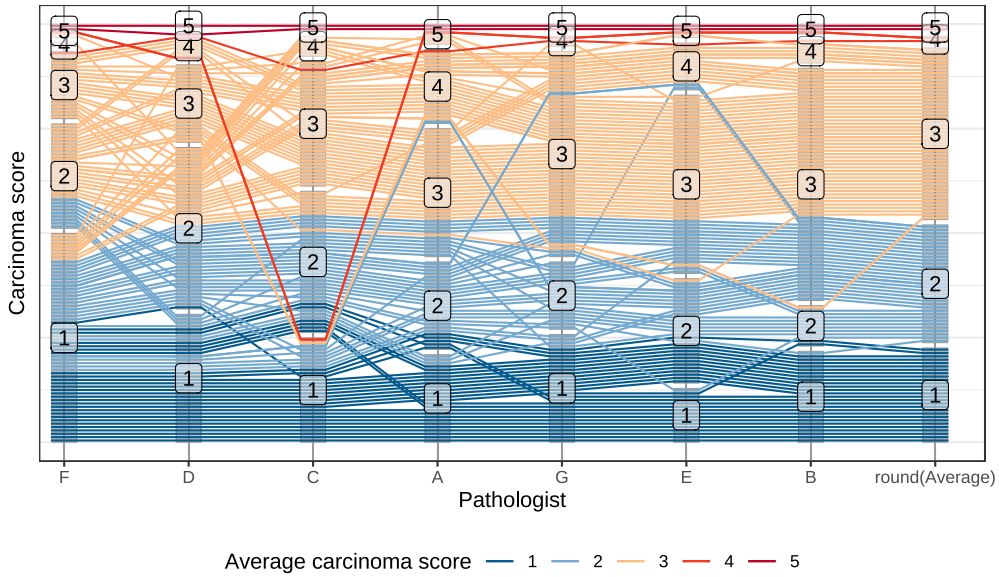


Figure 12: *Closer look at pathologists' evaluations on a more detailed scale from 1 (Negative) to 5 (Invasive Carcinoma). Rounded average scores are mapped to color to help distinguish severity of scan evaluations.*

similarity in evaluations is particularly striking between pathologists A, G, E, and B.

In this example, the generalized parallel coordinate plot gives us a visual tool for assessing the similarity between evaluations by different pathologists that moves beyond a mere correspondence of scores to an analysis that is based on ranks. The hierarchical sorting used in `pcp_arrange` assigns ranks to each observation. This provides additional information about the agreement between pathologists, which is graphically represented as the variability in line slope (that is, whether the y coordinate on each vertical axis is similar). When the poly-lines are relatively flat, this means that pathologists agree on the relative severity of the carcinoma in the scan. Obviously, we can assess 'flat-ness' of the poly-lines numerically as the variance of the calculated variable `pcp_y`. Figure 13 highlights the controversial scans, and provides additional visualizations assessing the frequency of difficult scans and the variability in `pcp_y` and in the numerical scores assigned.

5.3 Clustering with PCPs

PCPs can also be used to assess, explain, and explore statistical methods. In the penguins example, we can use k-means clustering on all numeric body measurements and investigate which observations are generally captured in each of the clusters, as well as which categorical

