Displaying Correlations using Position, Motion, Point Size or Point Colour

Serge Limoges
Colin Ware
William Knight
School of Computer Science University of New Brunswick
P.O. Box 4400 Fredericton, New Brunswick CANADA E3B 5A3

Abstract

An empirical study is presented which explores the use of motion, point size and grey scales for the display of statistical data; specifically, the perception of correlations between variables. The task is to measure the subject's ability to perceive differences between high correlation and low correlation for a two dimensional plot presented in different ways. One variable is always mapped to position with respect to the x axis while the other is mapped to one of six different display parameters depending on experimental condition; namely: position *with respect to the y axis, frequency, phase, and amplitude of oscillation, greyscale value of the data points and size of the data points. Human observers appear to be quite insensitive to the relative frequency of moving points but quite sensitive to the correlation of phase angle. Some of the potential advantages and disadvantages involved in using motion for the display of multivariate data are discussed.*

***Keywords*:** data display, motion coding, information visualization

1 Introduction

There is little doubt that using graphics in the display of statistical data can be invaluable in both exploratory data analysis and in the presentation of results (Tukey, 1977). However, when we look at the range of graphical techniques which have been studied, we find them to be surprisingly static for the most part. Statisticians have used position on the page, point shape, point size, colour, texture, the expression on little faces and other display variables to represent data variables. Typically none of these are allowed to vary over time. But the human visual system is extremely sensitive to the motion of objects through space (Cutting 1986) and it seems that a investigation of how motion can be used to convey data parameters is likely to be well worth the effort. The experiment presented here represents our first efforts to investigate how motion may be used for data display.

It is not fair to say that motion has been entirely ignored in data display. The kinetic depth effect is a phenomenon whereby if a three dimensional cloud of points is rotated and projected onto a flat screen, their three dimensional spatial arrangement can be perceived. The three dimensional percept thus obtained is extremely compelling and this phenomenon has been exploited extensively in such products as molecular modeling packages (Wright, 1972) and flight simulators. The kinetic depth effect has also been exploited in packages designed for the exploration of multidimensional scatterplots (Donoho and Gasko, 1988). Typically in these applications a multidimensional scatterplot is rotated about various axes in order to exploit the kinetic depth effect, which thereby allows a three dimensional perceptual window into the data space. This technique can be combined with colour to provide up to three additional perceptual dimensions (Ware and Beatty, 1988). Notice that the above techniques employ motion to enable the observer to perceive a single additional data dimension, a three dimensional space, rather than a two dimensional space is revealed.

It is possible to imagine that motion may be capable of conveying more than a single additional data dimension. Consider a data point which is oscillating sinusoidally in a horizontal plane. The frequency of oscillation may be varied, the amplitude of the oscillation may be varied, and the phase of the oscillation may be varied, with respect to some other oscillating point. We cannot thereby display three data dimensions because frequency and phase are not independent - if two points are oscillating at different frequencies the phase angle between them has little meaning - but this does allow the display of two data dimensions. Moreover, we can also imagine the data point oscillating

independently in a vertical direction. In this way up to four data dimensions may be encoded and displayed using motion.

The purpose of this paper is to present the results of a first experiment concerning the use of motion for data display and to present some informal observations we have made in the course of our investigations.

2 Perceived Correlation

In order to measure the effectiveness of a display technique, it is necessary first to decide on a canonical task. For our canonical task, we measure the effectiveness of motion in revealing correlations between variables. More specifically, we measure the human ability to discriminate a high correlation (above $r^2 = 0.5$) from a low correlation (below $r^2 = 0.9$, where r refers to the product moment correlation. We use r^2 as our metric because studies have shown that perceived correlation is more closely related to r^2 than to r (Pollack,1960), although there is also evidence that neither r nor r^2 are good predictors of relationships perceived in scatterplots (Strahan and Hansen, 1978).

3 Plotting Package

A plotting package was devised to investigate various display parameters. This package lets the user map up to 11 dimensions of a multivariate space to the following display parameters:

- X position
- Y position
- X amplitude α_x
- Y amplitude α_y
- X frequency ϕ_x
- Y frequency ϕ_y
- X phase θ_x
- Y phase θ_y
- X point size
- Y point size
- Grey scale value of point

Motion and x,y position are controlled using the following equations

$$C_x = X + \alpha_x . sin(\phi_x (t+\theta_x))$$

$$C_y = X + \alpha_y . sin(\phi_y (t+\theta_y))$$

Where (C_x, C_y) are the current coordinates, (X, Y) is the point about which the data point moves, (α_x, α_y), (ϕ_x, ϕ_y), and (θ_x, θ_y)define the amplitude, frequency and phase of oscillation, given that the variable t represents time. Data points are plotted as rectangles which can vary independently in width and height. The grey value of each data point is determined by a linear mapping between the monitor black and the monitor white or it can be set to black.

With this plotting package, we can read in samples of multivariate data containing up to twelve variables, and map these data dimensions arbitrarily to the display parameters listed above. We also synthesize data, having known properties, and look at this data using various combinations of motion, position, size and grey scale parameters.

