Visualization and Analysis of Multi-variate Data: A Technique for All Fields

T. Mihalisin, J. Timlin, and J. Schwegler
Mihalisin Associates, 600 Honey Run Rd., Ambler PA 19002
(215) 646-3814

Abstract

"Scientific" Visualization has been driven by the need to visualize data sets that involve a large number of points and/or many "dimensions" or variables. However, many, or, perhaps all, fields of endeavor that use numeric data (including, finance, market research, management, manufacturing, process control, risk analysis, social services, health sciences, social sciences, physical sciences, computer science, applied mathematics, the engineering disciplines, and a host of other fields) deal with data sets of this type. Asking who can benefit from data visualization is like asking who can benefit from math or statistics. We present a hierarchical technique for visualizing truly multi-dimensional data that can be applied to any or all of these fields. The emphasis will be on visual statistical analysis of either discrete variables or continuous variables that have been sampled on, or binned to, a regular n-dimensional lattice.

Introduction

Although the need to analyze large data sets of high dimensionality (more than three variables) has been championed primarily by scientists and engineers it is a need that is shared by a wide range of non-scientists in business, government and academia. For instance, financial data, census data, survey data of all types, socio-economic data and medical records are intrinsically multi-variate and can involve tens or even hundreds of millions of data points, and are of interest to a wide audience of scientists and non-scientists alike.

The need to visualize and to a large extent to visually analyze data as graphics rather than text numbers is obvious when one speaks of millions or more data points. However, a satisfactory method of graphically presenting data involving more than three variables is not obvious and in fact has occupied many investigators over the last several decades. A range of techniques involving projections of the data onto two variable subspaces has evolved. Non-projective techniques have also been invented, such as, the parallel axis technique and the use of icons or glyphs. Each of these techniques has been found to be of some utility for visualizing certain types of multi-variate data. Moreover, the projective techniques and the parallel axis technique can be used to provide some types of visual data analysis. However, a satisfactory general method for plotting large multi-variate data sets which allows the user to interactively visualize and to visually analyze his or her data in real time has not emerged from these techniques.

We have invented a new way of plotting large multi-variate data sets which meets these goals. The technique was first reported in a form applicable only to data for which the variables could be classified into two distinct sets, namely a set of independent variables, and a set of dependent variables with the former sampled at values that form a regular n-dimensional lattice. The technique involves the mapping of the n independent variable dimensions to a single hierarchical horizontal axis with a single dependent variable being plotted on the vertical axis. A set of hierarchical data driven symbols representing the behavior of the data at points, along lines, in planes, in 3 spaces, 4 spaces, etc. can be defined in a variety of ways. The use of the first instance of the technique, and in particular, its rules for defining the hierarchical data-driven symbols will only be covered briefly here since it and its use in visually fitting multi-variate data, finding minima, maxima, zeros, regions of stability, dominant and irrelevant variables, etc. has been covered previously. Similarly, extensions of the technique by us and others to the case of creating two subsets of independent variables, one subset represented hierarchically on the horizontal axis and the other subset represented hierarchically on the vertical axis will not be discussed.
This approach uses color or gray scale to represent the dependent variable and may or may not invoke a set of rules for hierarchical data representations. This technique is more appropriate for qualitative visualization of large databases rather than visual data analysis.

In this article we wish to discuss the general applicability of the single hierarchical axis technique and to explore new ways of representing the hierarchical data-driven symbols that are particularly well suited to a variety of visual analysis tasks including fitting, visual display of distribution and cumulative distribution functions and conditional probabilities, visual determination of significant mean differences, visual factor and correlation analysis and multi-variate "what-if" scenarios. In all instances, the techniques are extremely fast because simple 2-d graphics are used to represent all data dimensionalities and the technique only draws a maximum of ~1000 data symbols per screen refresh even though they may represent the behavior of millions or even hundreds of millions of data points.