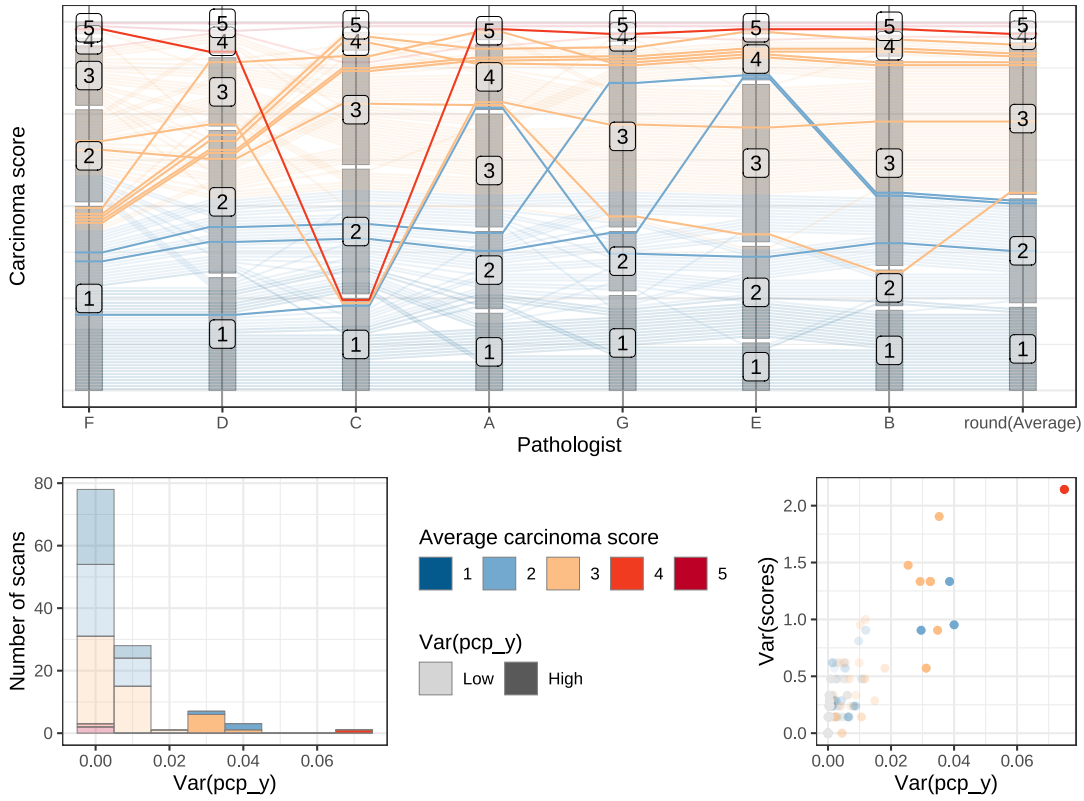


Figure 13: *Scans with a high variability in line segments are highlighted. While we might initially assume that a high line variability is directly associated with a high variability between pathologists' scores, we see from the scatterplot at the bottom right that the correlation between these two measures is not perfect. The difference lies in the tie-breaking approach: the y values for two scans with the same score on one axis are adjusted based on the scores by the other pathologists.*

variables are most associated with membership in each cluster.

k-means clustering assigns cluster labels arbitrarily based on random cluster centers. In order to maintain a persistent ordering over different values of k we reorder the cluster labels by the value of `body_mass_g`. This helps us to compare between k and $k+1$ clusters. Figure 14 shows the numeric measurements along with the assigned clusters, with categorical variables species and sex on the right. Each line is colored by the assigned cluster, allowing us to determine how the categorical variables relate to the quantitative variables and the resulting clusters. When $k = 2$, Figure 14a shows that the largest difference in the observed data is between Gentoo penguins and the other two species. When $k = 3$, in Figure 14b, the additional cluster separates the Adelie and Chinstrap penguins into two groups with a few misclassifications; this additional cluster is based on the length of the bill (which we can follow due to the clear connection between data values in the generalized PCP). Adding a fourth cluster, as in Figure 14c splits Adelie penguins into males and females, though again there are some penguins that are misclassified. The addition of a fifth cluster in Figure 14d splits Chinstrap penguins into male and female. Once we add a sixth cluster in Figure 14e, we finally split the Gentoo penguins by sex as well, though again this clustering is not perfect.

What is clear from this exercise is that Adelie and Chinstrap penguins are much more similar to each other than they are to Gentoo penguins, but that there is still noticeable sexual dimorphism within each species.

We also see from the figure that some of the separation into sexes is lost from one clustering to the next. This is typical for non-hierarchical clustering algorithms. Rather than refining a previous cluster, a switch from k clusters to $k+1$ clusters starts the clustering process anew. If the signal in the data to separate into k clusters is not strong or is ambiguous, we will see this reflected in the results; observations might be quite arbitrarily put together into groups, or a group of observations might be split into multiple clusters. In the Hartigan-Wong [Hartigan and Wong, 1979] algorithm used here for the clustering, points are assigned to random clusters in the initialization. In order to assess the effect of this non-deterministic start on the results, it is good practice to investigate the cluster stability by repeating the clustering multiple times for the same number of classes k (if $k > 1$). Figure 15 shows a comparison of the results from multiple runs of the k-means algorithm for $k = 6$. The lines in this figure are colored by species and sex. We see that the splits by species are relatively stable – there are only a few cases across all results in which individuals end up in clusters with individuals from another species, and if they do, it

is the same individuals across different results. Splits by sex show more variability: Chinstrap penguins rarely split into male/female clusters, while Gentoo penguins shows a relatively stable separation into males and females. The Adelie population has subsets of individuals that are separated into males only, females only, and a third, more variable subset of a combination of the two.

