Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions

Eser Kandogan
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
eser@us.ibm.com

Abstract

Visualizing multi-dimensional data has tremendous effects on science, engineering, and business decision-making. A new visualization technique called Star Coordinates is presented to support users in early stages of their visual thinking activities. Star Coordinates arranges coordinates on a circle sharing the same origin at the center. It uses simply points to represent data, treating each dimension uniformly at the cost of coarse representation. Current implementation of Star Coordinates provided valuable insight on several real data sets for cluster discovery and multi-factor analysis tasks. The work on Star Coordinates will continue on developing advanced transformations that will improve data understanding in multi-dimensions.

Keywords: Multi-dimensional Visualization, Data Mining, Cluster Visualization.

1. Introduction

Most scientific, engineering, and business data is multi-dimensional; i.e. datasets contain typically more than three columns of data. The use of visualizations to make sense of data is a powerful technique to support users' decision-making activities by stimulating visual thinking. Researchers have proposed a number of approaches to visualizing multi-dimensional datasets. Among the pioneering work is Bertin's Permutation Matrices [1] in which data are visualized in rows and columns of cells containing simple graphical depictions. By rearranging rows and columns, users try to form clusters –typically on the diagonal of the matrix– to understand the distribution of data. Chernoff's [2] use of faces to represent multi-dimensional data is also among the most well known work in the area. Tufte [3] also provides many compelling examples of multi-dimensional visualizations. More recent work include Inselberg's Parallel Coordinates [4,5], Worlds within Worlds [6], Table Lens [7], VisDB [8], Dynamic Queries [9], Attribute Explorer [10], and [11,12,13,14,15].

The motivation of the Star Coordinates work is to find easy to understand multi-dimensional visualizations that support users in the early stages of their data-understanding tasks. The purpose is thus not numerical analysis but to gain insight. It is assumed that once the user has a good overall understanding of the data, they will know where to look for numerical details for further analysis. Thus, it is expected that users are likely to tolerate loss of information in the initial steps of the data understanding process. Then, through dimensionality reduction and use of other visuals to represent data they can numerically support the knowledge they extracted previously. The use of simple visuals, uniform across dimensions is crucial in the early stages, as users need to explore and compare a number of different options rapidly.

The inspiration for the present work came from Bertin's Permutation Matrices (PM), which allows users to *rearrange* rows and columns to discover patterns and clusters from *coarse* graphical depictions of data. In PM, data in each cell are represented using simple visuals. One implementation of PM uses only black and white colored cells to indicate significant values. Using this simple but coarse representation, users can grasp the distribution of data and clusters without needing exact data values. Users can tolerate loss of information for the sake of gaining insight into the data.

Another source of inspiration was Parallel Coordinates, which intelligently *positions coordinates* in parallel, *treating each dimension uniformly*. As a result of this, users can make comparisons among dimensions easily. Uniform treatment becomes critical as the number of dimensions increase, as it is difficult if not impossible to make comparisons when dimensions are mapped to different visuals (shape, color, etc.). In Parallel Coordinates each data element is represented as a line passing through the coordinate axes (at the value of the element for that dimension) as opposed to its dual representation, point. While Parallel Coordinates is a very powerful technique -especially for modeling relations- these visualizations require user expertise and knowledge of mathematical methods.

This paper is organized as follows. First, the basic mathematics behind Star Coordinates is described, followed by interface features of the current implementation. Second, several data-understanding scenarios for cluster discovery and multi-factor analysis are presented using real datasets. Next, Star Coordinates is compared with related work summarizing of the strengths and weaknesses of Star Coordinates. The paper concludes with a sketch of future work.

2. Star Coordinates

Inselberg points out very appropriately that orthogonality uses up space rapidly [5]. On the other hand what orthogonality provides to the users, for example in two-dimensional scatter plots, is the ability to easily find exact data values for each point by simple projections over the coordinate axes. This task becomes more difficult in three dimensions. In fact on a static 2d projection of a 3d scatter plot, a data point may correspond to any data value parallel to the view axis. Only with the aid of interactive 3d transformations such as rotations and translations can users make sense of the data distribution. Star Coordinates in principle attempts to extends this idea to dimensions higher than 3.

2.1. Basics

The basic idea of Star Coordinates is to arrange the coordinate axes on a circle on a two-dimensional plane with equal (initially) angles between the axes with an origin at the center of the circle (Figure 1). Initially, all axes have the same length. Data points are scaled to the length of the axis, with the minimum mapping to the origin and the maximum to the other end of the axis. Unit vectors are calculated accordingly.

