Dimensional Anchors: A Graphic Primitive for Multidimensional Multivariate Information Visualizations

Patrick Hoffman Georges Grinstein David Pinkney

Institute for Visualization and Perception Research
Computer Science Department
University of Massachusetts Lowell
Lowell, MA 01854
978-934-3384
{phoffman, grinstein, dpinkney}@cs.uml.edu

ABSTRACT

We introduce a graphic primitive, called a dimensional anchor (DA), which facilitates the creation of new visualizations and provides insight into the analysis of information visualizations. The DA represents an attempt to provide a unified framework or model for a variety of visualizations, including Parallel Coordinates, scatter plot matrices, Radviz, Survey Plots and Circle Segments. A dimensional anchor is constructed by assigning values to parameters associated with various geometric graphic elements that encode the basics of the above visualizations. We define a visualization vector space in which all of the above visualizations and many new ones are represented by vectors. These encodings make it possible to perform a Grand Tour traveling from Parallel Coordinates to Survey Plot, and visiting many other visualizations in between.

Keywords

Visualization, graphics, information, dimensional anchors, multidimensional, multivariate.

1. BACKGROUND

There has been some previous work on graphic primitives for visualizations [3], [9], [10], [11]. Our dimensional anchors (DAs) sit at a higher conceptual level than Bertin's graphical language of marks (points, lines, and areas) and are not meant to create visualizations automatically for presentations, such as MacKinlay's primitives which encode "expressiveness" and "effectiveness". There are aspects of dimensional anchors that are similar to other graphic primitives that map or encode aspects of the data to graphical marks or objects. For example, DAs share some functionality with the "encoders" used in Sage [10] and could be classified as a VMTO (Visual Mapping Transformation Operator) as defined by Chi and Riedl in their taxonomy [4].

The data sets used in the examples that follow are the Car and Fischer Iris flower data sets, both from the UC Irvine Machine Learning Repository, and the Simple Seven data described in [6].

2. BASE VISUALIZATIONS

Four base visualizations can be encoded with the dimensional anchor parameters: scatter plot, Parallel Coordinates [7], Survey Plot, and the Radviz spring visualization. Survey plots are similar to the Permutation Matrix [3], except the lines representing data values extend outward from a center point instead of being left justified. Radviz, described in [5], attaches to each data point fixed springs each of which is also attached at points around a circle. The springs represent dimensions of the data. The data points are displayed at the position where the sum of the spring forces is zero. The spring force K for each spring is the value of the data point for that dimension. Circle Segments is described in [1]. We use a modified version where one dimension using a color scale is interspersed among the other dimensions, which use a gray scale.

3. DIMENSIONAL ANCHORS

3.1 Parameters of Dimensional Anchors

We have selected nine parameters to describe the graphics properties of each dimensional anchor. This small set of parameters can specify the four base visualizations. A dimensional anchor is a geometrical primitive, and is typically configured as a straight line. An *anchorpoint* is defined as the coordinate value or position on a dimensional anchor corresponding to the data value of the mapped data column; this is similar to the X or Y coordinate values in a graph or scatter plot. Further descriptions will be detailed in the examples below.

The nine DA parameters we selected are:

p_1: size of the scatter plot points

p_2: length of the perpendicular lines extending from individual anchorpoints in a scatter plot

p_3: length of the lines connecting scatter plot points that are associated with the same data point

p_4: width of the rectangle in a survey plot

p_5: length of the parallel coordinate lines

p_6: blocking factor for the parallel coordinate lines

p_7: size of the radviz plot point

- p_8: length of the “spring” lines extending from individual anchorpoints of a radviz plot
- p_9: the zoom factor for the “spring” constant K

3.2 Geometry of Dimensional Anchors

A dimensional anchor is usually configured as a straight line, but it could also be a polyline or an arbitrarily shaped curve. The arrangement of some number of dimensional anchors determines the basic layout of the visualization, which we define as the *visualization geometry*. By limiting the DA geometries to simple curves, we have a definition, or specification, of a visualization that is simple yet powerful enough to generate many of the standard multidimensional visualizations used today. Other novel visualizations can also be generated.

3.3 Basic Dimensional Anchor

At its simplest, a dimensional anchor can represent one of the axes in a two-dimensional scatter plot. It is associated with a dimension or variate from a data set. The data values for the associated dimension are mapped to the axis in the standard manner where the minimum and maximum values usually correspond to points near the ends of the axis. Labels and scale tick marks can also be associated with the dimensional anchor. These regular spacings along an axis are normally called the coordinate values. Mapped data points, which we call *anchorpoints*, represent the coordinate values (points along a dimensional anchor) that correspond to the distribution of the data points for the column associated with the dimensional anchor. A simple example is a one-dimensional anchor with lines extended from the anchorpoints (see Figure 1). The vertical colored lines show the distribution of the data (the Cars data set) for the miles per gallon (mpg) attribute. The colored lines emanate from anchorpoints, whose coordinate values correspond to the value of the data point being represented. For example low mpg cars are on the left and high mpg cars are on the right.