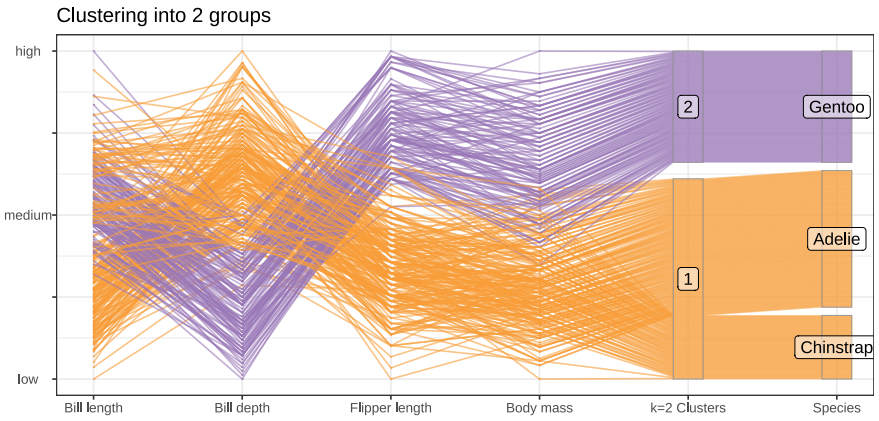
6 Design of the ggpcp API

Now that we have demonstrated the `ggpcp` API and several examples of how `ggpcp` can be used to create plots which allow users to investigate and explore multivariate data in new ways, it is useful to take a moment to reflect on the design of the `ggpcp` API and how it differs from past implementations of parallel coordinate plots.

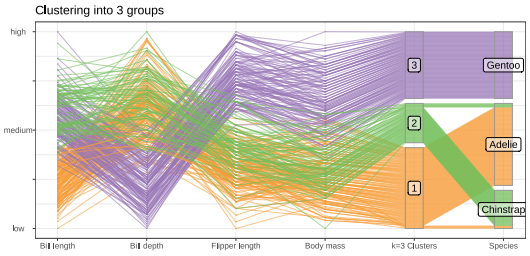
Using the `ggplot2` grammar of graphics API as a foundation has several benefits: users of ggplot2 and the tidyverse come with a general understanding of how an API is set up and **should** work. Aesthetically, users can make use of all of the 'ggplot2' functionality for customizing plots. Functionally, 'ggpcp' interfaces seamlessly with 'ggplot2' and can therefore leverage existing infrastructure, such as facetting, as used in Figure 8.

In addition, designing `ggpcp` using the `ggplot2` framework expands the functionality available to users without much additional code, thanks to other packages, such as `plotly` [Sievert, 2020] and `gridSVG` [Murrell and Potter, 2020], which extend `ggplot2` to create interactive graphics for the web.

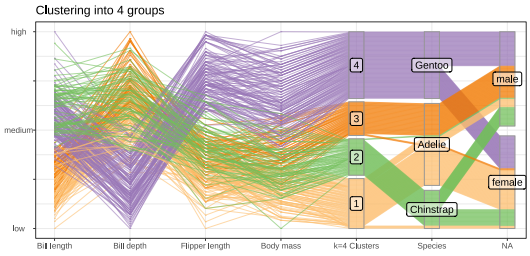
While we make use of the general `ggplot2` API for plotting the data, `ggpcp` makes the additional decision to separate the data wrangling from the plotting. This is a slight deviation from the `ggplot2` extension approach, as data summaries and modifications in ggplot2 are implemented in *statistics* functions. These functions, named 'stat_xxx', where 'xxx' is usually the name of the corresponding geom to capture the close relationship between the 'geom_xxx' and the 'stat_xxx' functions. For example, 'stat_boxplot' calculates the summaries necessary for drawing a boxplot, such as the mean, quartiles, and IQR. 'stat_bin' is the default statistics associated with histograms: continuous variables are binned and a frequency count is being visualized. A 'stat_pcp' data function would therefore have to deal with the translation from a the user-friendly wide dataset to the technically motivated long form of the data that allows a direct mapping of one variable to the x axis (implemented as 'pcp_x') and another variable to the y



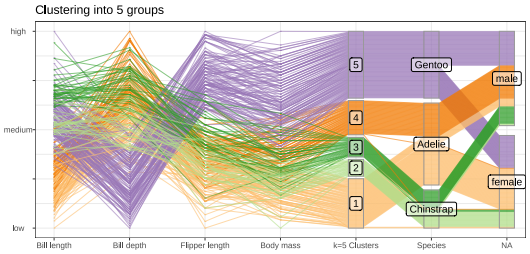
(a) *When $k = 2$, Gentoo penguins are separated from Adele and Chinstrap penguins.*



