

# High-Dimensional Visualization Using Continuous Conditioning

William C. Wojciechowski

Department of Statistics  
Rice University  
Houston, TX 77005-1892

David W. Scott

Department of Statistics  
Rice University  
Houston, TX 77005-1892

## Abstract

Many methods for visualizing hypervariate data have been employed. A general technique is to display a subset of the total variable space. The variable space is divided into two disjoint sets. The graphing subspace and the conditioning subspace. Only the variables in the graphing subspace are displayed. A subset of the conditioning space is defined and an observation is visible only if it is an element of the subset. The subset is usually a point or box and an indicator function is used to determine if an observation appears on the graphic. Continuous conditioning provides several extensions. First, there are no restrictions on the subset. Second, the indicator function is replaced by a general distance function. Finally, any visual characteristics, such as color, can be determined by the distance an observation is from the subset. These generalizations allow for a wide variety of visualization methods.

## 1 Introduction

Continuous conditioning is one of many different techniques available for viewing high-dimensional data. Many other researchers have created, studied, and had success with high-dimensional graphics. An incomplete list includes Asimov (1985), Carr (1988), Chernoff (1973), Cleveland (1993), Hurley and Buja (1990), Scott (1992) and Wegman (1990). As the technique is still being developed, a comparison to other methods is not presented. Further work is needed to decide what place continuous conditioning holds in visualizing hypervariate data.

To aid in understanding our method, we categorize the characteristics of continuous conditioning in the taxonomy of Buja, Cook, and Swayne (1996). Their system classifies the characteristics of techniques and not the techniques themselves. The main division is into two categories, *rendering* and *manipulation*. In the rendering category, continuous conditioning falls in the scatterplot

subcategory. Manipulation is concerned with operating on individual plots and with organizing many plots. Since continuous conditioning is used to find features of interest, our method falls in the *finding gestalt* manipulation subcategory. Putting the two categories together, continuous conditioning is a method that renders scatterplots to find structure in a hypervariate dataset. While this implementation uses scatterplots for rendering, continuous conditioning is not limited to a particular visual medium. Furthermore, all plot methods display some features of a dataset. Therefore, gestalt finding does not separate continuous conditioning from other hypervariate visual methods. Our method is different by its generality. The manipulation procedure is generalized and the abstraction provides a mechanism to tailor the rendering and manipulation to a particular application.

To give an example, coplots (Cleveland 1993) are a specific case of continuous conditioning. Let the dataset consist of 3 continuous variables,  $X_1$ ,  $X_2$ ,  $X_3$ , and let  $X_{i,j}$  represent the value of variable  $j$  for observation  $i$ . Define the graphing subspace to be the  $X_1$  and  $X_2$  plane. Let the rendering method be a scatterplot having  $X_1$  as the abscissa and  $X_2$  as the ordinate. Let the line represented by  $X_3$  be the conditioning subspace. The conditioning subset is one of several line segments. The line segments are obtained by using the equal-count algorithm (Cleveland 1993) with  $X_3$  as the conditioning variable. Represent the set of line segments by  $S_1, \dots, S_k$  and the current conditioning subset as  $S_c$ . Let  $X_{i,3}$  be the value of  $X_3$  for observation  $i$ . The distance function is

$$\delta(X_{i,3}, S_c) = \begin{cases} 1 - I_{S_c}(X_{i,3}) & \text{if } X_{i,3} \in S_c \\ I_{S_c}(X_{i,3}) & \text{otherwise} \end{cases}$$

where

$$I_{S_c}(X_{i,3}) = \begin{cases} 1 & \text{if } X_{i,3} \in S_c \\ 0 & \text{otherwise} \end{cases}$$

The point  $(X_{i,1}, X_{i,2})$  is drawn on the scatterplot only if  $\delta(X_{i,3}, S_c) = 0$ . A separate scatterplot is drawn for each  $S_i$  ( $i = 1, \dots, k$ ) and all of the plots are arranged

in a static layout. Dynamic coplots (Wojciechowski and Harner 1995) is an animated extension. One scatterplot is rendered and animation is achieved by repeatedly setting  $S_c$  to a new line segment and drawing the corresponding scatterplot. The discontinuous nature of the indicator function prevents a smooth transition. While this is a specific example of continuous conditioning, our method can be used to enhance coplots.