Multi-variate data and the hierarchical axis

By truly n-dimensional data we mean a list of n-tuples, i.e., n associated numbers. One may think of the n-tuples as representing n variables. These variables may be thought of as either a set of n variables with no classification as to "dependent" or "independent" or as n-1 independent variables and 1 dependent variable which may or may not be single-valued. One cannot divide the variables into m independent variables and n-m-1 single-valued dependent variables and still have an n-dimensional data set. In this case, one has instead n-m instances of data sets of dimension m+1.

One truly robust way of representing, say n-1 independent variables and 1 dependent variable is to list the n tuples vertically, hence forming n columns by r rows where r is the number of n-tuples. This format can be used to represent any and all possible samplings of points in the n dimensional space, from the most highly regular array or lattice to the most randomly scattered points imaginable. When dealing with data points, even of low dimensionality, for example d=3, the classification of the set of points as representing a trajectory or a surface or a volume is not well defined even though viewing the data points may create a psychological impression of a line or surface or volume. The appropriate interpretation may or may not be known to the investigator.

In the technique to be discussed we may plot data by choosing one of the columns (variables) to be the dependent variable and the remainder to be the independent variables. If the values of the independent variables already form a discrete regular lattice we will proceed to our technique to be discussed shortly. If the independent variables are instead continuous in nature and sampled in a more or less random fashion we can form equal interval bins in the n-1 dimensional space of independent variables and in effect map the independent variable space to a regular lattice. In this approach each bin in the n-1 dimensional space may contain zero or a non-zero number of the original data points. We may of course choose to bin all n variables and to create a new dependent variable, i.e., the number of points in each bin. We will assume that the data of interest is manifestly lattice-like as is the case for categorical variables and for many computationally derived data sets or that it can be cast to a lattice via binning.

The mapping of the multiple independent variables to one hierarchical axis is indicated in Figure 1.
1 for the case of 3 independent variables. There are in fact \(3! = 6\) ways of doing this mapping corresponding to which variable is fastest running, which is second fastest, and which is slowest (see Fig. 1). One may think of the data as consisting of a collection of points or lines parallel to the fastest running variable axis or planes parallel to the fastest/second fastest variable plane or as one rectangular parallelepiped. We will give examples for the case where there are no missing values, i.e., an \(n\) dimensional surface, \(n\) being 4 in the case of plotting a single dependent variable over the 3 dimensional space of Fig. 1.

The hierarchical mapping of the independent variables to the horizontal axis seems to imply that the maximum number of points in the \(n\) dimensional space is constrained to be of the order of 1000 or so, i.e., the number of pixels along the horizontal of a monitor. However, we are interested in representing millions or even hundreds of millions or more data points. This problem will be overcome by inventing a set of hierarchical (color or gray-scale coded) symbols that will represent the data not only at individual data points but also along data lines, and in 2d, 3d, 4d, etc. data subspaces. If the number of data points exceeds the number of horizontal pixels we will only see the hierarchical data symbols for data lines and/or data 2d subspaces and/or data 3d subspaces, etc. A set of tools including a subspace zoom tool will allow us to view any region in the space to any desired resolution so that no information is lost.

**Hierarchical data symbols and variable widgets**

Shown in Figure 2 are three ways of rendering the hierarchical data symbols using a small simple data set. The set has three variables \(A, B\) and \(C\). \(A\), \(B\) and \(C\) have 3 values namely 0, 1, 2. The data was generated by the function \(D = (A^2 + B)e^{C}\). Here \(D\) is the dependent variable plotted along the vertical and \(A, B\) and \(C\) are the independent variables plotted along the horizontal axis in the hierarchical fashion described with \(A\) being the fastest running followed by \(B\) and then \(C\).