However, although these exploratory exercises provide us with intuitions about the utility of motion for information display, we also wish to provide empirical evidence concerning the efficacy of motion in allowing us to discriminate statistical properties of data, for example, correlations between variables. In our first attempt to formally evaluate the use of motion, we devised the simple experiment presented below. This tests a variety of data presentation

techniques including motion parameters, grey scales, point size and conventional scatterplots, in how well they allow observers to discriminate between degrees of correlation.

4 Experiment

The purpose of this experiment was to determine how effective various display techniques are in permitting an observer to determine whether a given bivariate data sample has a high or low correlation between variables. We displayed the data in six different ways. One of the data variables was always mapped to position with respect to the X-axis, while the other was mapped to one of the following:

1. Y axis - this yields a conventional scatterplot. There is no motion in this plot. Points are scaled to fill the diplay window.
2. Y amplitude - data points move vertically with a fixed frequency and phase. A positive correlation is perceived by points on the right moving through a greater vertical range than points on the left. The amplitude varies between zero and the height of the display window.
3. Y frequency - data points move vertically with a fixed amplitude and phase. A positive correlation is perceived by points on the right oscillating faster than points on the left. The frequency varies between 0 and 0.5 Hz.
4. Y phase - data points move vertically with a fixed amplitude and frequency. A positive correlation is perceived by points on the right moving out of phase with points on the left; alternatively a wave is seen travelling across the scattered points. The phase angle varies between 0 and 7_r radians.
5. Y block size - data points are plotted as like histogram bars of constant width. A positive correlation is percieved by points on the right being taller than points on the left. The data points do not move. The heights vary between Minimum point size = 0.1 cm and Maximum point size = 3.5 cm.
6. Grey scale - data points are given a grey value on a linear (gamma corrected) sequence between monitor black and monitor white. A positive correlation is perceived by points on the right being lighter than points on the left. Grey value varies between monitor black and monitor white.

4.1 Stimulus Parameters

In addition to the display parameters which we explicitly manipulate, there are a number of other display parameters which may effect the efficacy of a given display technique. To give an example: the amplitude of motion can affect discrimination of frequency, if the points are moving a small amount it is hard to perceive how rapidly they are oscillating:

In order to give each different display method the best chance, we attempted to optimize the display for each of the experimental conditions. This was done informally by the two authors of this paper sitting in front of the display and, for each display condition, attempting to optimize such variables as display window, amplitude and frequency. The results of this exercise are reflected in the ranges given to the experimental variables listed above and to the default values for each of the 6 conditions which are listed in the table below.

Cond	Window Width x height (cm)	Ampl % of window used	Freq (Hz)	Phase Angle (rad)	Point size Width Height	Point Grey Value
1	13x12.5	0.0	0.0	0.0	0.5x0.5	Black
2	13x12.5	*	1.0	0.0	0.5x0.5	Black
3	6.5x12	100	*	0.0	0.5x0.5	Black
4	13x12.5	100	0.4	*	1.0x1.5	Black
5	13x3.5	0.0	0.0	0.0	-.2x*	Black
6	10.5x7	0.0	0.0	0.0	0.2x2.0	Black

The background colour for all 6 experiments was a neutral grey midway between monitor black and monitor white. The entries in the table "*" indicate that one dimension of the bivariate data was mapped to this display parameter.

4.2 Measuring the Effectiveness of the Display Techniques

We devised the following task to measure the relative effectiveness of the display techniques. In a given experimental condition, the subject was presented with a random sequence of plots containing 17 degrees of correlation (r^2 = 0.1, 0.2, ..., 0.9 in steps of 0.05). The subject's task was only to decide whether the correlation was high or low - a forced choice binary decision.

Subjects were tested in 4 separate sessions on different days. In each session the observer was tested in each of the 6 conditions, with order of conditions randomized. Each condition involved 85 trials consisting of 5 presentations of each of the 17 degrees of correlation.

4.3 Training

For each of the 6 conditions, each subject was first shown an example of a low, middle, and highly correlated chart. The subject was also told what characteristics to look for when deciding if the chart had low or high correlation value. No time constraint was put on the subject's responses.

4.4 Data Generation

A set of m n-vectors was generated; the correlation matrix of these vectors (not the distribution from which they were drawn) was a predetermined matrix, E. (The algorithm is as follows: Generate n vectors from the multivariate normal distribution with zero mean and identity covariance matrix. Apply the Gram-Schmidt procedure to the data, and rescale it, thereby making the data correlation matrix exactly I, then multiply the vectors by any square root of E, we used the triangular square root.)

4.5 Subjects

The 10 subjects used in the experiment were graduate and undergraduate computer science students at the University of New Brunswick. All had completed an elementary statistics course and were all familiar with the concept of correlation between variables. Subjects were paid to participate.