The extension presented in this paper is to replace  $\delta$  with a continuous distance function and to define the transparency and possibly other properties of a point as a continuous function of the distance. A fully transparent point will not be visible, of course. The transparency of a point is then a continuous composite function with  $X_{i,j}$  and  $S_c$  as arguments. Therefore, if  $S_c$  is slightly modified, the transparency of the observations are slightly modified. A smooth animation is produced by continually modifying  $S_c$ . Furthermore, the scatterplot itself may be thought of as a composite function. As long as each function in the composition is continuous, smooth animation can be achieved by continually modifying the arguments. This abstraction may be extended by allowing any plot type to be defined as a composition of continuous functions. From this perspective, statistical graphics can be studied under continuity's proper setting, topology. This is an active area of research.

## 2 General Method

The details of continuous conditioning are presented. To begin with, the variable space is divided into two disjoint parts, the graphing subspace and the conditioning subspace. There is no *correct* separation of the variable space. As with other statistical methods, variable selection depends on the application, and the difficulty increases with the number of dimensions. The graphing subspace is displayed on the computer screen. It is not necessary to display the conditioning subspace unless it aids in comprehending the graphing subspace. A subset of the conditioning subspace is defined. Then the graphic is a continuous composite function of the subset and the observations. The subset of the conditioning subspace is constantly changed by a small increment. The definition of small increment is given in the next section. By continuity, the graphic will also change by a small increment and a smooth animation effect is produced. Selecting the subset and procedure to modify the subset is an interesting problem this area requires further work.

### 2.1 Definitions

Mathematical notation will facilitate the description of the technique and will be used in the remaining sections.

- $\Omega$  is the variable space and  $X$  is an  $n \times d$  data matrix such that  $X \subset \Omega$ .  $X$  consists of  $n$  observations and  $d$  variables.
- $\Omega^g$  is the graphing subspace of  $\Omega$  and  $X^g$  is an  $n \times d_g$  matrix, where  $X^g \subset \Omega^g$ .
- $\Omega^c$  is the conditioning subspace of  $\Omega$  and  $X^c$  is  $n \times d_c$  matrix, where  $X^c \subset \Omega^c$ .
- $X^c$  and  $X^g$  are disjoint subsets of the columns of  $X$  and  $d_g + d_c \leq d$ .
- $S \subset \Omega^c$  is the conditioning subset and  $\Omega^S$  is the space of all  $S$ .
- $\phi$  is a distance function defined on  $\Omega^S$ .
- $X_i^c$  is the  $i$ th row of  $X^c$ .
- $\delta_i = \delta(X_i^c, S)$  is the distance  $X_i^c$  is from  $S$ .
- $\pi_i = f(\delta_i)$  is a function from  $\mathbb{R}^1$  to  $\mathbb{R}^p$ , where  $p$  is the number of graphic properties depending on  $\delta_i$ .

The last item defines a statistical graphic as a continuous composite function depending on the observations and the conditioning subset. Many plot types and conditioning techniques can be implemented using this general notation.

### 2.2 Algorithm

The animation steps are summarized as: initialize  $S$ , calculate the graphic properties, draw the graphic, update  $S$ , and repeat. The explicit steps are

1. Let  $S_0$  be the initial value of  $S$ .
2. Calculate  $\pi_i$  for  $i = 1 \dots n$ .
3. Draw the graphic using the set of  $\pi_i$ 's.
4. Set  $S = S^*$  where  $S^*$  is such that  $\phi(S, S^*) = \epsilon$ .
5. Go to step 2.

As the steps are repeated, the graphic is constantly modified and an animation is produced.

### 3 Examples

Four datasets are presented. Two datasets are simulated and the others are from S-PLUS<sub>TM</sub>. Each uses a scatterplot to display both the graphing and conditioning subspaces. The datasets and animations are described, but no graphics are presented in the paper. The animations are included as Quicktime<sub>TM</sub> movies.