In Figure 2 we have coded the variables by grayscale with \(A\) being lightest and \(B\) darkest. In the color plate \(A\) is white, \(B\) is blue and \(C\) is red. In Fig. 2 the widths of the rectangles are correlated with the corresponding variables. That is, the white rectangles correspond to points which are sampled

![Hierarchical data symbols and variable widgets](image)

Figure 2. The function \(D=(A^2+B)e^{C}\) with \(A, B, C = 1, 2, 3\). Here \(A\) is the fastest running variable and \(C\) is the slowest. Three methods i.e. the min/max, sum and mean methods, are shown. (See color plate, page 421.)
by moving with the white variable as the fastest running. These rectangles are of width one, i.e., they represent one data point in the space. The blue rectangles represent the behavior of the data over a line in the white direction at a specific value of the blue variable. They are of width 3 and represent the behavior along the line of 3 points. Each red rectangle is of width \( 3 \times 3 = 9 \) and represents the behavior of the data over a plane, i.e., an A, B or white, blue plane at the specific value of the red variable. This association is easily extended to an arbitrary number of independent variables.

In Fig. 2 the heights and vertical positions of the hierarchical rectangles are determined by 3 different rules. Each rule is consistent with the concept that the vertical position and/or height of a particular rectangle should reflect the behavior of the dependent variable in the corresponding subspace, i.e., point, line, plane, 3 space, etc. In the top panel of Fig. 2 the bottom of a rectangle is drawn at a height that corresponds to the minimum value of the dependent variable in the corresponding subspace and the top of the rectangle is drawn at the maximum. This method is called "min/max rendering". In the middle panel of Fig. 2 the bottom of the rectangle is drawn at \( D=0 \) and top is drawn at the sum of \( D \) within the corresponding subspace if the sum is positive. If the sum is negative, the top is at \( D=0 \) and the bottom is at the value of the sum. This method is called "sum rendering". In the bottom panel of Fig. 2 the bottom of the rectangle is drawn at the mean in the corresponding subspace minus the standard deviation of the mean while the top is drawn at the mean plus the standard deviation of the mean. Hence, the vertical location of the center of the rectangle denotes the mean and the vertical extent of the rectangle reflects the uncertainty in the mean (here chosen as 2 standard deviations of the mean). Clearly, other criteria can be used to drive the vertical extent of these rectangles. This method is called "mean rendering."

The three rules for determining the vertical locations and/or extents of the rectangles shown in Fig. 2 are simply illustrative. They were chosen because they have obvious utility for common multi-variate analysis tasks. Other rules may be invented to suit a particular task. The software system under development supports user defined rendering rules.

As discussed in earlier articles, one needs a variety of tools that allow one to permute variables, i.e., change which variable is fastest, or 2nd fastest, etc., zoom to a subspace or lower dimension, i.e., fix a variable or variables to a specific value, zoom to a subinterval of one or more variables, sample the space in a more coarse-grained fashion, etc. Fig. 3 schematically defines a variety of tools. In addition to the tools shown in Fig. 3, one may clone the graph at any point. One may create as many clones as one likes. The clones may be positioned and sized via the mouse. Each tool has an inverse so that one may recover the initial view of the data. Each tool may be applied to any or all variables once or repeatedly. All of the tools are fully integrated so that one may execute any combination of tools on any number of variables. All of the tools may be applied to any rendering method and one may toggle between rendering methods at will.

The animation tool and the general zoom and sub-space zoom tools and their inverses may be applied with auto-scaling of the vertical axis (dependent variable axis) turned on or off. This feature proves to be essential for a variety of tasks including fitting, conditional probability determinations and correlation.