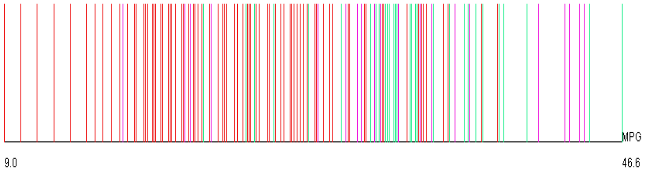

Figure 1 A single dimensional anchor - miles per gallon - color type of car

The data column associated with a dimensional anchor is called the *DA Data Column*. Several other columns and parameters can impact the representation of the display.

The *DA Color Column* determines how the lines, marks, or polygons generated from a DA are colored. For example in Figure 1 the lines are colored according to the type column (American - red or dark, Japanese - green or light, and European - purple).

The *DA color scale* is the color scale used by the DA.

The *DA Sort Column* determines how the anchorpoints are sorted along a DA. It is the same as the *DA Data Column* for most situations, but is usually different when used to represent the Survey Plot visualization.

The nine parameters associated with a DA form a vector *P* and control how the DA interacts with other DAs to form graphical constructs such as points, lines and advanced visualizations. A DA vector and geometry arrangement define a specific visualization. Several examples will now be presented.

3.4 The Scatter Plot Parameters

We have defined three DA parameters (p_1, p_2 and p_3) associated with the construction of scatter plots. One possible construction of a scatter plot requires that a perpendicular line extend outward from an anchorpoint on a DA. An anchorpoint associated with the same data point on another DA (another column of the same data set) also has a perpendicular line extending outward. If the two perpendicular lines meet, the point of intersection becomes the plotted point of the scatter plot. It should be noted that the perpendicular lines extend outward from *both sides of the DA*. For an example of perpendicular lines extending out on both sides of a DA and intersecting to form scatter plot points, see Figure 4 through Figure 6. Parameter p_1 controls the size of the scatter plot point and ranges from 0 (drawing no point) to 1 (displaying a maximum size point). The size, shape or color of a point in a scatter plot can be associated with other dimensions in the data set. In our implementation the *DA Color Column* determines the color of the point, and the shape of the point is fixed as a circle. The intersection of perpendicular lines extending from anchorpoints works with any arrangement or any number of dimensional anchors. For example in a Parallel Coordinates arrangement, all of the perpendicular lines extending from the DAs are parallel and do not usually intersect. If they do, it is because data values are identical when normalized.

Parameter p_2 controls the length of the perpendicular lines extending from the DA to the scatter plot point. p_2 ranges from 0 (no lines drawn) to 1 (a full line drawn from the DA anchorpoint to the scatter plot point). The intersection of these perpendicular lines, whether drawn or not, defines the position of the scatter plot point.

In Figure 2, the Iris flower data set is shown using Sepal Length along the horizontal DA (X-axis) and Petal Width along the vertical DA (Y-axis). The size of the scatter plot point is close to the maximum (p_1 = .8). The lines extending from the anchor points are controlled by p_2 and are only 20% of the maximum length. Tufte [12] suggested that these line extensions be included to enhance the utility of a scatter plot by showing the distribution of data along each axis. In Figure 3 the same data and the same arrangement of DAs is shown, but this time p_1=.1 (producing small points) and p_2=1.0 (resulting in fully drawn lines).

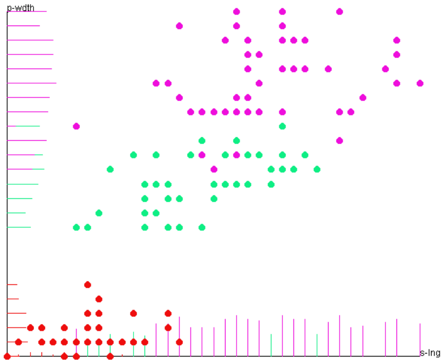

Figure 2 DA scatter plot using two DAs P = (0.8, .2, 0, 0, 0, 0, 0, 0, 0)

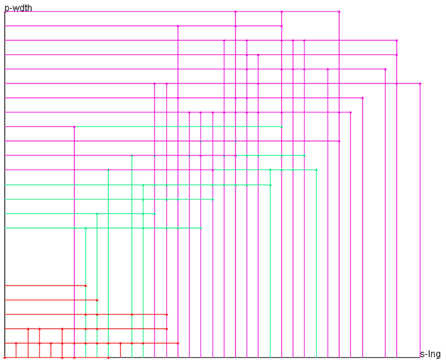

Figure 3 Two DA scatter plot P = (0.1, 1.0, 0, 0, 0, 0, 0, 0, 0)