Two graphic properties are modified in each example, color and transparency. Therefore,  $f(\delta_i)$  is a function from  $\mathbb{R}^1$  to  $\mathbb{R}^2$  and  $\pi_i$  is a  $1 \times 2$  vector. Let  $\pi_{i,1}$  be the color and  $\pi_{i,2}$  be the transparency of the  $i$ th observation. As  $\delta_i \rightarrow 0$ , the point becomes blue and solid. As  $\delta_i \rightarrow \infty$ , the point becomes white and transparent.

Each animation has plots on the left and the right. The left plot is the graphing subspace and the right plot is the conditioning subspace.  $S$  is represented by the points that range from blue to white in the conditioning subspace. The color of the points in the graphing subspace correspond to their counterparts in the conditioning subspace. As  $S$  is updated, the corresponding graphing subspace modifications can be observed.

For each of the examples we use the following notation. Let  $X_i$  represent the  $i$ th variable and  $X_{i,j}$  represent the value of the  $j$ th variable for the  $i$ th observation.

#### 3.1 Simulated Multivariate Normal

The first example is provided to demonstrate how simulated random data appears using continuous conditioning. The dataset is simulated from a three dimensional standard normal distribution.  $X^g$  consists of  $X_1$  and  $X_2$  and  $X^c$  is  $X_3$ . The conditioning subset,  $S$ , is defined to be a one-dimensional point and  $\delta(X_i^c, S) = |X_i^c - S|$ . The distance function on the set of  $S$  is  $\phi(S_1, S_2) = |S_1 - S_2|$ , where  $S_1, S_2 \in \Omega^S$ . In practice, the range of  $S$  is between the minimum and maximum values of  $X_3$  and the distances are scaled to reflect the finite limits of the data.

The left graphic is a scatterplot of  $X_1$  and  $X_2$  and the right graphic displays  $S$ . Observing the right plot,  $S$  can be seen to *slide* between the minimum and maximum values of  $X_3$ . Points *close* to  $S$  are blue and solid. Points that are far from  $S$  are transparent and white. As  $S$  is modified, the animation is produced. Since the three variables are independent, no trends are apparent.

#### 3.2 Rotated sin Wave

This example uses simulated data to see if the continuous conditioning method can detect an obvious pattern in the data. The dataset consists of three variables.  $X_1$  is uniformly distributed on  $(0, 2\pi)$  and  $X_{i,2} = \sin(X_{i,1}) + \epsilon$ ,

where  $\epsilon$  is random error.  $X_3$  is uniformly distributed between  $(0, \pi)$  and controls the rotation of the sin wave. The graphing subspace displays  $X_1$  on the abscissa and  $X_2$  on the ordinate. The conditioning subspace is plotted as a line segment.  $S$ ,  $\delta_i$ , and  $\phi$  are the same as in the first example. If the graphing subspace is plotted without conditioning on  $X_3$ , there is no visible pattern in the data. When conditioning is used, the rotated sin wave is easily observed. The animation is in the `rotate.mov` movie file. The structure of the dataset is readily seen using a three-dimensional scatterplot. However, the example is for proof of concept. It is also recognized that simulated data containing more subtle trends should be examined.

#### 3.3 Environmental Data

For this example, the dataset is four dimensional and is taken from the `environmental` dataset found in S-PLUS<sub>TM</sub>. It shows continuous conditioning detecting a trend that is not known apriori as in the second example.  $X^g$  consists of  $X_1$ ,  $X_2$ , and  $X_3$  and  $X^c$  consists of  $X_4$ .  $S$ ,  $\delta_i$ , and  $\phi$  are the same as in the first two examples except  $X_4$  replaces  $X_3$ .  $X_1$ ,  $X_2$  and  $X_3$  are plotted in a three-dimensional scatterplot. The movie file for this example is `enviro.mov`. As the animation runs, two groups are seen. One cluster is seen when  $S$  is near the small values of  $X_4$  and another group is seen for intermediate and large values of  $X_4$ .