### The sum rendering method, distributions, conditional probabilities and correlations

Shown in Figure 4 is a small data set of low dimensionality drawn using the sum rendering
method. Here the dependent variable (plotted on the vertical) is the number of data entities (for example, companies) while the independent variables A, B and C take on values $A_1$, $A_2$, $A_3$; $B_1$, $B_2$, $B_3$ and $C_1$, $C_2$, $C_3$, respectively. Hence the 3 dimensional independent variable space represented on the hierarchical horizontal axis contains $3 \times 3 \times 3 = 27$ points. Here A is the fastest running variable followed by B and finally C is the slowest. In general, the total number of data entities (for example, companies) may far exceed the number of data points $= 27$ or may be far less than the number of data points. The number of entities with a specific set of

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Population data using the sum method showing 9 $N$ vs. $A$ distributions for 9 values of $(B, C)$, 3 $N_{SA}$ vs. $B$ distributions for 3 values of $C$ and a single $N_{SAEB}$ vs. $C$ distribution. (See color plate, page 421.)}
\end{figure}

$A, B$ and $C$ variable values, say $A = A_i$, $B = B_j$, and $C = C_k$, will be denoted as $N(A_iB_jC_k)$. Each of the narrowest rectangles of width one, in Fig.4 represents one (of the set of 27) $N(A_iB_jC_k)$ numbers and has a height proportional to $N(A_iB_jC_k)$. These rectangles are white. Each of the rectangles of width = 3 (blue/darkest gray-scale) represents the cumulative distribution $\Sigma^3_{i=1}(N(A_iB_jC_k))$. There are 9 such rectangles corresponding to the $3 \times 3 = 9$ values of $B_jC_k$. Each has a height of $\Sigma^3_{i=1} N(A_iB_jC_k)$. Each

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{The 6 possible sum method graphs for the data of Figure 4 that correspond to the $3! = 6$ possible ways of ordering $A$, $B$ and $C$ as the fastest (I), second fastest (II) and slowest (or third fastest III). (See color plate, page 421.)}
\end{figure}
of the rectangles of width=9 corresponds to \( \sum_{j=1}^{3} \sum_{i=1}^{3} n(A_i, B_j, C_k) \). Hence the height of these rectangles vs. \( C_k \) constitutes the doubly cumulative distribution function \( \sum_{j=1}^{3} \sum_{i=1}^{3} n(A_i, B_j, C_k) \). These rectangles are red (intermediate gray-scale) in color. There are 3 rectangles of this type.

Hence, Fig. 4 visually presents 9 distribution functions of the form \( N \) vs. \( A \), i.e., one for each of the 9 choices of fixed variables \( B \) and \( C \). These are the 9 sets of 3 white rectangles inside blue rectangles (white inside dark gray-scale). Fig. 4 also shows 3 single cumulative distributions \( N_{\Sigma A} \) vs. \( B \), i.e., one for each value of \( C \).

It is important to note however, that Fig. 4 does not show all possible distributions. This is due to the ordered nature of the variables along the hierarchical axis. In order to view, for example \( N \) vs. \( B \) or \( N_{EB} \) vs. \( C \) or \( N_{EBEC} \) vs. \( A \), the variable ordering would have to be \( B \) fastest, \( C \) 2nd fastest and \( A \) slowest. Shown in Fig. 5 are the 3!=6 possible permutations of the variables and the resulting hierarchical graphs. This collection of graphs can be made manually be using the permute and clone tools and positioning the clones manually or may be defined as a standard tool macro. This collection of graphs shows all possible distributions and single and double cumulative distributions. Moreover, it shows how distributions and cumulative distributions depend parametrically on other variables. For instance, one can follow how \( N \) vs. \( B \) evolves for differing \( A \) but fixed \( C \), etc.

