Multivariate Visual Explanation for High Dimensional Datasets

Scott Barlowe, Tianyi Zhang, Yujie Liu, Jing Yang
Dept of Computer Science
University of North Carolina at Charlotte
sabarlow,tzhang3,yliu39,jyang13@uncc.edu

Donald Jacobs
Dept of Physics and Optical Science
University of North Carolina at Charlotte
djacobs1@uncc.edu

Abstract

Understanding multivariate relationships is an important task in multivariate data analysis. Unfortunately, existing multivariate visualization systems lose effectiveness when analyzing relationships among variables that span more than a few dimensions. We present a novel multivariate visual explanation approach that helps users interactively discover multivariate relationships among a large number of dimensions by integrating automatic numerical differentiation techniques and multidimensional visualization techniques. The result is an efficient workflow for multivariate analysis model construction, interactive dimension reduction, and multivariate knowledge discovery leveraging both automatic multivariate analysis and interactive multivariate data visual exploration. Case studies and a formal user study with a real dataset illustrate the effectiveness of this approach.

Keywords: visual analysis, multivariate analysis, dimension reduction, multivariate model construction, multivariate visualization.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—User Centered Design; G.3 [Mathematics of Computing]: Probability and Statistics—Multivariate Statistics

1 Introduction

The analysis of multivariable behavior and model construction for quantitative relationships among a large number of variables is important in a myriad of applications. For example, economic forecasting is dependent on the relationships among unemployment, interest rates, consumer confidence, and inflation. Predicting solvent formulation for protein storage, an example important to pharmaceutical technology, depends on the nature of the protein, and the temperature, pH, ionic strength, and co-solute concentrations of the solvent. Often, it is desirable to identify significant affecting factors from empirically accumulated data, and discover hidden relationships that transcend a neural network approach. To deduce an explanation of the behavior of monitored characteristics, multivariate data must be examined in context with all attributes simultaneously.

Multivariate analysis is a mature topic in the field of statistics. Numerous statistical methods are available, such as linear regression [6], generalized additive models [11], and response surface analysis [3]. Despite ample computational power of modern computers, application of automatic statistical methods for constructing correlation models scales poorly to increasingly massive, high dimensional multivariate datasets. Two important reasons are that analysts find it difficult to apply domain knowledge critical for understanding complex data, and their perceptual ability to discern relationships is lost when using an automatic analysis approach.

Information visualization approaches allow users to gain insights from complex abstract data using their perceptual abilities and domain knowledge, but are they helpful for multivariate analysis? In the field of information visualization, a major category of techniques named multivariate visualization does exist. It aims to help users analyze multivariate datasets. However, existing multivariate visualization techniques are quite limited in helping multivariate analysis when there are more than three variables involved. For example, in Figure 1, a four variable dataset with one thousand data items is shown in a pixel-oriented display [16], parallel coordinates [13], and scattorplot matrices [5]. These are among the most popular, most widely used multivariate visualization techniques. However, can you tell from any of the displays the simple relationship among the four dimensions: $y_0 = x_0x_1 + x_2$? Just as Amar and Stasko pointed out in their Infovis 2004 paper [1], there is an urgent need for multivariate visualization systems to support determination of correlation among multiple variables. A *Worldview Gap*, namely the gap between what is being shown and what actually needs to be shown to draw a straightforward representational conclusion for decision making [1], is to be filled.

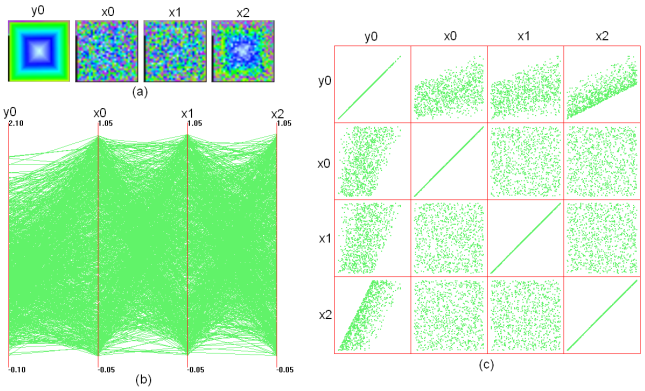


Figure 1: A 4-d dataset in (a) a pixel-oriented display, (b) parallel coordinates, and (c) a scatterplot matrix. (a) and (b) were generated using XmdvTool [28].