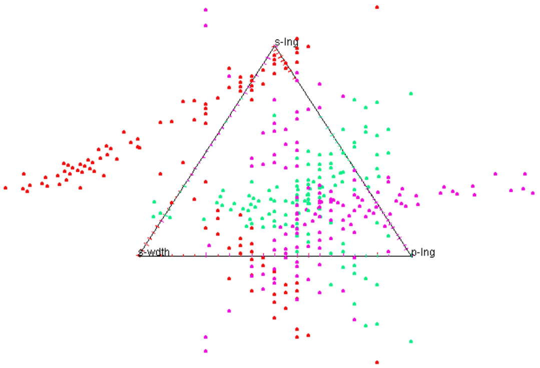

Figure 4 Three DAs in an equilateral triangle, Iris data set – A three-dimensional scatter plot P = (0.6, 0, 0, 0, 0, 0, 0, 0, 0)

In addition to the perpendicular configuration, dimensional anchors can be arranged arbitrarily such as polygonally, radially, or in an L-shape or crisscross pattern. In Figure 4, an equilateral triangle arrangement of three DAs is shown.

When N dimensional anchors are used (in a data set with N variates or dimensions), there may be up to N points plotted for each data point in the visualization using the scatter plot parameters. It sometimes helps the visualization to have these N points connected. Figure 4 is an example of three dimensional anchors arranged in an equilateral triangle pattern. Here the data is the Fischer Iris data and the three colors represent the three types of flowers. This three-dimensional scatter plot uses three DAs, resulting in three display points generated for each data point (at the intersection of the perpendicular lines extending from the DAs). Notice that intersections (and therefore points) can occur outside the triangle.

Parameter p_3 controls the length of the lines connecting all displayed scatter plot points associated with one real data point (data record). In Figure 5 the display points associated with the same data point are connected (p_3=1.0). In general, both p_2 and p_3 generate N-sided polygons if the DAs are configured as regular polygons.

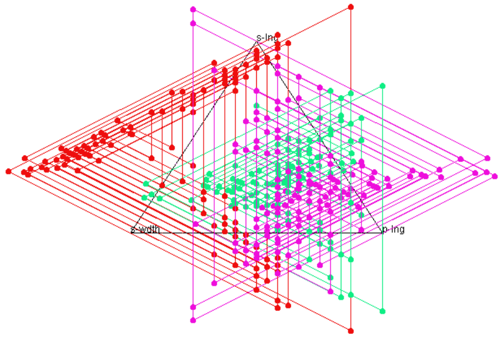

Figure 5 Three DAs with P = (.6, 0, 1.0, 0, 0, 0, 0, 0, 0)

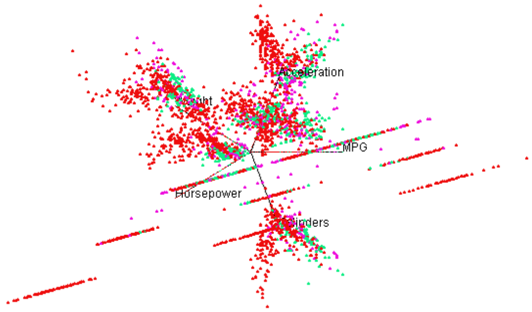

Figure 6 A five DA scatter plot – radial layout P = (.5, 0, 0, 0, 0, 0, 0, 0, 0)

Several visualizations, generated from various arrangements of DAs and using the scatter plot parameters, are shown in Figure 2 though Figure 8. It is interesting to note that discrete variables, such as the Cylinders and Year dimensions, produce distinct straight lines. Also, most of the scatter plot points remain inside the polygon in the 6 DA plot (Figure 7). The reason is that the three pairs of DAs are opposite and parallel to each other in the hexagon. Since they are parallel, perpendicular lines extending from them will not meet (unless the anchor points are identical), and therefore no scatter plot points will be drawn.

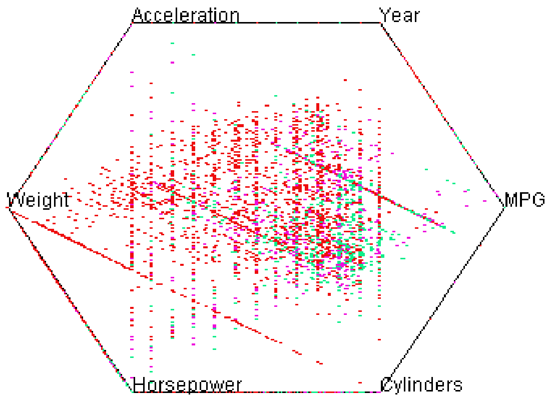

Figure 7 A six DA Scatter plot - P = (.5, 0, 0, 0, 0, 0, 0, 0, 0)