It is important to recognize that Figs. 4 and 5 may also be used to determine conditional probabilities. For instance consider the 1st red rectangle (far left red or intermediate gray-scale rectangle) of width 9 in Fig. 4. It corresponds to \( C=C_1 \). If one normalizes the height of this red rectangle to unity, then the height of the 1st blue (darker gray-scale) rectangle inside the 1st red rectangle is the probability of obtaining \( B=B_1 \) irrespective of the value of \( A \), for \( C \) fixed at \( C=C_1 \). Similarly, the height of the 2nd white rectangle inside the 1st blue rectangle inside of the first red rectangle is the probability of obtaining \( B=B_1 \) and \( A=A_2 \) subject to the constraint \( C=C_1 \). Similarly, if one normalized the height of the 1st blue rectangle inside the 1st red rectangle to be unity, then the height of the same white rectangle would be the probability of obtaining \( A=A_2 \) subject to the constraint that \( C_1=C_2 \) and \( B=B_1 \), etc. In fact, viewed in this way, Fig. 5 shows all possible conditional probabilities of this sort consistent with the original data. However, one could also ask more complicated questions such as the following: what would be the probability of obtaining either \( C_1 \) or \( C_2 \) if only values \( A_1 \) and \( A_2 \) (i.e., not \( A_3 \) were allowed etc. The fully integrated nature of the general zoom, decimate and subspace tools (with auto-scaling on or off) allows one to quickly answer these questions visually for a mind-boggling number of cases.

We should note from Fig. 5 that \( A \) and \( C \) appear to be distributed uniformly while \( B \) appears to have an exponential distribution. This is probably most obvious if one focuses one's attention on the slowest running variables and hence the double cumulative distributions. If the number of values for \( A \) were substantially larger while the number of \( B \) and \( C \) values remained fixed one would be better able to ascertain whether certain trends seen in non-cumulative distributions are significant. For example, consider the case of \( N \) vs. \( A \) for \( C=C_2 \) and \( B=B_1 \) in Fig. 4. If ten white rectangles showed an exponential decay instead of three the likelihood of it being a fluctuation would diminish significantly, i.e., by a factor of \( 2^7 \).

It is important to recognize that Figs. 4 and 5 may also be used to determine conditional probabilities. For instance consider the 1st red rectangle (far left red or intermediate gray-scale rectangle) of width 9 in Fig. 4. It corresponds to \( C=C_1 \). If one normalizes the height of this red rectangle to unity, then the height of the 1st blue (darker gray-scale) rectangle inside the 1st red rectangle is the probability of obtaining \( B=B_1 \) irrespective of the value of \( A \), for \( C \) fixed at \( C=C_1 \). Similarly, the height of the 2nd white rectangle inside the 1st blue rectangle inside of the first red rectangle is the probability of obtaining \( B=B_1 \) and \( A=A_2 \) subject to the constraint \( C=C_1 \). Similarly, if one normalized the height of the 1st blue rectangle inside the 1st red rectangle to be unity, then the height of the same white rectangle would be the probability of obtaining \( A=A_2 \) subject to the constraint that \( C_1=C_2 \) and \( B=B_1 \), etc. In fact, viewed in this way, Fig. 5 shows all possible conditional probabilities of this sort consistent with the original data. However, one could also ask more complicated questions such as the following: what would be the probability of obtaining either \( C_1 \) or \( C_2 \) if only values \( A_1 \) and \( A_2 \) (i.e., not \( A_3 \) were allowed etc. The fully integrated nature of the general zoom, decimate and subspace tools (with auto-scaling on or off) allows one to quickly answer these questions visually for a mind-boggling number of cases.

We should note from Fig. 5 that \( A \) and \( C \) appear to be distributed uniformly while \( B \) appears to have an exponential distribution. This is probably most obvious if one focuses one's attention on the slowest running variables and hence the double cumulative distributions. If the number of values for \( A \) were substantially larger while the number of \( B \) and \( C \) values remained fixed one would be better able to ascertain whether certain trends seen in non-cumulative distributions are significant. For example, consider the case of \( N \) vs. \( A \) for \( C=C_2 \) and \( B=B_1 \) in Fig. 4. If ten white rectangles showed an exponential decay instead of three the likelihood of it being a fluctuation would diminish significantly, i.e., by a factor of \( 2^7 \).

In Fig. 6 we show a simple example of how correlation between two variables is visually indicated by the sum rendering method. Here a population \( N \) is plotted for 2 variables, \( A \) and \( B \), with 7 bins and 3 bins respectively. The first set of 7 small rectangles (for \( B=1 \)) corresponds to the distribution \( N \) vs \( A \) for \( B \) fixed at \( B=1 \). The second set of 7 small rect-
angles corresponds to N vs A for B fixed at B=2 and the third set corresponds to N vs A for B=3. Note that the most populated A value increases (shifts to the right) as one moves from B1 to B2 to B3. It is this shift that indicates that A and B are correlated.