The Star Coordinate (SC) system is basically a curvilinear coordinate system, which can be formally mapped to the Cartesian Coordinates (CC) by defining a two-dimensional point representing the origin $O_n(x,y) = (o_x, o_y)$ and a sequence of n two-dimensional vectors $A_n = \left\langle \vec{a}_1, \vec{a}_2, \ldots, \vec{a}_i, \ldots, \vec{a}_n \right\rangle$ representing the axes. The mapping of a data element (D_j) from a dataset D to a point (P_j) in the two-dimensional Cartesian Coordinates is determined by the sum of all unit vectors $\vec{u}_i = (u_{xi}, u_{yi})$ on each coordinate multiplied by the value of the data element for that coordinate, as shown below:

$$P_j(x,y) = \left(ox + \sum_{i=1}^{n} u_{xi} \cdot (d_{ji} - \min_i), oy + \sum_{i=1}^{n} u_{yi} \cdot (d_{ji} - \min_i) \right)$$

where,

$$D_j = (d_{j0}, d_{j1}, \ldots, d_{ji}, \ldots, d_{jn}), \quad |\vec{u}_i| = \frac{|\vec{a}_i|}{\max_i - \min_i},$$

$$\min_i = \min\{d_{ji}, 0 \le j < |D|\}, \max_i = \max\{d_{ji}, 0 \le j < |D|\}$$

This is simply an extension of typical 2d and 3d scatter-plots to higher dimensions with normalization. However, it introduces some ambiguity, as do 3-dimensional visualizations. A single data point may correspond to a number of data values. The approach taken here is to provide operations on the visualization that will help users resolve these ambiguities as it is done in most 3-dimensional visualizations. Large scattered real datasets also decrease this ambiguity problem by the way data is distributed. Clusters or patterns in the datasets are preserved in the resultant visualization. In the following section some basic operations that increase users' data understanding are described.

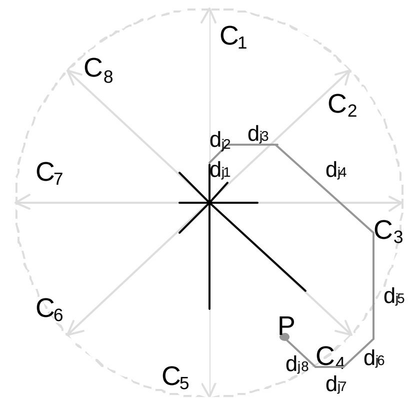

Figure 1. Calculation of data point location for an 8-dimensional dataset.

2.2. Operations

Users can apply a number of transformation and selection operations on the visualization. The set of transformations is currently limited to single- and multi-axis scaling and rotation.

Scaling transformations allow users to change the length of an axis, thus increasing or decreasing the contribution of a particular data column on the resultant visualization. This makes it a natural interaction to collapsing and expanding hierarchical clusters. To scale users simply pick the end point of an axis and push or pull towards or away from the origin. The data values are remapped to the new axis length.

Rotation transformations change the direction of the unit vector of an axis, thus making a particular data column more or less correlated with the other columns. To rotate, users pick the axis from any point on it and drag to set the new direction to be the vector from the origin to the drag point. Rotation changes only the direction of the unit vector. Both rotation and scaling can be applied on multiple selected axes to examine the combined effects of multiple columns at once.

Users can also query data values of a particular data point by moving the mouse over the point, which displays all corresponding data values. Users can mark data points by either selecting individual ones or by selecting all in a rectangular area. Marked data points are painted in a different color making them easier to follow when new transformations are applied thereafter. This is a useful operation for examining how clusters are redistributed under new parameters as a result of the transformations.

Users can also select value ranges on one or more axes and mark the corresponding data points in the visualization. This operation allows users to understand how particular factors play a role in the distribution of data.

3. Applications

Star Coordinates has been evaluated on a number of real datasets. It has been found to be particularly useful in gaining insight into hierarchically clustered datasets. Figure 2 shows an example dataset that contains car specs (e.g. mpg, cylinders, weight, acceleration, displacement, origin, horsepower, year, etc.) on approximately 400 cars manufactured world-wide. After playing (scaling, rotating, turning off some coordinates) for a while with the "cars" dataset the user easily discovers that there are 4 major clusters in the data as shown in the top of the figure.