#### 3.4 Hamster Data

The data for this example is a five dimensional dataset taken from the `hamster` dataset in S-PLUS<sub>TM</sub>. It demonstrates a case where  $S$  is not a single point. In this example,  $S$  is an ellipse and the color and transparency of the points depends on the distance a point is from the ellipse. The ellipse corresponds to a Mahalanobis distance calculated using the sample covariance matrix and mean vector. The graphing subspace depicts three of the five dimensions and the conditioning subspace displays two dimensions.

$X^g$  consists of  $X_1$ ,  $X_2$ , and  $X_3$ . The conditioning variables are  $X_4$  and  $X_5$ . The graphing variables are plotted in a three-dimensional scatterplot and the conditioning variables are plotted in a two-dimensional scatterplot.  $S$  is defined to be

$$S = S(\rho) = \{X : X \in \Omega^c, (X - \bar{X})^T \hat{\Sigma} (X - \bar{X}) = \rho\},$$

where  $\bar{X}$  and  $\hat{\Sigma}$  are respectively the sample mean vector and sample covariance matrix of  $X^c$  and  $\rho$  is a given number. This is an ellipse corresponding to a Mahalanobis distance of  $\rho$ . The distance between  $S$  and  $X_i^c$

is

$$\delta(X_i^c, S) = |\mu(X_i^c) - \rho|,$$

where  $\mu(X_i^c)$  is the Mahalanobis distance from  $X_i^c$  to  $\bar{X}$  calculated with  $\hat{\Sigma}$ . Let  $S_1$  and  $S_2$  be two ellipses as defined above and let  $\rho_1$  and  $\rho_2$  be the respective Mahalanobis distances. The distance function on  $\Omega^S$  is defined as

$$\phi(S_1, S_2) = |\rho_1 - \rho_2|.$$

Observing the `hamster.mov` movie file, the ellipse expands and contracts as it is modified. As the ellipse moves towards its maximum expansion, it can be seen where the outliers in the conditioning subspace lie in the graphing subspace, assuming the data follows a multivariate normal distribution.

## 4 Summary and Conclusions

The continuous conditioning method has been introduced and the general algorithm was presented. Although an indicator function eliminates the continuity, it was seen how coplots can be defined as a special case of continuous conditioning. Four examples were presented. The first example used simulated data to observe how independent variables would appear using continuous conditioning. Using another simulated dataset, the second example helped establish proof of concept by observing a known trend. In the third example, two groups in a real dataset were found. The final example exhibited a conditioning subset that was not a single point and how to modify that subset to produce the animation. Overall, continuous conditioning is in its developmental stages and the examples show promise for this new method. Defining  $S$  and modifying  $S$  such that informative plots are produced is an open challenge.

## Acknowledgements

This research was supported in part by the National Science Foundation grant NSF 99-71797.

## References

Asimov, D. (1985), "The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM Journal on Scientific and Statistical Computing*, 6, pp. 128-143.

Bujas, A., Cook D., and Swayne D. F. (1996), "Interactive high-dimensional data visualization," *Journal of Computational and Graphical Statistics*, 5, pp. 78-99.

Carr, D. B. and Nicholson, W. L. (1988), "EXPLOR4: A program for exploring four-dimensional data using stereo-ray glyphs, dimensional constraints, rotation, and masking", in *Dynamic Graphics for Statistics* eds. W. S. Cleveland and M. McGill, Belmont, CA: Wadsworth, pp. 309-329.

Chernoff, H. (1973), "The Use of Faces to Represent Points in  $k$ -Dimensional Space Graphically," *Journal of the American Statistical Association*, 68, pp. 361-368.

Cleveland, W. S. (1993) *Visualizing Data*, Hobart Press.

Hurley, C., and Buja, A. (1990), "Analyzing High-Dimensional Data with Motion Graphics," *SIAM Journal on Scientific and Statistical Computing*, 11, 1193-1211.

Scott, D. W. (1992), *Multivariate density estimation: Theory, practice, and visualization*, Wiley.

Wegman, E. J. (1990), "Hyperdimensional data analysis using parallel coordinates", *Journal of the American Statistical Association*, 85, pp. 664-675.

Wojciechowski, W. C. and Harner, E. J. (1995), "Dynamic Coplots", *Proceedings of the 27th Symposium on the Interface*, pp. 274-278.