In this paper, we present our preliminary research toward supporting determination of correlation among multiple dimensions in multivariate visualization systems to fill the Worldview Gap. We call this type of research *Multivariate Visual Explanation* (MVE), following the multivariate explanation notion proposed by Amar and Stasko [1]. An MVE system explicitly reveals the hidden multivariate relationships to users in a straightforward way. In particular, the MVE techniques presented in this paper are targeted at answering the following questions important to multivariate analysis:

Given a multivariate relationship, how does an independent variable affect the dependent variable in the context of the remaining independent variables? Is the effect positive or negative? Is the effect strong or weak? Are there any independent variables closely entangled in their effects and thus need to be considered together in the analysis? Can we identify ignorable variables and infer interdependence between variables to achieve rational dimension reduction? From strategic analysis, can we construct a model to quantify the relationships between dependent and independent variables?

A system that can answer the above questions will not only ben-

IEEE Symposium on Visual Analytics Science and Technology
October 21 - 23, Columbus, Ohio, USA
978-1-4244-2935-6/08/$25.00 ©2008 IEEE

efit multivariate analysis, but also benefit multivariate visualization techniques by determining significant factors and entangled dimensions for dimension reduction in high dimensional data visualization. In addition, visually aided model construction will benefit automatic multivariate analysis since a good model is critical to ensure effectiveness and efficiency. As shown in Figure 1, it is almost impossible to do this job using pure visualization techniques. A natural approach is to integrate numerical methods with visualization techniques.

In our preliminary work, we chose to use partial derivatives calculated by a numerical differentiation method to reveal the local effects of the independent variables on the dependent variables, and allow users to gain a global view of the effects through multivariate visualization techniques. We selected this approach for three reasons. First, partial derivatives convey intuitive meanings and are familiar to analysts in domains such as physics and engineering. Second, partial derivatives can be easily displayed together with the original data in existing multivariate visualization techniques. Third, compared with other analysis methods that generate summary results, partial derivatives reveal multivariate relationship details across the whole multidimensional space and the users can still gain summarized information through visual aggregation. Our approach consists of the following components:

Partial derivative calculation and inspection: A numerical differentiation method is used to calculate the partial derivatives of the dependent variables on their independent variables. The error bounds of the partial derivatives are visually examined (see Figure 3 for an example) to ensure the quality of the partial derivatives to be used in the following partial derivative visual exploration.

Visual exploration of partial derivatives: The partial derivatives are visually presented with the variables to help detect correlation among the variables. To reduce clutter, a step by step visual exploration pipeline is used to guide users in analyzing different types of correlations in different steps. Different views, such as the First Order Partial Derivative Histograms (see Figure 4 for an example) and the Independent Variable-First Order Partial Derivative Scatterplot Matrices (see Figure 7c for an example), are provided in these steps for determining different types of correlations.

This visual exploration process is tightly integrated into a multivariate visualization system (see Figure 6 for an example). All partial derivative views are coordinated with other multivariate visualizations available in the system such as parallel coordinates and glyphs. Thus users can perform interactions such as interactive dimension selection and data selection from the partial derivative views and propagate the selection results to other views in the system in support of meaningful dimension reduction and data filtering. The coordinated views and the interactions available in them such as N-D brushing [23] also enable flexible visual exploration of the partial derivative views.

Interactive multivariate model construction: An interactive model construction interface is provided to allow users to interactively construct correlation models for high dimensional datasets (see Figure 7 for an example). Coupled with the step by step partial derivative visual exploration pipeline, the interface allows users to generate models that can be used in automatic multivariate analysis which are useful for more than acquiring a qualitative impression about the correlations.

The major contributions of this paper are:

- A novel MVE approach that is tightly integrated into a multivariate visualization system is proposed. It supports determination of correlation among variables in high dimensional datasets. It leverages the visual exploration in other views provided by the system by allowing users to perform dimension reduction and data filtering using the MVE insights. It also allows users to interactively construct multivariate models for automatic analysis;
- A novel visualization approach is proposed to examine the quality of the partial derivatives. A novel visualization pipeline and informative displays for examining multivariate correlations and constructing multivariate models are proposed;
- A formal user study and case studies have been conducted. The case studies reveal how the proposed approach supports users to effectively detect multivariate relationships. The user study compared the proposed approach with scatterplot matrices in revealing multivariate relationships in a real dataset and its result was strongly in favor of the proposed approach.

The rest of the paper is organized as follows: Section 2 summarizes the related work; Section 3 briefly introduces the partial derivative calculation method we use and presents our visual differentiation error examination approach; Section 4 describes the visual exploration of the partial derivatives; Section 5 demonstrates the interactive model construction process; Section 6 presents the user study we conducted; and Section 7 gives our conclusions and future work.

2 Related work

There exist many automatic techniques for multivariate analysis. For example, regression analysis [6] establishes a linear relationship between an independent variable and a dependent variable in its simplest case. It has been extended to include multiple independent variables and describe more complex relationships, such as generalized additive models [11] for detecting nonlinear relationships. Response surface analysis [3] is a method of discerning multivariate relationships through model fitting and 3d graphs [17]. Our approach is different from the automatic techniques in that it enables users to intuitively examine multivariate relationships in detail by tightly integrating numerical approaches with interactive visual explorations.

