A grammar of graphics framework for generalized parallel coordinate plots

Author 1[1*] Author 2[2] Author 3[3]

[1] University 1; [2] University 2; [3] University 3

September 3, 2021

Abstract

Parallel coordinate plots (PCP) are a useful tool in 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 additional mappings to plot aesthetics, such as point size and color, or facetting. Parallel coordinate plots are one of the exceptions: in parallel coordinate plots we can actually visualize an arbitrary many 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 polylines (cf. **??**). 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 plot [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 to get a high-level overview of the marginal distributions involved, to identify outliers in the data and to find potential clusters of points. In the absence of those, Parallel Coordinate Plots are often critized for the amount of clutter they produce, resembling a game of mikado rather than organized data. This clutter is sometimes combatted 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].

However, parallel coordinate plots have some shortcomings. XXXX after bashing clutter in pcps, 'some shortcomings' feels like rubbing it in :) The biggest challenge comes when working with categorical

*Corresponding author: xxx

variables. In current solutions, levels of categorical variables are transformed to numbers and variables are then used as if they were numeric. This introduces a lot of ties into the data, and the resulting parallel coordinate plot becomes uninformative, as it only shows lines from each level of one variable to all levels of the next variable. Some versions of parallel coordinate plots have been specifically developed to deal with categorical data, e.g. parallel set plots [Kosara et al., 2006], Hammock plots [Schonlau, 2003], and common angle plots [Hofmann and Vendettuoli, 2013]. These solutions all have in common that they work with tabularized data and show bands of observations from one categorical variable to the next. Hammock plots and common angle plots provide solutions to mitigate the sine-illusion's effects [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]. These plots provide an extension to parallel sets that allows numeric variables to be included in the plot. Similar to parallel sets, this approach is also based on marginal frequencies for the categorical variables. Categorical parallel coordinate plots are the closest of these variations to our solution, but they are not implemented in the `ggplot2` framework and can therefore not be further extended.

Various packages in R [R Core Team, 2019] exist that contain an implementation of one of the parallel coordinate plots. The function "parcoord" in the `MASS` package [Venables and Ripley, 2002] makes use of the base plot system of R to draw parallel coordinate plots. The function "cpcp" in package `iplots` implements the parallel coordinate plot [Pilhöfer and Unwin, 2013]. Developments based on the grammar of graphics [Wilkinson, 2005] and the `ggplot2` [Wickham, 2016] framework are, e.g. the function 'ggparcoord' in `GGally` [Schloerke et al., 2018] or `ggparallel` [Hofmann and Vendettuoli, 2016] which provide an implementation of Hammock and common-angle plots.

Those packages based on `ggplot2` make use of `ggplot2`, but are actually wrapper of existing functions for highly specialized plots with tens of parameters, which do not allow the full flexibility of `ggplot2` and do not make use of `ggplot2`'s layer framework.

```
## Error in pcp_scale(., method = "globalminmax"):  could not find function "pcp_scale"
```

The remainder of the paper is organized as follows: section 2 describes the data processing for parallel coordinate plots.

2 Data management

The idea 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 treatment of ties is an aspect not generally addressed in the original parallel coordinate plots of Inselberg [1985] and Wegman [1990]. We have found a need to deal with ties, because ties are visually the main obstacle of allowing the viewer to follow an observation from axis to axis through the high-dimensional space. Once this assessment is broken for the individual observation, the we cannot reasonably expect to gain much more information by plotting more data in the plot. XXX this is clumsy - what I am really aiming for needs a bit more: the way plots should be used in order to make sense of complicated situations is to provide (1) a visual summary of the main trends and (2) allow us to identify observations that do not follow those trends. By mapping all plots that have the same value on one dimension into a single point on an axis, we hinder any higher-dimensional insights.

XXX test test

The modularization of the data wrangling process has the additional advantage to lay out the necessary elements in successive steps. Some of these steps are even optional - e.g. 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 there are any factor variables, and if we actively want to spread these observations out.

2.1 Variable selection

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 aesthetic and one data variable. 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 . However, by transforming the wide form of the data set into a long form [Wickham, 2007, Wickham et al., 2021, Wickham, 2014, 2021], we get to a form of the dataset in which we achieve 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 has purely the form of a data selection, while the data wrangling is going on behind the scenes in this function.