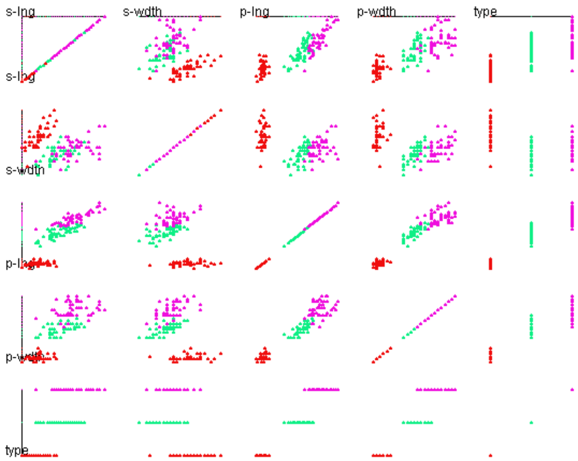

Figure 8 Five DAs x 2 in an L-shape generate a scatter plot matrix - P = (.5, 0, 0, 0, 0, 0, 0, 0, 0)

At first thought it would seem that one might need $2N^2$ DAs to generate the standard scatter plot matrix, since two dimensional anchors are needed for a scatter plot. However, one simply has to lay out 2N dimensional anchors in an L-shaped pattern to generate N^2 scatter plots (see Figure 8).

3.5 The Survey Plot Parameter

The parameter p_4, along with the *DA Sort Column*, is used to construct a Survey Plot [8] or a visualization similar to Circle Segments [1]. The p_4 parameter scales the size of a rectangle extending from an anchorpoint. The size is also based on the dimensional value at the anchorpoint. In this display, the anchorpoints are sorted along the DA according to the *DA Sort Column*. There is a threshold value for the maximum size of the Survey Plot rectangles. This threshold depends on the orientation of all the other dimensional anchors. The heuristic is that a DA rectangle can not intersect another DA rectangle. The rectangles can be scaled by the p_4 parameter up to this limit.

The Survey Plot and modified Circle Segments visualizations can be constructed by using the p_4 parameter, with its constraints, and an appropriate arrangement of dimensional anchors. The circle segments' arcs become straight lines, and extend out to a regular polygon instead of a circle, but the essentials of the visualization are still the same. See Figure 9 through Figure 13 for Survey Plot and Circle Segment like constructions.

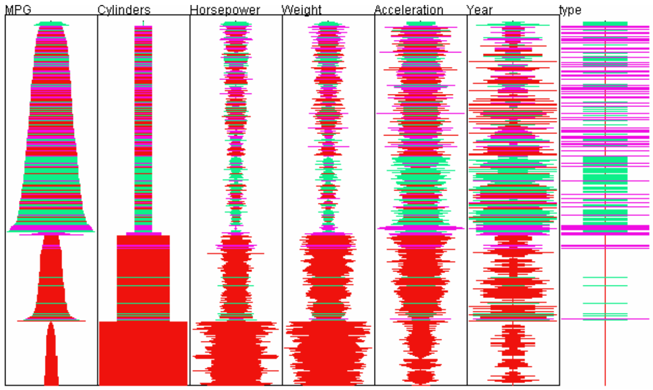

Figure 9 A Survey Plot generated from seven DAs

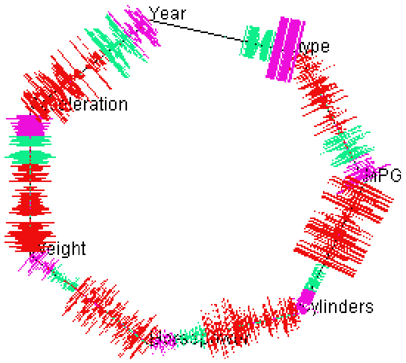

Figure 10 A Survey Plot, regular polygon configuration

Figure 11 through Figure 12 show the construction of a visualization similar to Circle Segments (Figure 13). First, we intersperse the classification dimension (type) with each of the seven Car data set dimensions. The columns are then sorted according to the classification dimension (*DA Sort Column*=type). We use a gray scale mapping for the seven dimensions and a rainbow color mapping for the classification dimension. Finally, by varying the p_4 parameter, the following visualizations are obtained. Figure 11 shows the visualization with a nominal value of .4 for p_4. In Figure 12 p_4=1, and the visualization is called CCCViz for Color Correlated Column Viz. In CCCViz, one can see whether a dimension (gray scales) visually correlates with a particular classification dimension (color scale). In this case, a correlation is seen in mpg, cylinders, horsepower, and weight with the American (red or dark) cars. The CCCViz and the Circle Segment visualization (Figure 13) are closely related. Some variations on CCCViz, such as reducing the width of the color columns, only having two color columns, or drawing horizontal lines at the class boundaries, could potentially enhance the visualization.