The mean rendering method vs traditional methods

In Fig.6 the correlation of A and B is indicated in the sum rendering method by a shift in the peak of N vs A as B varies. Next we want to show how correlation is indicated when one uses the mean rendering method to plot the mean of one variable (not the population N) on the vertical versus one or more variables on the horizontal. Consider data consisting of 4 variables A, B, C, and D and 1000 records (N=1000). We will divide A, B, C and D into 5 bins each to form a total of $5^4=625$ bins in our 4 dimensional A, B, C, D space. In Fig.7 we show a tool macro that displays, in a graph matrix style, the mean of each of the four variables on the verticals, versus each choice of the slowest running variable on the horizontal, with trivial cases such as $<A>$ vs A omitted. For instance, the top row shows the mean of A denoted as $<A>$ versus B then versus C and finally versus D. The strong correlation of $<A>$ and D, $<B>$ and D, $<C>$ and D, $<D>$ and B, and $<D>$ and C is obvious from this type of figure. Note that correlations between $<A>$ and B, $<A>$ and C, $<B>$ and A, $<B>$ and C, $<C>$ and A, $<C>$ and B, and $<D>$ and A (if present) are much weaker.

For comparison to more conventional non-hierarchical multi-variate graphing methods, consider Fig.8 which shows the same data as Fig.7 plotted in the traditional graph matrix manner\(^2\). That is, the
variables A, B, C and D are not binned and their means are not computed. Instead one simply plots A (not \(\langle A \rangle\)) versus B (not binned B intervals), A versus C, etc. Hence each of the 12 graphs of Fig.8 contains 1000 points. Clearly, the correlations are not as obvious in Fig.8 as they are in Fig.7. Traditional graph matrix methods have been augmented\(^2\) by dynamic "brushing" which allows one to select a subrange of any variable (e.g., A of Fig.8) and see only the corresponding points in all 12 of the panels of Fig.8. Similarly, one can limit two variables, etc. The hierarchical mean rendering method and in particular, the matrix of graphs displayed in Fig.7 can be set to show the means when one constrains two variables and only averages over the last remaining variable. Hence each rectangle of Fig.7 contains 5 narrower rectangles which correspond to averaging each variable over a line rather than a 2-dimensional subspace (the latter being the rectangles shown in Fig.7). Fig.9 shows the resulting figure consisting of 12 graphs each with 5 wide rectangles and 25 narrow rectangles. For example, the second panel of the top row in Fig.9 shows the mean of A over the CD subspace denoted now explicitly as \(\langle A \rangle_{CD}\) (instead of simply \(\langle A \rangle\)) versus B (wider rectangles) and the mean of A over lines parallel to D denoted as \(\langle A \rangle_D\) versus C and B (C being the faster running variable). In order to obtain similar information using the traditional graph matrix method (and brushing) would require that one retain (in one's mind, or, on hardcopy) the results of 320 brushings, i.e., a total of 320 4x4 graph matrices like the one shown in Fig.8.

Conclusions

The hierarchical graphing methods presented can be used to visually analyze data in a wide variety of ways. Although we have chosen to display small data sets involving only a few variables, the technique can be applied to much larger data sets and still be a real-time visual analysis tool. For instance, a subspace zoom for a dataset involving 10 variables and 1 million bins can be performed in a fraction of a second on a Sun Sparcstation. We are in the process of generalizing the system so that it will support user-defined hierarchical data symbol renderings.

Acknowledgments

The authors wish to acknowledge Temple University's support of this research and all of the associated patent filings.

References

1. See entire issue of IEEE Computer, August 1989, which is devoted solely to scientific visualization.