`pcp_select(data, ...)` allows a selection of variables to be included in the parallel coordinate plot. Variables can be specified by

- 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`

or any combination thereof. Variables can be selected multiple times and will then be included in the data and the resulting plot multiple times. 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_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_level`: character variable containing the factor levels of selected data variables. In case of numeric variables, the data values are stored (in textual form). Ordering of levels in factor variables (XXX not sure where to put this, yet): What we are doing with the ordering of levels in factor variables, is to stick with the basic interpretation of factor variables as a type of variable that has both labels and an 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. This means in particular, that the first level of a categorical variable is mapped to the lowest value along the y axis rather than the 'top' value as e.g. when mapped to 'fill' in a barchart. This might lead to a visual insonsistency between the orderings in the levels of a categorical variable and an accompanying color legend. In those situations we suggest to reverse the order in the legend by using the command `guides(color = guide_legend(reverse=TRUE))` as shown in the example in subsection 4.2.
- `pcp_class`: character variable containing the class information of a selected variable.
- `pcp_id`: integer variable identifying each observation in the original dataset. This variable will be used as grouping variable to identify which values should be connected by a line segment in the parallel coordinate plot.

XXX reordering variables in the parallel coordinate plot could be done before the selection by pcp_select or after it by re-ordering the levels in the pcp_x variable. We should probably include an example.

XXX univariate transformations, such as reversing an axis for a variable can be done using a mutate statement before the variable selection. example? changing order of levels on a categorical variable or a log transform or square root transformation work the same.

2.2 Scaling

`pcp_scale(data, method)` scales the values on each axis and determines the relative relationship of the axes to each other.

`method` is a character string specifying the method to be used when transforming the values of each variable into a common y axis. By default, the method `uniminmax` is chosen, which univariately scales each variable into a range of [0,1] with a minimum at 0 and the maximum at 1. `globalminmax` maps the values across all axes into a 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 a robust 95% confidence interval (based on the median absolute deviation) to an interval of 0 to 1.

?? shows two of the scaling methods at the example of the olive oil data [Cook and Swayne, 2007, Wickham et al., 2011, Forina et al., 1983]: measurements of fatty acids in 572 olive oils from three different regions in Italy are visualized as parallel coordinate plots. Similar to the findings in 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 find these conclusions.

```
## Error in pcp_arrange(.):  could not find function "pcp_arrange"
## Error in dt1$scaling <- "uniminmax":  object 'dt1' not found
## Error in pcp_arrange(.):  could not find function "pcp_arrange"
## Error in dt2$scaling <- "std":  object 'dt2' not found
## Error in rbind(dt1, dt2):  object 'dt1' not found
## Error:    You're passing a function as global data.
##   Have you misspelled the `data` argument in `ggplot()`
```

2.3 Breaking ties on categorical axes

`pcp_arrange(data, method, space)` provides a rescaling of values on categorical axes to break ties. `method` is a parameter specifying which variables to use to break ties. The two implemented methods are `"from-left"` and `"from-right"`, meaning that ties are broken using a hierarchical ordering using variables' values from the left or the right, respectively. The parameter `scale` 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.

?? 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 of 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 lightly greyed 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 **??** show two intermediary solutions of breaking ties in categorical variables: jittering and equi-spaced (unordered) values.

```
## Error in pcp_scale(.):  could not find function "pcp_scale"
## Error in pcp_arrange(.):  could not find function "pcp_arrange"
## Error in pcp_scale(.):  could not find function "pcp_scale"
## Error in pcp_scale(.):  could not find function "pcp_scale"
## Error in arrangeGrob(...):  object 'p1' not found
```

3 The Generalized Parallel Coordinate Plot

XXX We need to include some code somewhere to also show the 'how'.

4 Examples

4.1 Penguins

?? shows a first generalized parallel coordinate plot of the Palmer penguins data [Horst et al., 2020]. The data consists of body measurements, such as weight, flipper length, bill length, and depth, of three species of penguins. What can be seen is that Adelie penguins generally have smaller bill lengths than the other two species, while Gentoo penguins can be distinguished from the other two species by their relatively larger flipper lengths. The lines in **??** are colored by sex of penguins. What can be seen is 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 [@penguins2]. These penguins are represented by black lines. Based on these penguins' body measurements within the context of the other penguins, we can make some suggestions regarding their sex. **??** shows body measurements of penguins in generalized parallel coordinate plots facetted by species. Chinstrap penguins are excluded because all of their individuals in the data have a gender assigned. The general pattern of measurements of the Gentoo penguins suggests that all four individuals with missing sex information are female (for further evidence, we find from the original data that their nest partners are all sexed as male). For Adelie penguins the situation is not quite as clear-cut, but based on body weight and bill length measurements the three lightest penguins might be female, while the heaviest one could be male. The fifth penguin walks the line between typical male and typical female measurements.

```
## Error in pcp_arrange(.):  could not find function "pcp_arrange"
## Error in ggplot(., aes_pcp()):  object 'penguins_pcp' not found
```

```
## Error in pcp_arrange(.):  could not find function "pcp_arrange"
## Error in ggplot(., aes_pcp()):  object 'penguins_pcp' not found
```