The multivariate analysis tool we use is differentiation, which can establish detailed quantifiable relationships among multiple variables. Differentiation can be conducted numerically for discrete data items. Numerical differentiation methods include finite, polynomial interpolation, operator, lozenge diagrams, and undetermined coefficients [20]. Other approaches span spline numerical differentiation, regularization, and automatic differentiation. Recently, the contribution that nonuniform fast Fourier transforms [22] and wavelet transforms [26] can have in numerical differentiation has been examined. Numerical differentiation is intuitive, flexible, powerful, and is widely used in applications such as engineering and the physical sciences. For example, it was used in cell cycle network research to prove the usefulness of mathematical models in molecular networks [25]. Image processing has also benefited from partial derivatives. Partial derivatives have been used as image descriptors through higher-order histograms [21]. However, these and other similar approaches are typically relegated to specific application domains and do not provide a general framework in which partial derivatives and their visualizations are used in dynamic dimension reduction and model building. To the best of our knowledge, the MVE approach we proposed is among the first efforts toward using numerical differentiation techniques to enhance high dimensional data visualization.

There exist a few efforts toward understanding the relationships between pairs of dimensions in existing multivariate visualization systems. Among them, rank-by-feature [24] calculates measurements such as correlation and uniformity between pairs of dimensions and allows users to select two dimensional projections of a high dimensional dataset according to these measurements. Value

and Relation display [29] calculates the pair wise correlation between each pair of dimensions and visually presents the correlations to users through dimension positions in a two dimensional display using multi-dimensional scaling. Some traditional multivariate visualization methods, such as scatterplot matrices [5] and parallel coordinates [13], can also visually reveal correlations between pairs of dimensions. However, few approaches effectively convey the correlation between two dimensions in the context of other dimensions. Among the few exceptions, the conditional scatterplot matrix [8] depicts partial correlations among variables, but it is hard to scale to high dimensional datasets.

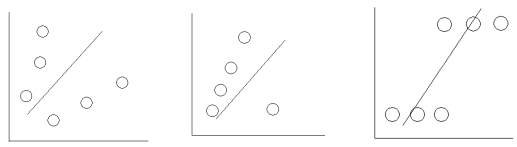
Model construction and selection, namely the construction and selection of appropriate predictive or explanatory models for automatic multivariate analysis, is an important research topic in multivariate analysis since the effectiveness and efficiency of a large portion of multivariate analysis algorithms heavily depend on the underlying model used. Numerous algorithms have been proposed for model construction and model selection. Examples of model selection methods include bootstrap and backward elimination [7], nonlinear bounded-error estimation [4], and visualization [12]. For high dimensional datasets, most automatic algorithms lose their effectiveness and efficiency due to the large number of candidate models and the number of dimensions per model. Our approach provides users a transparent model construction process with the help of both automatic numerical differentiation calculation and human perceptional abilities and domain knowledge. Our approach is different from previous visual model selection approaches, such as D2MS [12], by integrating visualization into the preprocessing steps through which users can interactively filter unimportant dimensions.

Dimension reduction is an important topic for a wide range of research fields such as data compression, pattern recognition, cluster and outlier detection, multivariate analysis, as well as visualization. For example, dimension reduction can reduce the number of candidate models or the number of dimensions per model and thus leverage the model construction and selection process in multivariate analysis. In visualization, projecting a high dimensional dataset to a lower dimensional space can also effectively reduce the clutter of visualizations. Commonly used dimension reduction techniques include principle component analysis [15], multidimensional scaling [19], and Self Organizing Maps [18]. Major drawbacks of these automatic techniques are that they yield results that have little intuitive meaning to users and that they may yield huge information loss for high dimensional datasets. A few visualization approaches, such as VHDR [30], have been proposed allowing users to interactively select dimensions for constructing lower dimensional spaces. Unfortunately, only pairwise dimensional relationships are considered in VHDR and thus its capability of manual dimension reduction is largely limited.

3 Partial derivative calculation and inspection

3.1 Partial derivative calculation

Our multivariate visual explanation approach is based upon partial derivatives calculated using numerical differentiation. The derivative of a dependent variable, y, as the independent variable, x, changes is approximated as $\Delta y/\Delta x$. The relationship is geometrically interpreted as a local slope of the function $y(x)$. This idea is extended to partial derivatives where multiple independent variables are analyzed. In partial differentiation, the derivative of the variable of interest is taken while all other independent variables are held constant. This can be repeated until a quantitative relationship to the dependent variable can be found for each independent variable. For example, partial differentiation of the relationship $y = x_1x_2 + x_3$ yields $y_{x_1} = x_2$, $y_{x_2} = x_1$, and $y_{x_3} = 1$. Here $y_{x_i} \equiv \partial y/\partial x_i$. Furthermore, the non-zero second order partial derivatives yield $y_{x_1,x_2} = y_{x_2,x_1} = 1$ where $y_{x_i,x_j} \equiv \partial y_{x_i}/\partial x_j$.