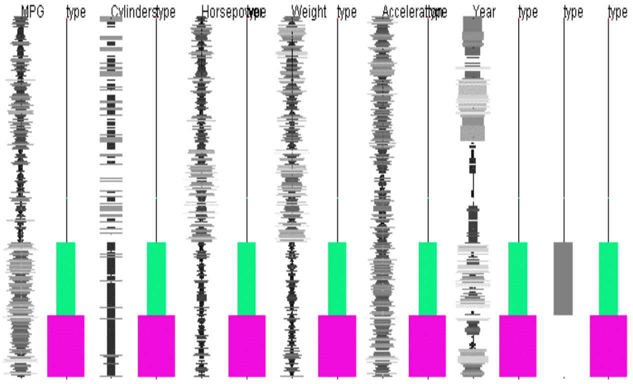

Figure 11 DAs alternated with an identical class DA - gray scale and color scale P = (0, 0, 0, .4, 0, 0, 0, 0, 0)

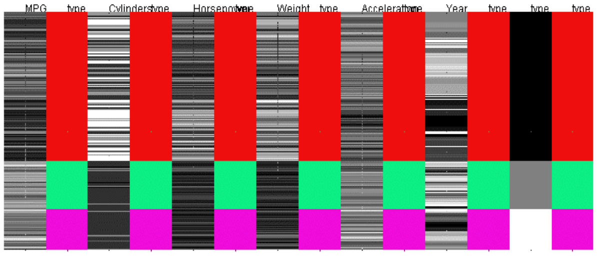

Figure 12 CCCViz DAs with P = (0, 0, 0, 1.0, 0, 0, 0, 0, 0)

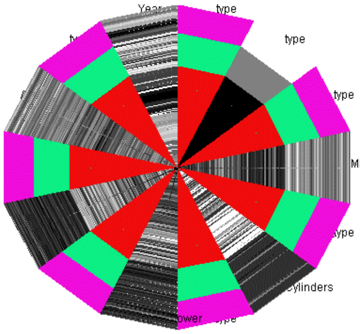

Figure 13 DAs in a radial pattern to generate Circle Segments P = (0, 0, 0, .4, 0, 0, 0, 0, 0)

3.6 Parallel Coordinate Parameters

The simplest concept of dimensional anchors and anchorpoints arises from observing the construction of the Parallel Coordinates [7] visualization. The dimensional anchors are arranged vertically and parallel, and a line from one DA anchorpoint is drawn to another. The length of these connecting lines is controlled by the DA parameter (p_5). If all anchorpoints on all DAs are exhaustively connected (each anchorpoint connected to N-1 other anchorpoints), some interesting visualizations are generated (see Figure 14 and Figure 15). We term these fully connected DA visualizations *Mesh Plots*. To transform Figure 14 into a standard Parallel Coordinates visualization, an additional parameter must be defined. Parameter p_6 represents how many DAs a p_5 connecting line can cross. When set to 0, this blocking parameter leads to the familiar Parallel Coordinate visualization. Parallel Coordinates in a circle (superimposed Star Glyphs) can be generated using the p_5 and p_6 parameters when the DAs are arranged as spokes radiating from the center of a circle (as in Radviz).

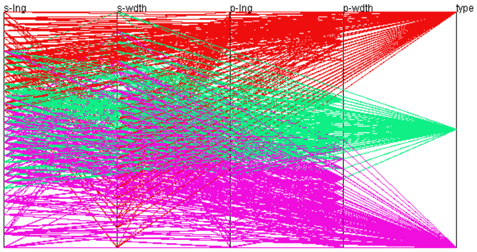

Figure 14 DAs in PC configuration P = (0, 0, 0, 0, 1.0, 1.0, 0, 0, 0) Mesh Plot

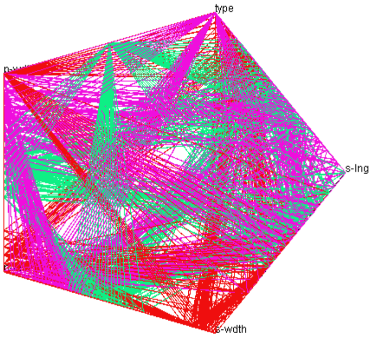

Figure 15 DAs in a regular polygon P = (0, 0, 0, 0, 1.0, 0, 0, 0, 0) Mesh Plot

3.7 Radviz Spring Force Parameters

If the anchorpoints along a dimensional anchor are considered to be fixed points where invisible springs are attached to a movable data point, then a visualization similar to Radviz can be created. Parameter p_7 is used to control the size of the point that is placed in the display where the spring force sum is 0. An "enhanced" Radviz display is created when dimensional anchors are arranged as an N-dimensional regular polygon. One of the limitations of the original Radviz is that data points with different data or dimensional values can overlap in the center of the circle. When the fixed spring points are spread out along the DA in a regular polygon, the chance of points overlapping is reduced because the springs are less likely to be aligned (with respect to the data point). If the DAs are compressed to points and are uniformly distributed around the circumference of a circle, the original Radviz display is created. The enhanced DA Radviz is a more "efficient" visualization, in that it better utilizes the total area of the circle (or regular polygon). At higher dimensions (greater than 30) the anchorpoints are compacted and the visualization becomes similar to the original Radviz. Parameter p_8 is used to draw lines extending from the spring anchorpoints to the displayed point. Parameter p_9 is used as a zoom factor in the display. The zoom factor is simply the scaling factor used when calculating the spring force. The force is summed using the data point values for each dimension. Figure 16 shows an example of the original Radviz with three overlapped points (located near the center of the circle). In Figure 17, an enhanced Radviz visualization is shown with spring lines drawn (p_8=1.0), and all seven points are now visible. Figure 17 is an example of a "spread" polygon, in which the sides of the polygon do not touch. The spread polygon configuration separates the DAs further than a regular polygon and has the advantage of resolving ambiguities at the intersections of the DAs.