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

?? 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 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. However, we see that this disagreement is not random. When pathologists are ordered (by moving the corresponding axes) left to right from fewest number of overall carcinoma diagnoses to highest number, we see that generally for a slide more pathologists make a carcinoma diagnosis from left to right.

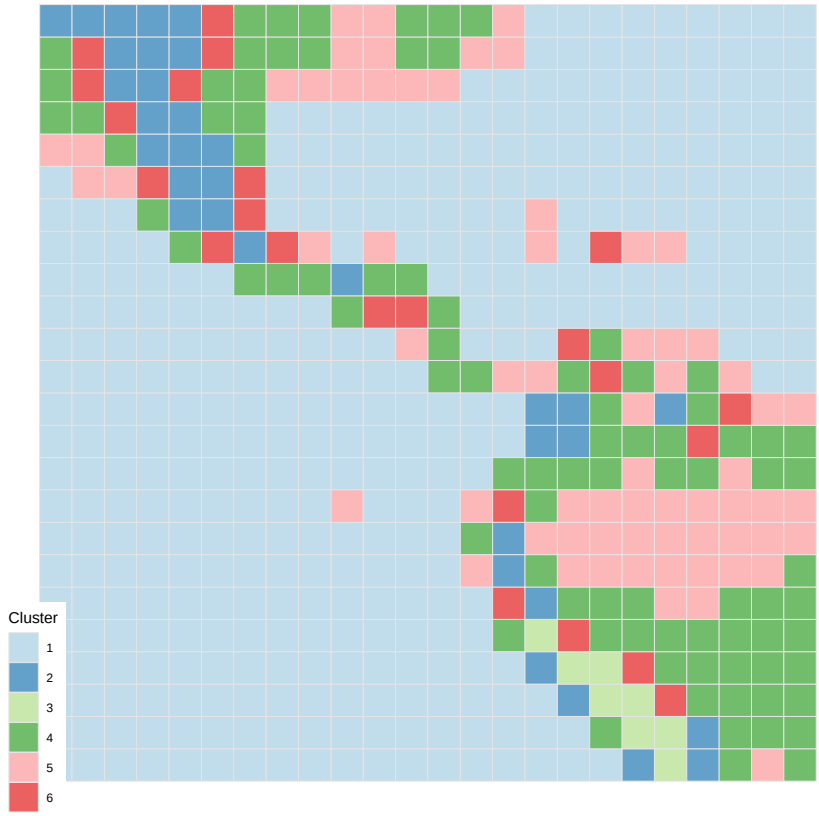

Figure 1: *Tile plot of the (gridded) geographic area underlying the data. Each tile is colored by its cluster membership.*

```
## Error in pcp_arrange(.):  could not find function "pcp_arrange"
```

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.

4.3 ASA Data expo 2006

In this example, we re-visit a data set that was used for the ASA Data Expo in 2006. The 'nasa' data, made available as part of the 'ggpcp' package provides an extension to the data provided in the 'GGally' package Schloerke et al. [2018]. It consists of monthly measurements of several climate variables, such as cloud coverage, temperature, pressure, and ozone values, captured on a 24x24 grid across Central America between 1995 and 2000.

Using a hierarchical clustering (based on Ward's distance) of all January measurements of all climate variables and the elevation, we group locations into 6 clusters. The resulting cluster membership can then be summarized visually. Figure 1 shows a tile plot of the geography colored by cluster. We see that the clusters have a very distinct geographic pattern.