(a) Multi-Trends (b) Outlier (c) Piece-Wise Defined Function

Figure 2: Typical Errors [9, 14]. (a) Data with two different trends is oversimplified into a straight line. (b) One outlier dramatically skews the result. (c) A piece-wised defined function is treated as a line.

There exist many methods to obtain partial derivatives [9, 14], and thus it is not the focus of this paper. We only briefly introduce the partial derivative calculation for completeness. We extract a local tangent on every point in a high-dimensional space. For point P specified by n variables, $x_0, x_2 ... x_{n-1}$, the set of neighboring points within a threshold t is defined as:

$$\text{Set}(P) = \{p | \|P - p\|^2 < t\} \quad (1)$$

The data items are segmented into small groups of neighboring points determined by a threshold based on dimension value ranges and adjusted according to the amount of acceptable error. A tangent line for each central point is calculated based on its neighbors using linear regression for every data entry and independent variable. To find higher order derivatives, the set of values consisting of the slopes of tangent lines are repeatedly substituted in place of the original set of points. Although the assumption of continuity may fail for discrete data, numerical differentiation often continues to provide useful characteristics in exploring relationships.

3.2 Partial Derivative Inspection

Errors can be introduced in the partial derivative calculation, as shown in Figure 2. In Equation 1, we have to trade off between the significance of errors and the speed of the algorithm by setting the boundary searching threshold t. In addition, care must be taken to ensure that the function under consideration is well behaved being differentiable at the points of analysis. Since errors can overwhelm users with distracting and inaccurate information in the following visual explorations, we propose a novel visualization approach for partial derivative quality inspection. This approach not only allows users to visually examine the quality of partial derivatives calculated, but also enables users to filter out low quality partial derivatives from the following visual explorations. Furthermore, users can adjust the partial derivative calculation parameters to improve the overall quality based on the visual feedbacks provided by the visualization.

First, we calculate the partial derivative error for each point. Among choices such as the sum of squared errors, the sum of absolute errors, and the maximum value of errors, we find that the sum of squared errors achieves the best performance for many real datasets in practice. The sum of squared errors for P is calculated using Set (P) and is expressed as:

$$E = \sum (y_i - \widehat{y_i})^2 \quad (2)$$

Then errors are visually shown using an extension of parallel coordinates to allow users to interactively examine the errors. Figure 3 shows an example of such a display. It shows the original dimensions, partial derivatives, and differentiation errors of a synthesized segmented dataset named SegData, defined as: $y = 8x_0 + x_1$ if $x_0 \geq 0.6$ and $x_1 \leq 0.3$ and $y = x_0 - 7x_1$ otherwise for x_0 and x_1 randomly distributed on [0,1].

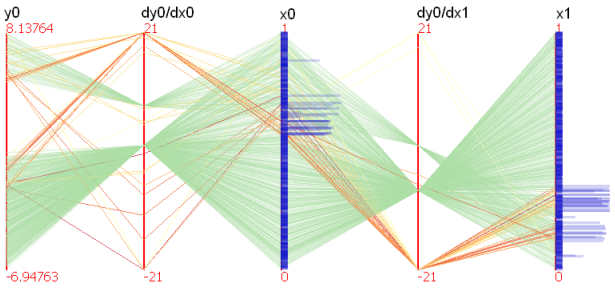


Figure 3: The error display of a segmented dataset. The lengths of the horizontal blue bars and the colors of the data items indicate the error bounds of the partial derivatives.

In Figure 3, the data items are colored by the average errors of their derivatives. Each horizontal blue bar attached to a projection of a data item on an axis indicates the first order partial derivative error at that data item. The longer the bar, the larger the error is. It can be seen from the figure that there are large errors around the segmentation boundary. The error display reminds users to drop derivatives with large errors in the MVE. Since errors can often be reduced by adjusting the threshold t used in Equation 1, this visualization is useful in helping improve the quality of the derivatives calculated by adjusting t.

Besides this method, the errors can be treated as extra dimensions and visualized together with the variables. Figure 6a shows such an approach where x0_Error and x1_Error give the errors of y_{x_0} and y_{x_1} respectively. The data items with low quality derivatives were unselected from the display so that users can focus on high quality derivatives in the visual exploration. They can also be filtered out to avoid distracting the users.