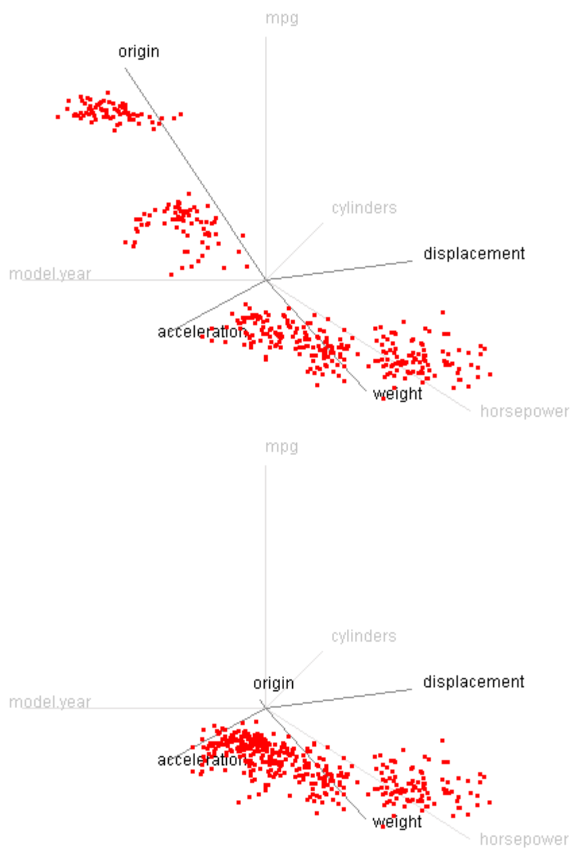

Figure 2. Cluster analysis on car specs.

Scaling the "origin" coordinate moves only the top two clusters, which indicates that these clusters represent the origin of the cars, specifically European and Japanese cars. Down-scaling the origin further reveals that these two clusters join one of the other clusters (American-made cars of similar specs.) forming a new cluster, which can be identified as low weight, low displacement, high acceleration cars. The remaining cluster thus represents American-made heavy, low acceleration, high displacement cars. Within few minutes users can identify how the data is clustered and gain an understanding of the basic characteristics of these clusters.

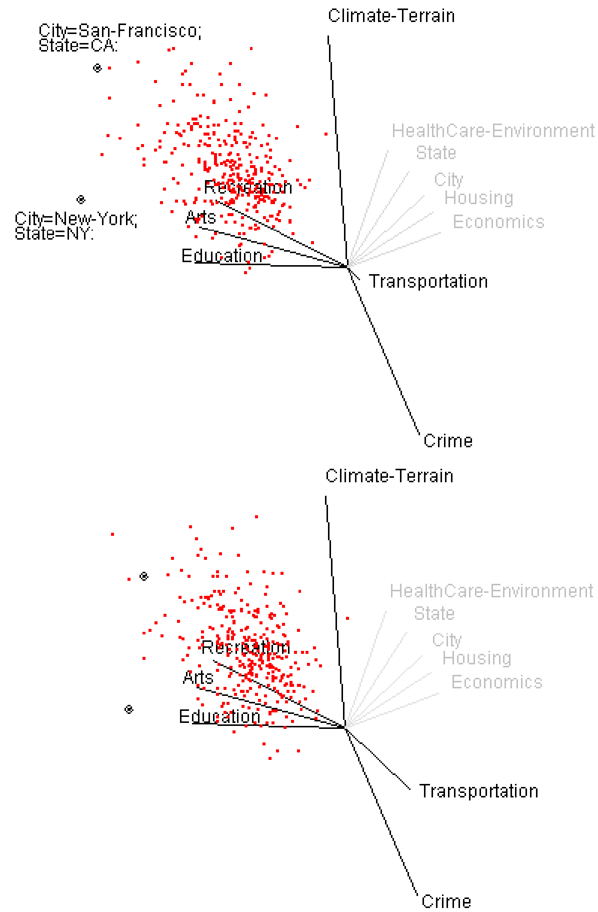

Figure 3. Multi-factor analysis on city criteria.