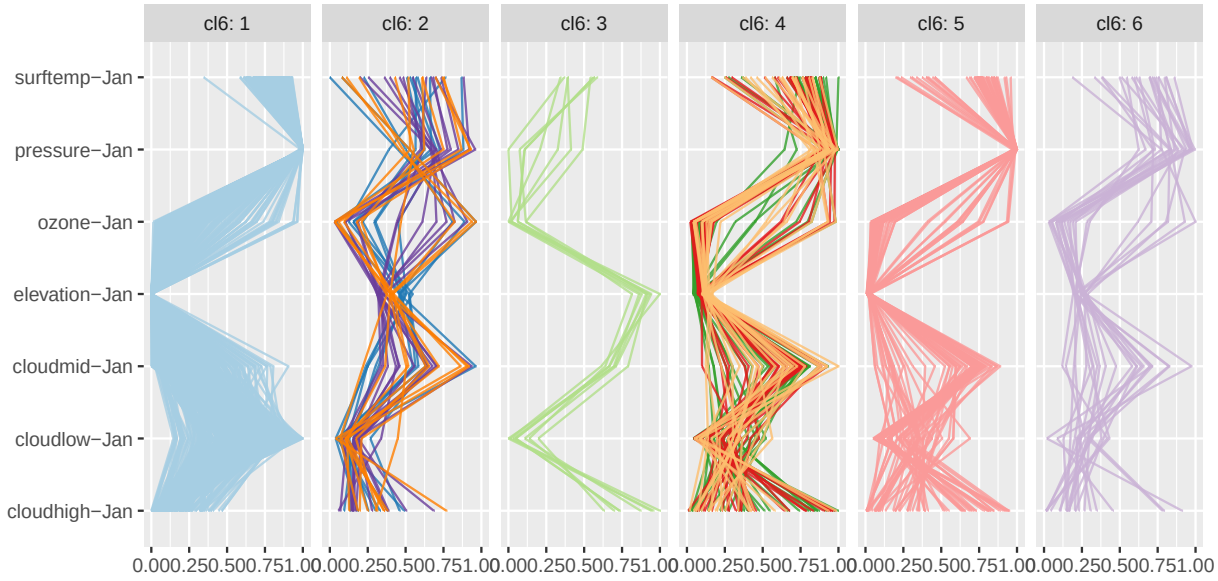

Figure 2: *Overview of all variables involved in the clustering.*

XXX changed the data for the clustering - need to adjust the description From the parallel coordinate plot in Figure 2 we see that cloud coverage in low, medium and high altitude distinguishes quite succinctly between some of the clusters. (Relative) temperatures in January are very effective at separating between clusters in the Southern and Northern hemisphere. The connection between the US gulf coast line and the upper region of the Amazon (cluster 2) can probably be explained by a relatively low elevation combined with similar humidity levels.

A parallel coordinate plot allows us to visualize a part of the dendrogram corresponding to the hierarchical clustering.

```
## Error in pcp_scale(.):  could not find function "pcp_scale"
```

Using the generalized parallel coordinate plots we can visualize the clustering process in plots similar to what Schonlau Schonlau [2002, 2004] coined the clustergram, see **??** and **??**.

```
## Error in pcp_arrange(.):  could not find function "pcp_arrange"
```

Along the x-axis the number of clusters are plotted with one PCP axis each, from two clusters (left) to 10 clusters (right most PCP axis). Each line corresponds to one location, lines are colored by cluster assignment in the ten-cluster solution. This essentially replicates the dendrogram while providing information about the number of observations in each cluster as well as the relationship between successive clustering steps.

5 Results

6 Discussion

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.

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`.

H. Hofmann and M. Vendettuoli. *ggparallel: Variations of Parallel Coordinate Plots for Categorical Data*, 2016. URL `https://cran.r-project.org/package=ggparallel`. R package version 0.2.0.

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/`.

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/`.

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`.

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, 2019. URL https://www.R-project.org/.

B. Schloerke, J. Crowley, D. Cook, F. Briatte, M. Marbach, E. Thoen, A. Elberg, and J. Larmarange. *GGally: Extension to 'ggplot2'*, 2018. URL https://CRAN.R-project.org/package=GGally. R package version 1.4.0.

M. Schonlau. The clustergram: a graph for visualizing hierarchical and non-hierarchical cluster analyses. *The Stata Journal*, 2(4):391–402, 2002.

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.

M. Schonlau. Visualizing hierarchical and non-hierarchical cluster analyses with clustergrams. *Computational Statistics*, 19(1):95–111, 2004.

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.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, 4 edition, 2002. ISBN 0-387-95457-0. URL http://www.stats.ox.ac.uk/pub/MASS4/.

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

H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 2007. URL http://www.jstatsoft.org/v21/i12/paper.

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. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2 edition, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

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, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.7.

L. Wilkinson. *The Grammar of Graphics*. NY: Springer, New York, 2 edition, 2005.