4 Visual Exploration of Partial Derivatives

After the partial derivatives are calculated and inspected, they need to be visually presented to users in the context of the original data to facilitate users in detecting correlation among multiple variables. This is a challenging task: the partial derivatives are a data body that can be larger than the original data; and the data volume is even larger when the partial derivatives are considered together with the original data. The following example shows how large the data volume will be: Assume that there is a 4 dimensional dataset that contains one dependent variable, namely y, and 3 independent variables, namely x_0, x_1, and x_2, and the calculated partial derivatives up to the second order. Then 9 extra dimensions will be added into the dataset that include: y_{x_0}, y_{x_1}, y_{x_2}, y_{x_0,x_0}, y_{x_1,x_1}, y_{x_2,x_2}, y_{x_0,x_1}, y_{x_0,x_2}, and y_{x_1,x_2}. Thus, rather than exploring a 4 dimensional dataset, we now need to explore a 13 dimensional dataset! This number increases significantly when more variables are considered.

Before we reached our final solution, we had a few failed trials on visualizing the extended datasets that augmented the partial derivatives. Our first attempt was a modified scatterplot matrix. We tried to dedicatedly arrange the scatterplots of all pairs of dimensions in the extended dataset in the same display to reveal correlations. Since there were too many scatterplots that cluttered the display, we employed the graph-theoretic scagnostics technique [27] to capture the outliers, shape, trend, density, and coherence of the scatterplots and colored them according to a measurement of user interest. However, our informal user studies showed that it was a failed trial since users were completely overwhelmed by so many scatterplots and so many possible correlations among the variables. An important lesson from this attempt is that the various relationships among the partial derivatives and variables reveal different types of correlations among the variables. We should not mix them together since this seriously challenges the working memory of users.

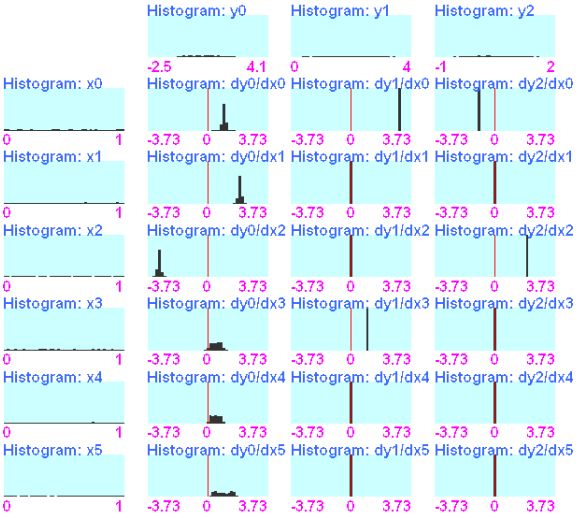


Figure 4: The first order partial derivative histogram view of the ThreeSix dataset

Our final solution is a step by step visual exploration pipeline where different patterns are examined in each step. The rules of this pipeline are: (1) different types of correlations are examined in different steps; (2) correlations that are easier to be detected will be examined before the more complex correlations; and (3) once the correlation between an independent variable and the dependent variable is decided, that independent variable will be excluded from further steps so that the users can focus on the variables with unknown correlations. For each step in the visual exploration pipeline, one or more views are provided.

The view for the first step is a highly condensed first order partial derivative histogram display. Figure 4 shows the histogram view for a synthesized dataset named ThreeSix. It has 3 dependent variables (y_0 - y_2), 6 independent variables (x_0 - x_5), and 1000 data points. In this figure, all independent variables are examined together for all the dependent variables. In the top row are the histograms of the dependent variables. In the left most column are the histograms of the independent variables. The rest of the views are histograms of the first order partial derivatives $\partial(y_i)/\partial(x_j)$, where y_i is the dependent variable whose histogram is shown in the top row for the same column and x_j is the independent variable whose histogram is shown in the left most column for the same row.

The histograms of the first order partial derivatives reveal important information about the data: a first order partial derivative dimension with mostly positive/negative values reveals a positive/negative effect of the independent variable on the dependent variable. If the scales are properly selected, the significance of the independent variables on the same dependent variable can be compared by the shapes of their partial derivative histograms. In our approach, we set the scales in this way: we assume that all values of the independent variables from which the analyzed datasets are sampled are randomly distributed in known value ranges, and the variables are normalized into [0, 1] ranges from their real value ranges. Since for a given data point, the value of a first order partial derivative reflects the change of the dependent variable per unit change of the independent variable when all the other independent variables are held constant, the higher the absolute values of the partial derivatives from the derivative calculation, the stronger the independent variable impacts the dependent variable. To enable users to directly compare the absolute derivative values for com-