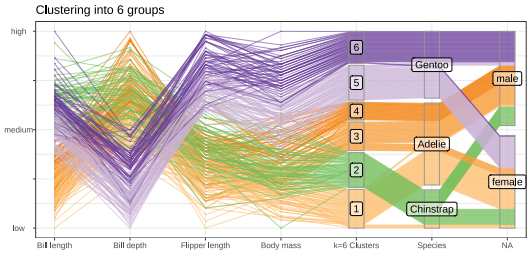
(b) *The third cluster splits the former second cluster into two based on the length of the bill. The three species are almost perfectly separated in using three clusters.*



(c) *The fourth cluster splits Adelie penguins (mostly) into males and females of the species.*



(d) *The fifth cluster splits the group of Chinstrap penguins (mostly) into females and males. Only a handful of individual penguins are grouped with the wrong species or the wrong sex.*



(e) *With the introduction of a sixth cluster, all previous clusters change: Gentoo penguins are split into males and females. The sex separation for Chinstrap penguins gets lost, but Adelie males get split into two separate clusters.*

Figure 14: *An overview of the use of parallel coordinate plots to examine which variables contribute to clustering and to identify individuals who are misclassified.*

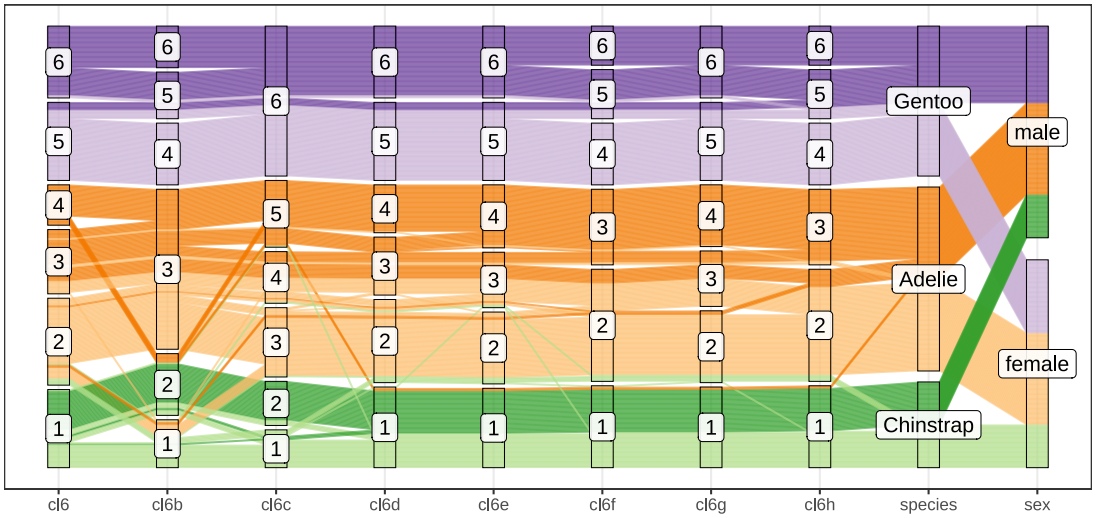


Figure 15: *Comparison of eight k-means runs for $k = 6$. Color of lines is given by species and sex. The differences between the clusters are introduced by the different random seeds in the inital cluster centers.*

axis (implemented as ‘pcp_y‘).

Obviously, we decided against an implementation of this ‘stat_pcp‘ function, resulting in the need to wrangle the original data into the correct format ahead of the call to ggplot. Splitting data wrangling and plotting provides several benefits.

1. **Speed:** moving the data wrangling out of the inner workings of ‘ggplot2‘ removes the necessity to repeat this step each layer. This results in a considerable speed up for larger datasets or more intricate plots with multiple layers.

2. **Transparency and Flexibility:** modularizing and exposing the data wrangling pipeline into individual steps creates a better conceptual framework and gives users more flexibility to make changes: users can interact with and modify intermediate results at each step.

