POINTS OF VIEW

Power of the plane

Two-dimensional visualizations of multivariate data are most effective when combined.

High-dimensional data pose a significant analytical and representational challenge. One instinctual response has been to represent data in three-dimensional (3D) space in order to capture additional information[1]. Given the common medium utilized for science communication, great utility can be achieved by pushing the communicative power of the endless 2D planes that surround us in the form of pieces of paper, computer monitors and video projections.

Data visualization methods such as parallel coordinate plots and scatter plots displayed in an array can be highly useful 2D visualization techniques for high-dimensional data. They represent data using location on a plane, and each has its own strength for highlighting different aspects of the data. Many data analysis tasks involve looking for clusters, trends and outliers, and well-chosen and well-designed 2D plots can be highly advantageous in revealing patterns in data.

A fundamental 2D plotting technique is the use of parallel coordinates (**Fig. 1a**,**b**). The characteristic appearance of these plots comes from their unique coordinate system: the coordinates are parallel rather than orthogonal to each other. Each vertical axis depicts a different dimension with data values scaled between a minimum and a maximum (**Fig. 1a**). Data points belonging to the same row are connected by line segments, which allows individual data features to be shown in the context of the overall data set.

Parallel coordinates can handle a variety of data types simultaneously. For example, gene expression data and other quantitative multivariate data over time or multiple conditions are often visualized using a special case of parallel coordinate plots in which each dimension is of the same type and all axes are scaled to the same range (**Fig. 1b**). This approach enables accurate comparisons across dimensions. In addition, these types of plots can also represent data sets that contain categorical, ordinal or quantitative dimensions.

By relying on robust graphical encodings, parallel coordinate plots make certain data relationships clear. For example, the appearance of many crossing lines between a pair of axes indicates an inverse relationship between the corresponding dimensions, whereas parallel (or nearly parallel) lines could suggest correlation between variables represented by neighboring axes (**Fig. 1a**,**b**). These types of features are easy to see in parallel coordinate plots. However, these plots are not well suited for data dominated by categorical information or data ranges that pass through only a small number of values, as data occlusion becomes a problem.

When using parallel coordinates, ensure that the axis height and the distance between the axes are adjusted so that the average of the absolute values of all angles is close to 45 degrees. The aspect ratio of the overall plot influences the angle at which line segments appear between axes. Proper shaping of parallel coordinate plots will improve the viewer's perception of the axe's orientation and make it easier to spot line crossings—useful for tracing individual profiles.

Scatter plot matrices are another common planar visualization method for multivariate data (**Fig. 1c**). In this plotting technique, pairwise relationships between all dimensions of a data set can be readily explored using a grid of scatter plots that represent all pairwise combinations.

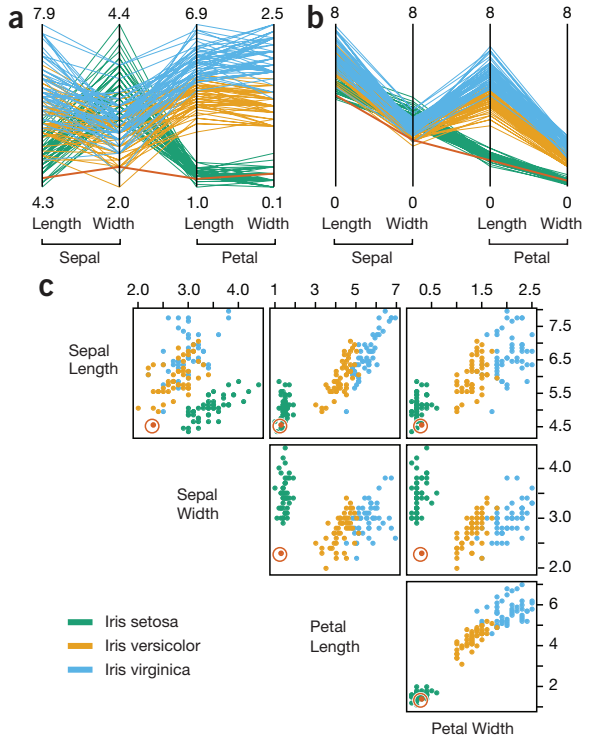

Figure 1 | Visualizations of the Iris data set popularized by R.A. Fisher[2]. (**a**) Parallel coordinate plot with unscaled axes. (**b**) The same data as in **a** plotted using scaled axes. (**c**) Scatter plot matrix. The same data feature is highlighted in red to illustrate how data across dimensions is represented in the two visualization types.

The choice between a parallel coordinate plot and a scatter plot matrix depends on the analytical task to be supported. The fundamental difference in the approaches is how they represent individual data features across multiple dimensions. A data point in a parallel coordinate plot is depicted as a single line or a profile (**Fig. 1a**,**b**). Together, the 'bundles' of lines point out clusters, and outliers therefore become apparent. A scatter plot matrix, on the other hand, represents a data feature as a series of points that are not connected across the scatter plots, making it difficult to draw conclusions about individual data features (**Fig. 1c**). However, scatter plot matrices can be used to efficiently identify pairwise correlations and other relationships between all dimensions in the overall dataset based on the characteristic shapes of the point clouds.

These methods complement each other and will deliver the best results when used in an interactive setting in which multiple coordinated visualizations of the same data set are available. Along with heat maps and dimensionality reduction tools, fundamental 2D visualization methods can be powerful approaches to multivariate data.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Nils Gehlenborg & Bang Wong

1. Gehlenborg, N. & Wong, B. *Nat. Methods* **9**, 851 (2012).
2. Fisher, R.A. *Annals of Eugenics* **7**, 179–188 (1936).

Nils Gehlenborg is a research associate at Harvard Medical School and the Broad Institute. Bang Wong is the creative director of the Broad Institute and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.