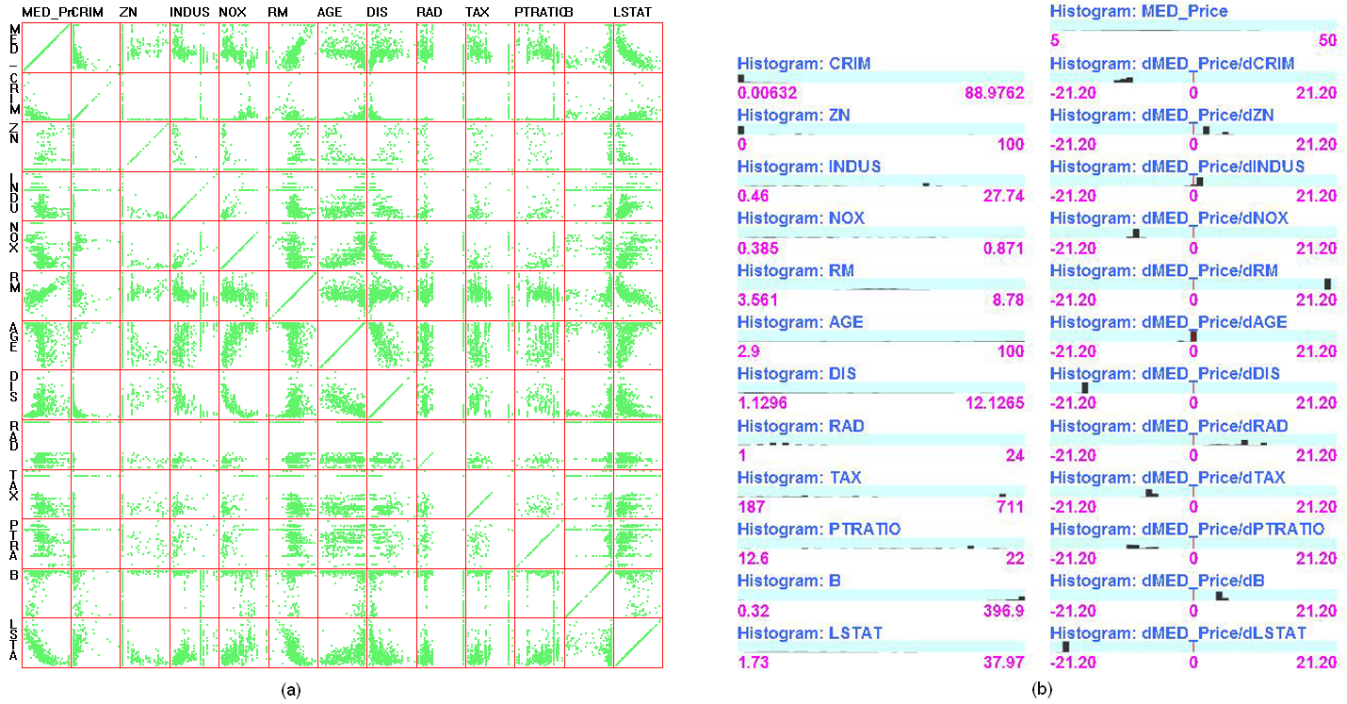


Figure 5: The Boston Neighborhood Housing Price dataset [2] in (a) a scatterplot matrix and (b) first order partial derivative histogram. (a) was generated using XmdvTool [28].

paring the variable impacts, the value ranges of all the derivative histograms are set to be the same. Although in many real applications the random distribution assumption does not hold and we simply normalize the independent variables by their actual value range in our system, the histogram view still allows users to estimate and compare the impacts of the independent variables. The histograms of the dependent variables and independent variable provided in the histogram view not only serve as an index of the derivatives, but also allow users to examine the distribution of the independent variables for judging their impacts on the dependent variables observed from the partial derivative histograms.

Figure 5 shows an interesting example with a real dataset, namely the Boston Neighborhood Housing Price (BNHP) dataset [2], which is a corrected version of the Boston house-price data [10]. It contains 506 data items and 14 variables. A dummy variable with huge derivative errors is dropped from the display although it is considered when calculating the derivatives of other variables. The variables displayed are listed as follows:

- Med-Price(y): Median value of owner-occupied homes in $1000's
- CRIM(x_0): per capita crime rate by town
- ZN(x_1): proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS(x_2): proportion of non-retail business acres per town
- NOX(x_3): nitric oxides concentration (parts per 10 million)
- RM(x_4): average number of rooms per dwelling
- AGE(x_5): proportion of owner-occupied units built prior to 1940
- DIS(x_6): weighted distances to five Boston employment centers
- RAD(x_7): index of accessibility to radial highways
- TAX(x_8): full-value property-tax rate per $10,000
- PTRATIO(x_9): pupil-teacher ratio by town
- B(x_{10}): $1000(Bk-0.63)^2$ where Bk is the proportion of blacks by town
- LSTAT(x_{11}): - % lower status of the population

We view Med-Price as the dependent variable and the others as the independent variables. The correlation among them is explored. Figures 5a and 5b show the BNHP dataset displayed in a scatterplot matrix and the proposed histogram view respectively. In the histograms, values increase from the left side to the right side, and the red lines indicate the zero value. Many correlations hidden in the scatterplot matrix are explicitly revealed in the histogram view. For example, we found that PTRATIO had a negative correlation with housing prices, i.e., the higher the pupil-teacher ratio by town, the lower the housing prices. Also, the accessibility to radial highways (RAD) had a positive correlation with the housing prices, and the weighted distance to the five employment centers (DIS) had a negative correlation with the housing prices. Such correlations can hardly be detected from the scatterplot matrix.