3. **Reducing the function clutter:** most parallel coordinate plot functions come with a LOT of parameters – many of them aimed at controlling the exact layout of the plot, dealing with considerations such as ’showPoints’ for drawing points or ’boxplot’ for drawing a boxplot. These parameters become unnecessary in the ggpcp implementation because of the direct availability of the ggplot2 layer system. Should there be a need for boxplots on top of the default PCPs, a simple call to ‘geom_boxplot‘ will draw them, and depending on whether this call happens before or after a call to ‘geom_pcp‘, the boxplots end up behind or in front of the lines. Modularizing the code for data wrangling and visualizing

means that parameters for parallel coordinate plots can be placed directly in their relevant functions. This helps with interpreting parameters as well as removing the need for calling the same parameters in multiple layers to ensure that all layers are based on the same data.

The use of data wrangling functions outside of the plotting apparatus allows us to format the data in a consistent way (and execute that operation only once). This is somewhat analogous to the way the `sf` package [**?**] handles plotting based on different geometries: by customizing the format of the data that is passed into the ggplot extensions, we can write plotting functions that conform to the expected use of ggplot2 layers that use data with a different underlying structure.

7 Discussion

This paper describes generalized parallel coordinate plots, which extend parallel coordinate plots to include categorical variables. This extension is a significant development: GPCPs are useful in a wide variety of scenarios where standard PCPs were insufficient; in addition, the handling of categorical variables introduced as part of the `ggpcp` implementation of GPCPs opens up many new areas for PCP-related research.

The most consequential feature of GPCPs as implemented in the `ggpcp` package is the ability to follow a single observation through multiple categorical and continuous axes. This continuity provides a perceptual advantage over other alternatives even for plots of only categorical variables, as it is possible to visually assess $N > 2$-dimensional contingent relationships using GPCPs, as demonstrated in Figure 15. Throughout the examples in this paper, we have attempted to showcase the impacts of this visual continuity through a range of different applications of GPCPs. We have attempted to assemble a broad set of such examples, but we expect that GPCPs will be useful in many other applications.

As a consequence of the line continuity in generalized PCPs, we also highlight the importance of four different types of ordering: axes, factors, lines, and plotting. While previous papers have examined the impact of axis ordering, there is a large increase in the importance of factor, line, and plotting order to preserve the line continuity afforded by GPCPs. The `ggpcp` framework allows the ordering of factors, lines, and plotting to carry additional information which affects the user's ability to understand the data. These factors, along with the use of color in PCPs, deserve much more investigation and consideration than the brief overview provided in this paper,

which is focused on demonstrating the `ggpcp` API. The implementation of GPCPs provided in the `ggpcp` package neatly separates data management from visual rendering while leveraging the `ggplot2` API. These features contribute to the power and flexibility of the `ggpcp` package.

Online Material

This manuscript has been created in the reproducible Rweave format [Xie, 2014, 2015] in R [R Core Team, 2022] using the RStudio IDE (version Spotted Wakerobin). All original files, data, and code can be found at `https://github.com/srvanderplas/ggpcp-paper`. The 'ggpcp' package is available from CRAN. We would like to thank all the contributors to open software, in particular, the authors behind the 'tidyverse' packages [Wickham et al., 2019]. We acknowledge that a lot of this work is only possible with the help of unpaid volunteers and developers.

References

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Hoboken, 2 edition, 2002.

D. Cook and D. F. Swayne. *Interactive and Dynamic Graphics for Data Analysis With R and GGobi*. Springer Publishing Company, Incorporated, 1st edition, 2007. ISBN 0387717617.

R. H. Day and E. J. Stecher. Sine of an illusion. *Perception*, 20:49–55, 1991.

M. d'Ocagne. Coordonnées parallèles et axiales : Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles. *Gauthier-Villars*, page 112, 1885. URL `https://archive.org/details/coordonnesparal00ocaggoog/page/n10`.

M. Forina, C. Armanino, and S. Lanteri. Classification of olive oils from their fatty acid composition. *Food Research and Data Analysis*, pages 189–214, 01 1983.

H. Gannett. General summary showing the rank of states by ratios 1880, plate 71. In *Scribner's statistical atlas of the United States, showing by graphic methods their present condition and their political, social and industrial development*. Charles Scribner's Sons, New York, 1880.

K. B. Gorman, T. D. Williams, and W. R. Fraser. Ecological sexual dimorphism and environmental variability within a community of antarctic penguins (genus pygoscelis). *PLOS ONE*, 9(3):1–14, 03 2014. doi: 10.1371/journal.pone.0090081. URL `https://doi.org/10.1371/journal.pone.0090081`.

J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100, 1979. doi: 10.2307/2346830. URL http://dx.doi.org/10.2307/2346830.