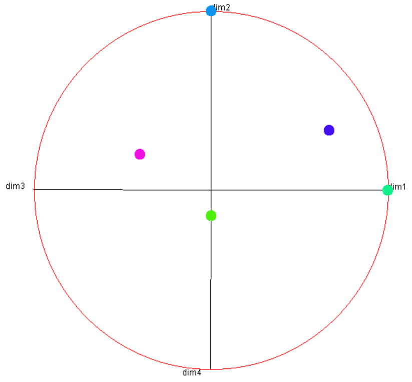

Figure 16 Original Radviz - three overlapping points

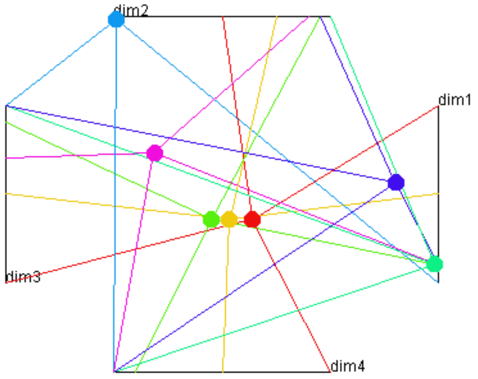

Figure 17 DAs spread polygon P = (0, 0, 0, 0, 0, 0, .5, 1.0, .5)

3.8 Dimensional Anchor Layout

So far we have described the parameters associated with dimensional anchors that make them capable of creating visualizations. The fact that DAs can be arranged with any arbitrary size, shape or position gives them flexibility and permits a large variety of visualization designs. By definition, a dimensional anchor can be any sequence of curves, such as line segments, an arc or a Bezier curve. With this definition, perpendicular line extensions can still be constructed. We have only investigated some standard arrangements, such as those used by Parallel Coordinates, Circle Segments or Regular Polygons. There are many possible arrangements of DAs on a display, such as curves arranged in the form of a lens. Hyperbolic circular displays could have unique data mining features. Perhaps these configurations could help produce visualizations focused on specific data patterns. Dimensional anchors shaped in the form of polynomial or logarithmic functions might also have useful properties.

Various arrangements of DAs can produce a partial ordering of the data. For example, the visualization in Figure 18 was made by eight DAs arranged in a crisscross pattern on the Iris Flower data set. Only the spring parameters p_7 and p_9 are used, and the display results in a simple diagonal pattern, since every "spring" has an opposite spring symmetrically across the diagonal. The particular order of the points in this display is not clear. It is not the same as four simple DAs using p_7 and p_9 in a straight line, as in Figure 19. The eight DAs in a crisscross pattern form a good discriminator of the Iris flower classes, whereas the four DAs in Figure 19 do not. This spring ordering, resulting from the configuration of the dimensional anchors, requires further investigation.

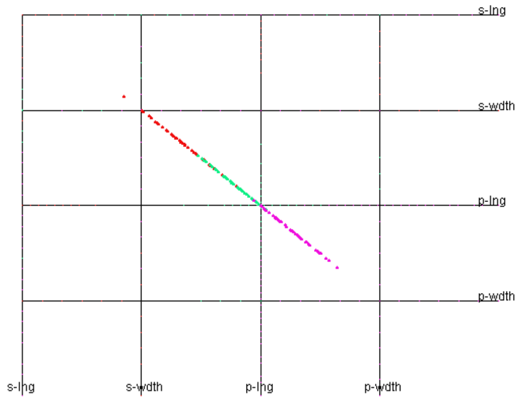

Figure 18 DAs arranged in a crisscross pattern P = (0, 0, 0, 0, 0, 0, .4, 0, .5)

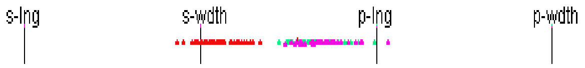

Figure 19 Four small DAs in a Parallel Coordinate configuration P = (0, 0, 0, 0, 0, 0, .4, 0, .5)

A visualization made with Bezier curves is shown in Figure 20. A combined visualization with scatter plot, survey plot and spring parameters active is shown in Figure 21.