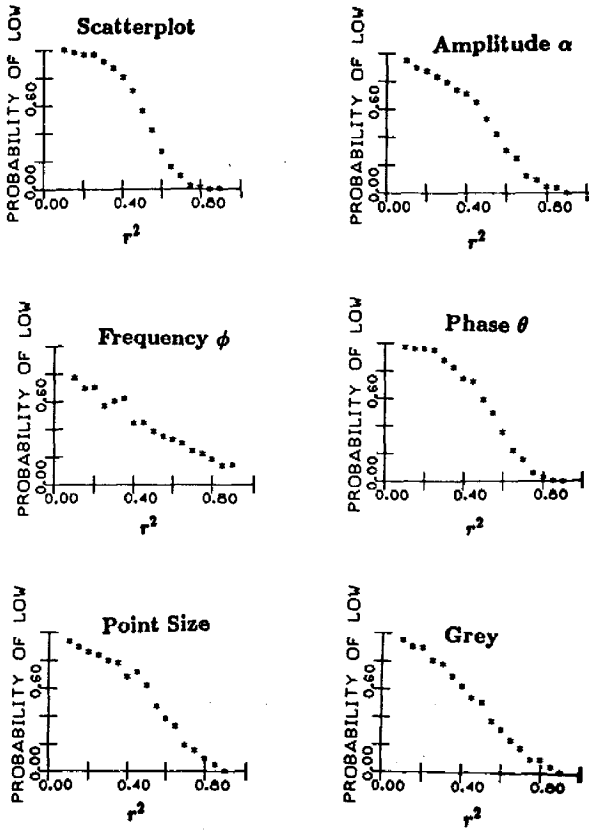

Figure 1

5 Results and Disussion

The results are summarized in Figure 1 which gives data for each of the 6 conditions averaged across subjects and trials. Each plot shows the proportion of trials for which subjects responded “low” in each of the 17 r^2 values. Perfect performance would show up as a step function, $p = 1.00$ for $r^2 < 0.5$, $p = 0.0$ for $r^2 > 0.5$. and by eye it can be determined that the conventional scatterplot is best and the frequency plot is worst. "Best" and "worst" refering to how well the display technique allowed subjects to discriminate a high correlation from a low correlation.

To obtain statistical ordering of the six techniques, we first transformed the probabilities (averaged over trials, but not subjects) using an arcsin transformation (Winer, 1971). This had the effect of converting the distributions into functions which were approximately linear. The gradient was then used as an estimate of the effectiveness of the display method; a steep gradient indicating an effective display. The gradients fall into three indistinguishable sets - Scheffe's test for multiple comparisons was used for this (Scheffe, 1953).

Scatterplot Phase
Amplitude Grey Size
Frequency

According to Scheffe, the phase plot and the conventional scatterplot are the best, while the frequency plot is worst. The grey scale, size and amplitude plots fall into an intermediate category.
This result should reassure us that the conventional scatterplot is a good technique for displaying bivariate data. However, it does not reassure us about other commonly used techniques, namely, the use of vertical bars as in histograms, or the use of grey value to display a data dimension.

Considering the motion conditions, it appears that for sinusoidal motion observers are considerably more sensitive to to the phase angle between moving points than they are to the frequency of motion. A theoretical explanation for this may come from the supposition that one of the primary duties of the human motion detection system is to discover spatial relationships in the environment from the way the image objects of objects move relative to one another across the retina as we move around in the environment (Cutting, 1986). The important information here is in relative motion, which is also the information preserved in the phase manipulation, and to a lesser extent in the amplitude manipulation.

While these results are interesting, this does not mean that we should immediately convert to the use of phase in data display. We can see several possible disadvantages in using motion. The principal of these is that motion is likely to interfere with the perception of position; when a point oscillates about a central location the perception of that central location may lose precision. Such interference may well be less when point size or point colour are used as display parameters.

One strong recommendation does come from this work and this is that when time varying data are displayed using cyclic motion, the important variables should be mapped to the phase or amplitude of motion, and not to the frequency of motion, if this is at all possible

6 References

Cutting, J.E., 1986, Perception with an Eye for Motion. The Massachusetts Institute of Technology.

Donoho, A.W., Donoho, D.L., Gasko, M., 1988, MacSpin: Dynamic Graphics on a Desktop Computer. IEEE, Computer Graphics and Applications, Vol. 8, No, 4, 51-58.

Pollack, L, 1960, Identification of Visual Correlational Scatterplots. Journal of Experimental Psychology, Vol. 59, No. 6, 351-360.

Scheffe, H.A.,1953, A Method for Judging all Possible Contrasts in the Analysis of Variance. Biometrika, Vol. 40, 87-104.

Strahan, R.F., Hansen, C.J., 1978, Underestimating Correlation from Scatterplots. Applied Psychological Measurements, Vol. 2, 543-550.

Tukey, J.W., 1977, Exploratory Data Analysis. AddisonWesley, Reading, Massachusetts.

Ware, C., Beatty, J.C., 1988, Using Color Dimensions to Display Data Dimensions. Human Factors, Vol. 30, No. 2, 127-142.

Winer B.J., 1971, Statistical Principles in Experimental Design. McGraw-Hill, p. 400.

Wright, J., 1972, An Interactive Computer Graphic System for Molecular Studies, Ph.D. Dissertation, Dept. of Computer Sci., Univ. of N.C., Chapel Hill, N.C.