Breaking the Clutter: Handling Numerical Ties in Generalized Parallel Coordinate Plots

Denise Bradford, Susan VanderPlas, PhD and Heike Hofmann, PhD

Abstract

Parallel coordinate plots (PCPs) provide a versatile means of visualizing multivariate data, especially when dealing with high-dimensional datasets. The ggpcp package extends the layered grammar of graphics `ggplot2` to generate parallel coordinate plots from tidy data, emphasizing numerical ties between variables. This paper discusses the conceptual underpinnings, functionality, and applications of `ggpcp`, illustrating how it facilitates nuanced insights into the relationships among numerical variables. The integration of multiple datasets and advanced tie-breaking techniques further expands its utility.

Introduction

Parallel coordinate plots (PCPs) are among the few visualization methods that effectively represent high-dimensional data. Originating from d'Ocagne and later modernized by Inselberg and Wegman, PCPs map each variable to a parallel vertical axis, connecting observations across axes using polylines. Despite their utility, traditional PCPs often struggle with representing categorical variables or a mix of categorical and continuous variables, limiting their broader applicability.

The `ggpcp` package, as part of the grammar of the graphics framework, addresses these limitations by extending the capabilities of PCPs to accommodate both categorical and numerical variables. Building on advanced tie-handling methods—including Incremental Offset, Jitter Spacing, and Cascade Rank Tie-Breaker—`ggpcp` enables analysts to derive meaningful insights from datasets with extensive ties.

Functionality

The `ggpcp` package offers several key features for creating and customizing parallel coordinate plots:

1. Data Preprocessing

`ggpcp` assumes input data is in a tidy format, where each variable is a column, and each observation is a row. Preprocessing functions like `pcp_select` and `pcp_arrange` facilitate the transformation of data into a format suitable for parallel coordinate plotting. These features also handle numerical ties through techniques such as:

- Incremental Offset: Adds small values to tied figures, maintaining the natural order without randomness.
- Jitter Spacing: Introduces slight random variations to tied values for visual separation.
- Cascade Rank Tie-Breaker: Resolves ties hierarchically by leveraging prior variable ranks.

2. Integration of Categorical and Numerical Variables

Building on generalized PCPs (GPCPs), `ggpcp` supports categorical variables by introducing equispaced and ordered arrangements. Techniques like "Tie Breaking the Band" apply incremental offsets or adaptive fractions to differentiate values within categorical and numerical ties.

3. Scaling and Normalization

The package supports various scaling methods, including uniminmax and robust normalization. These methods adaptively adjust axes to ensure all variables are visually comparable, even across datasets with diverse scales.

4. Multiple Dataset Integration

`ggpcp` supports the analysis of multiple datasets by aligning variables and managing tie-breaking strategies uniformly across data sources. This feature is particularly valuable for comparative studies or meta-analyses.

5. Line Rendering Order

To address overplotting, `ggpcp` provides options for controlling the order in which observations are rendered. Strategies like "small-on-top" or user-defined orders ensure critical patterns remain visible in dense datasets.