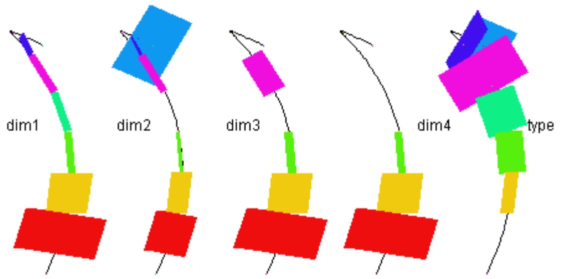

Figure 20 Survey Plot parameters with curved DAs

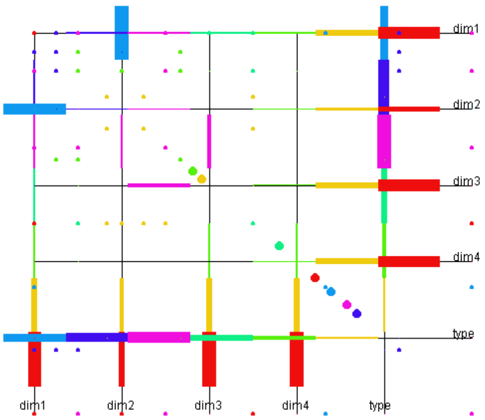

Figure 21 A crisscross pattern of DAs

4. COMBINATIONS OF VISUALIZATIONS

By treating the DA parameters as coordinates of vectors in a vector space, we can define linear combinations of visualizations, including interpolations between two visualizations. However, the combinations do not take into consideration the geometry of the dimensional anchors. If the DA geometry arrangement were parametized into a vector, a true interpolation between visualizations could be performed. Some of this work has been done [6] where we intuitively define DA arrangements that are interpolations of visualizations. For example, the radial spoke arrangement in Figure 22 can be defined as an arrangement in between Parallel Coordinates and Radviz. This visualization combines spring points and shortened Parallel Coordinate lines (if longer lines are used, the spring points are mostly obscured). This visualization combines features from both of the base visualizations. The Radviz clustering of the three types of cars (American, Japanese and European) is easily seen. The red points or darker points (American cars) are pulled more toward higher horsepower and cylinders. Features from Parallel Coordinates, such as the discrete values of the cylinders and year dimensions, are clearly seen along the DA. With the shortened Parallel Coordinate lines, one can get an idea of the distribution of each variate along the DAs (especially missing values).

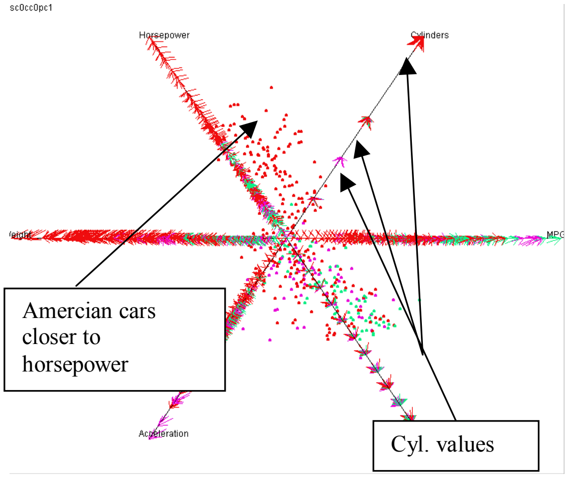

Figure 22 A PC + RV combination on the Car data set

Figure 23 is a variation of Radviz in a spread polygon arrangement with partial spring lines shown. This visualization (called X-Radviz) seems to provide the most information of any Radviz variant by reducing the point overlap and "showing" the dimensional data distribution with the spring lines. It can also be considered a combination of Parallel Coordinates and Radviz.