Another application where Star Coordinates proved useful is multi-factor analysis for decision-making. The example in this case is the "places" dataset, which contains ratings of major American cities with regards to a number of criteria such as climate, transportation, housing, education, arts, recreation, crime, health-care, and economics. The user arranges the coordinates in such a way that the coordinates for important desirable factors are pulled together in one direction and negative factors in the opposite direction. In the example in Figure 3, recreation, arts, education are grouped together with about the same importance (i.e. scaling factor). Climate is given special attention by increasing its scaling factor, yet it is made almost perpendicular to the other factors. This makes it easier to distinguish points with regards to the effect of the climate. Crime on the other hand is arranged so that it points almost in the opposite direction of the other factors indicating that it is not desirable. This layout immediately shows an outlier, namely New York City, which has all the best of the arts, recreation, etc. but high in crime and low in climate rating. San Francisco on the other hand has comparable arts, recreation, etc. but has a much better climate and lower crime. Playing further with the dataset, as shown in the lower figure, reveals that as the transportation becomes an issue other cities beat San Francisco in the combined measure. While the exact individual contributions of these factors are not immediately clear, the visualizations provide the user with an overview of how a number of factors are likely to affect the overall decision making.

4. Comparison

Due to space considerations only Parallel Coordinates is examined. Both Star Coordinates and Parallel Coordinates treat dimensions uniformly. While the space needed for Parallel Coordinates increases linearly with the number of dimensions (if inter-axis distance is kept the same), it is independent of the number of dimensions in Star Coordinates. However, as the number of dimensions increase, Star Coordinates does more compaction. In Parallel Coordinates, data elements are represented by lines showing exact data values. On the other hand, in Star Coordinates data are represented coarsely and by simpler and more space efficient points, which result in less cluttered visualizations for larger data sets.

5. Future Work and Conclusion

The work on Star Coordinates started only recently, thus there are a lot of possibilities to explore. Current experience with Star Coordinates with a number of datasets indicates that it is a viable approach to gain insight into multi-dimensional data. I would like to explore ways to improve it. One possibility is to provide users with transformations that animate the view in ways to make the data more understandable, much like rotation on an arbitrary line does in 3-dimensions. Transformations might be improved via extended trails that leave marks on the plane. Another possibility is to provide summaries of selected points and regions on the coordinate axes. These summaries will help users understand the characteristics of the selected data points (e.g. clusters, outliers, etc.) with respect to their data values. Another possibility is the colorization of data points according to their values on one or more coordinates. Once a number of such features are incorporated, I plan to run user studies to examine its viability on a number of data understanding tasks.

Acknowledgments

I would like to thank to Ismail Haritaoglu and Paul Maglio for their careful review of this paper.

References

[1] J. Bertin, *Graphics and Graphic Information Processing*, Walter de Gruyer & Co., Berlin, 24-31, 1981.

[2] H. Chernoff. The Use of Faces to Represent Points in k-Dimensional Space Graphically, *Journal of American Statistical Association*, 68, 361-368.

[3] E. R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut, 1983.

[4] A. Inselberg. Parallel Coordinates: A guide for the Perplexed, in Hot Topics Proc. of IEEE Conference on Visualization, pp. 35-38. IEEE Computer Society, Los Alamitos, CA, 1996.

[5] A. Inselberg, Multidimensional Detective. Proc. of IEEE Information Visualization ’97, pp. 100-107, 1997.

[6] S. Feiner and C. Beshers. Worlds within Worlds: Metaphors for Exploring n-Dimensional Virtual Worlds. Proc. UIST ’90, pp. 76-83,1990.

[7] R. Rao, and S. K. Card. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. Proc. CHI ’94, pp. 318-322, 1994.

[8] D. A. Keim, and H.-P. Kriegel. VisDB: Database Explorations Using Multidimensional Visualization. *IEEE Computer Graphics and Applications*, 40-49, September 1994.

[9] B. Shneiderman. Dynamic Queries for Visual Information Seeking. *IEEE Software*, 11(6), 70-77, 1994.

[10] Spence, R., and Tweedie, L., The Attribute Explorer: Information synthesis via exploration, *Interacting with Computers*, 11, 137-146, 1998.

[11] T. Mihalisin, J. Timlin, and J Schwegler. Visualizing Multivariate Functions, Data, and Distributions. *IEEE Computer Graphics and Applications*, 11(13), 28-35, 1991.

[12] W. Wright. Information Animation Applications in the Capital Markets. Proc. of IEEE Information Visualization ’95, 19-25, pp. 136-137, 1995.

[13] A. Spoerri. InfoCrystal: A Visual Tool for Information Retrieval. Proc. of IEEE Visualization ’93, pp. 150-157, 1993.

[14] J. J. van Wijk and R. D. van Liere. Hyperslice. Proc. Visualization ’93, pp. 119-125, 1993.

[15] M. Ankerst, D. A. Keim, and H.-P. Kriegel. Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets. Proc. Visualization '96, Hot Topic Session, San Francisco, CA, 1996.