Assuming that users want to reduce the number of dimensions displayed in a coordinated display to reduce clutter, it is convenient to accomplish this from the histogram view: since INDUS and AGE show pretty small effects on the housing prices, they are not interesting to the users and thus can be removed. The users can perform the dimension reduction easily from the histogram view by clicking these dimensions. In addition, the partial derivatives can be recalculated without the ignorable variables to eliminate the noise caused by them.

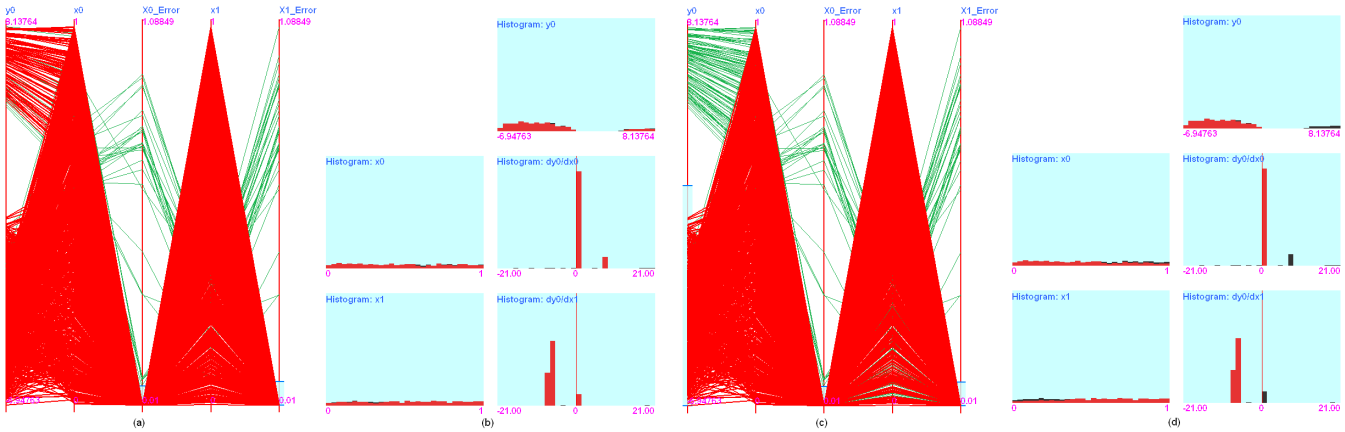


Figure 6: Coordinated Views of the SegData Dataset. (a) The parallel coordinates view of a dataset where the variables and their derivative errors are shown. Data items with good derivative qualities are selected and highlighted. (b) The selection is propagated to the histogram view. It shows that the noise in the histograms is the unselected low quality derivatives. (c) The selection in the parallel coordinates is modified so that only data items whose y_0 values fall into the lower value group are selected. (d) The corresponding histogram view shows that the selected data items have constant derivatives in both x_0 and x_1.

The histogram view not only reveals the significance of the independent variables, but also allows users to discover correlations: if y_{x_0} is a constant, then $y = A_0x_0 + f_0(x_1,..,x_n)$, where A_0 is a constant. In other words, if we find a first order partial derivative dimension with most values concentrated in a small value range (a histogram with a slim, tall bar), namely its value is a constant, the correlation between the independent variable and the dependent variable is discovered. For example, it can be seen from Figure 4 that all the correlations in the ThreeSix dataset belong to this type except y_{0x_3}, y_{0x_4}, and y_{0x_5}. Users can pick up the exceptions to examine them further in the following steps. Variables with known correlations will be removed from the following steps to reduce clutter and confusion.

A drawback of the histogram view is that it only conveys aggregated information. To overcome this drawback, we coordinate the histogram view with other multivariate views such as parallel coordinates and scatterplot matrices. In these displays the extended datasets or their subsets are displayed. Users can interactively select subsets of data items of their interest from these displays. The aggregated information of the selected subsets is then visually displayed in the histogram view (see Figure 6 for an example).

The view for the second step consists of scatterplots between the first order partial derivatives and the independent variables. A single dependent variable is examined in such a view. Figure 7c shows an example of the second step scatterplots. In this figure, each scatterplot is composed of a first order partial derivative dimension and an independent variable dimension. This view allows users to discover correlations. If y_{x_0} is linearly related to x_i $(0 \leq i < n)$, then there will be a straight line in the scatterplot of y_{x_0} versus x_i. Thus, $y = Bx_0x_i + A_0x_0 + f_0(x_1,..,x_n)$, where B is another constant. Interestingly, periodic correlations, such as $y_{x_0} = sin(x_0)$ can also be detected from this view.