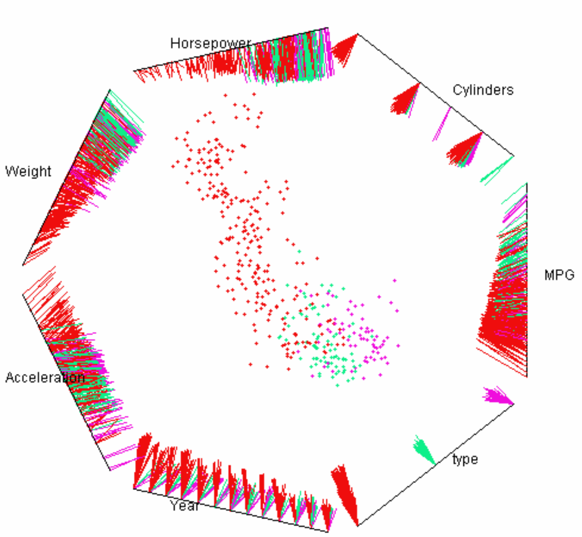

Figure 23 X-Radviz on the Car data set

5. THE VISUALIZATION SPACE

In the DA examples presented here, we have used 9 parameters that are shared by all the dimensional anchors in a visualization. This defines the size of our visualization space as at least $\Re^9$. Including the geometry of the DAs, and assuming we have at least three parameters defining the geometry, the size of our visualization space is at least $\Re^{12}$. It is possible to take a "Grand Tour" [2] through this visualization space. By varying the 9 parameters and animating the arrangement of the dimensional anchors, one can move through this space and look at an incrementally defined subset of the infinite visualizations possible. The previously described visualizations demonstrate a limited manual tour that has proved useful in finding new visualizations. We have implemented several tours with different data sets, with over 1400 visualizations in each tour. Some new visualizations and insights have resulted from these tours. For example, the spread polygon pattern in Figure 24 was discovered by a Grand Tour of an exon/intron data set. The visualization is a widely spread out polygon with scatter plot points and lines enabled. The partial scatter plot "slices" show a noticeable (darker/green) triangle in the center that has an underlying biological explanation, not easily found in other visualizations. Some Grand Tours of dimensional anchor visualization spaces are presented in [6].

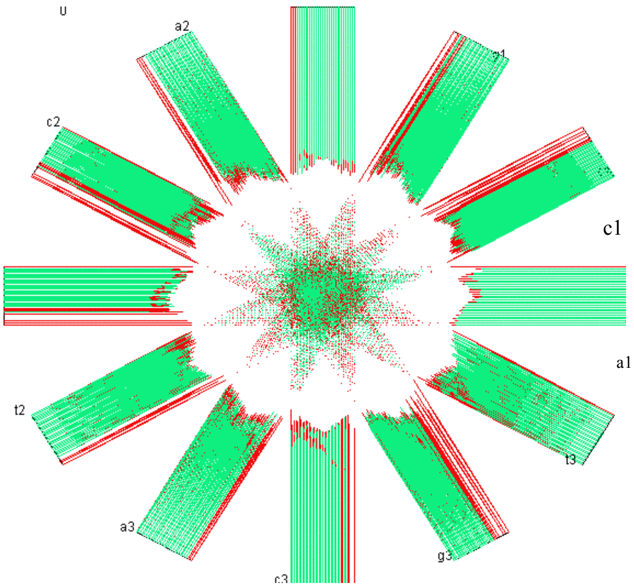

Figure 24 A spread polygon with scatter plot parameters - Exons/Introns

6. CONCLUSION

The Dimensional Anchor mechanism represents a new way of looking at information visualizations. It provides a formal representation of visualizations that supports creating a myriad of new visualizations which can then be viewed using a "grand tour" through that visualization space. We believe that with additional parameters and some unique DA arrangements, many more "standard" visualizations such as multi-line graphs, dimensional stacking, iconographics, and Kohonen self organizing maps can be generated from enhanced dimensional anchors. With these additional "base" visualizations, the DA visualization space will expand greatly and be ready for exploring.

New visualizations discovered from the grand tours are currently being evaluated. These and the previously described visualizations will be investigated and analyzed for their usefulness in visual data mining. Ongoing benchmarking and evaluation of various information visualizations will continue, especially in regards to which visualizations are best for classifying, clustering, outlier detection and other data mining features.

7. ACKNOWLEDGMENTS

This work is funded by the Institute for Visualization and Perception Research at the University of Massachusetts Lowell, and in part by grants from Pfizer Pharmaceutical and from the National Institute for Standards and Technology.

8. REFERENCES

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

[2] Asimov D., ``The grand tour: a tool for viewing multidimensional data", SIAM Journal on Scientific and Statistical Computing, 6 (1985) 128-143.

[3] Bertin J.:Semiology of graphics, W.J. Berg, Tr. Madison,WI: The University of Wisconsin Press, 1983.

[4] Chi E. H., and Riedl J. T., "An Operator Interacton Framework for Visualizations Systems", IEEE Visualization '98 - Info Vis Symposium, 1998.

[5] Hoffman, P., Grinstein, G., Marx, K., Grosse, I., Stanley, E., "DNA Visual and Analytic Data Mining", IEEE Visualization '97 Proceedings, pages 437-441, Phoenix, AZ, 1997.

[6] http://www.cs.uml.edu/~phoffman

[7] Inselberg A., "The plane with parallel co-ordinates", The Visual Computer 1, pp. 69-91, 1985.

[8] Lohninger H.: "INSPECT, a program system to visualize and interpret chemical data.", Chemomet. Intell. Lab. Syst. 22 (1994) 147-153 (http://qspr03.tuwien.ac.at/lo/)

[9] MacKinlay J., Automating the Design of Graphical Presentations of Relational Information. ACM Transactions on Graphics, Vol. 5, No. 2, April 1986, pp. 110-141.

[10] Roth, S. F., The SAGE Project. http://www.cs.cmu.edu/Web/Groups/sage/sage.html

[11] Senay, H., and Ignatius. E., A knowledge-based system for visualization design. IEEE Computer Graphics and Applications, pages 36-47, November 1994.

[12] Tufte, E. R., The Visual Display of Quantitative Information, Graphics Press, 1983.