J. Heinrich and D. Weiskopf. Continuous Parallel Coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1531–1538, 2009. doi: 10.1109/TVCG.2009.131. URL http://ieeexplore.ieee.org/document/5290770/.

J. Heinrich and D. Weiskopf. State of the Art of Parallel Coordinates. In M. Sbert and L. Szirmay-Kalos, editors, *Eurographics 2013 - State of the Art Reports*. The Eurographics Association, 2013. doi: 10.2312/conf/EG2013/stars/095-116.

H. Hofmann and M. Vendettuoli. Common Angle Plots as Perception-True Visualizations of Categorical Associations. *IEEE Transactions on Visualization and Computer Graphics*, 19 (12):2297–2305, Dec. 2013. doi: 10.1109/TVCG.2013.140. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6634157.

A. M. Horst, A. P. Hill, and K. B. Gorman. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*, 2020. URL https://allisonhorst.github.io/palmerpenguins/. R package version 0.1.0.

A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, Aug. 1985. doi: 10.1007/BF01898350. URL http://link.springer.com/10.1007/BF01898350.

R. Kosara, F. Bendix, and H. Hauser. Parallel Sets: interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, 2006. doi: 10.1109/TVCG.2006.76. URL http://ieeexplore.ieee.org/document/1634321/.

J. R. Landis and G. G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33(2):363–374, Jun 1977. doi: https://doi.org/10.2307/2529786.

D. A. Linzer and J. B. Lewis. poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10):1–29, 2011. URL http://www.jstatsoft.org/v42/i10/.

R. Mart and M. Laguna. Heuristics and meta-heuristics for 2-layer straight line crossing minimization. *Discrete Applied Mathematics*, 127(3):665–678, May 2003. ISSN 0166-218X. doi: 10.1016/S0166-218X(02)00397-9. URL https://www.sciencedirect.com/science/article/pii/S0166218X02003979.

K. T. McDonnell and K. Mueller. Illustrative Parallel Coordinates. *Computer Graphics Forum*, 27(3):1031–1038, May 2008. doi: 10.1111/j.1467-8659.2008.01239.x. URL http://doi.wiley.com/10.1111/j.1467-8659.2008.01239.x.

J. J. Miller and E. J. Wegman. *Computing and Graphics in Statistics*, chapter Construction of Line Densities for Parallel Coordinate Plots, pages 107–123. Springer-Verlag New York, Inc., New York, NY, USA, 1991. ISBN 0-387-97633-7. URL http://dl.acm.org/citation.cfm?id=140806.140816.

P. Murrell and S. Potter. *gridSVG: Export 'grid' Graphics as SVG*, 2020. URL https://CRAN.R-project.org/package=gridSVG. R package version 1.7-2.

A. Pilhöfer and A. Unwin. New Approaches in Visualization of Categorical Data: R Package extracat. *Journal of Statistical Software*, 53(7), 2013. doi: 10.18637/jss.v053.i07. URL http://www.jstatsoft.org/v53/i07/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

M. Scaife and Y. Rogers. External cognition: how do graphical representations work? *International journal of human-computer studies*, 45(2):185–213, 1996.

M. Schonlau. Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, 2003.

C. Sievert. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020. ISBN 9781138331457. URL https://plotly-r.com.

S. VanderPlas and H. Hofmann. Signs of the sine illusion–why we need to care. *Journal of Computational and Graphical Statistics*, 24(4):1170–1190, 2015. doi: 10.1080/10618600.2014.951547. URL https://doi.org/10.1080/10618600.2014.951547.

E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Assoiation*, 85:664–675, 1990.

H. Wickham. Tidy data. *Journal of Statistical Software, Articles*, 59(10):1–23, 2014. ISSN 1548-7660. doi: 10.18637/jss.v059.i10. URL https://www.jstatsoft.org/v059/i10.

H. Wickham. *tidyr: Tidy Messy Data*, 2021. URL https://CRAN.R-project.org/package=tidyr. R package version 1.1.3.

H. Wickham, D. Cook, H. Hofmann, and A. Buja. tourr: An R Package for Exploring Multivariate Data with Projections. *Journal of Statistical Software, Articles*, 40(2):1–18, 2011. ISSN 1548-7660. doi: 10.18637/jss.v040.i02. URL https://www.jstatsoft.org/v040/i02.

H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Y. Xie. knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, and R. D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. URL http://www.crcpress.com/product/isbn/9781466561595. ISBN 978-1466561595.

Y. Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL https://yihui.org/knitr/. ISBN 978-1498716963.