After the second step, scatterplots between the second order partial derivatives and the independent variable, and scatterplots between the second order partial derivatives and the first order partial derivatives are provided in different views to allow users to detect more complex correlations.

5 Interactive Model Construction

We support users in interactive model construction for further multivariate analysis through an interface tightly integrated with the step by step visual exploration pipeline. In particular, a dialog (see Figure 7a) is used to record the exploration results from the steps and summarize the final result. At the top of the dialog, there are multiple tagged pages. One page is the control page and the others are step pages. The control page allows users to select the dependent variable whose model is to be constructed and to set the dimension reduction propagation mode within and outside the pipeline. Each step page is associated with a visual exploration step and contains one or more lists recording the outcome from the visual exploration of that step. At the bottom of the dialog, there is a summary list summarizing outcomes from all steps so far and a button for model construction when the exploration is done.

The switch of views in the main display triggers the switch of the step pages in the dialog, and vise verse. Users can go through the pipeline starting from the first step and go back to a previous step at any moment during the visual exploration. In each step, users inspect the display to find histograms or scatterplots containing desired correlations. They then use simple keyboard input and mouse clicks to import the correlations into the dialog.

In the first step, a derivative histogram with a small value range around zero indicates that this independent variable is ignorable compared to other independent variables. A left click on a histogram sends the independent variable name into the *ignorable variable list* in the step one page in the dialog. A left click with the control key pressed removes a variable from the list. A derivative histogram with a small and concentrated positive/negative value range indicates that the independent variable is linearly related to the dependent variable while holding other variables constant. A right click on a histogram sends the independent variable name into to the *y = ax + f(other xs)* list. In the second step, variable pairs are sent to the *y = ax1x2 + bx1 + f(other xs)* list if data are distributed in a straight line in their derivative/variable scatterplot. We allow users to create their own lists for more complex correlations in the second step and the later steps.

In the default dimension reduction propagation setting, once a variable is imported into a list, it won't be shown in the views of the following steps. For example, after a linear correlation is detected in the first step, the independent variable and its derivatives will not be considered in the second step to avoid visual clutter and a model more complex than necessary. Users can also examine all variables in all steps by changing the settings from the control page. For views outside the pipeline, users can select to view all dimensions, dimensions with detected correlations only, or all di-

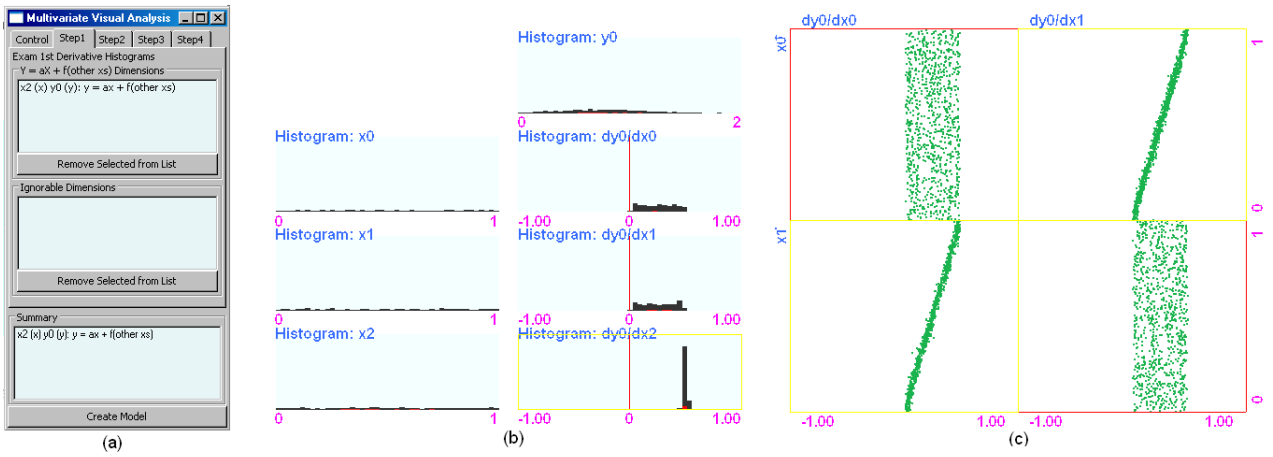


Figure 7: Interactive model construction for the $y_0 = x_0x_1 + x_2$ dataset. (a) model construction dialog in Step 1 (b) Step 1 display (c) Step 2 display

mensions except the ignorable dimensions, combined with manual dimension reduction. The summary list at the bottom of the dialog shows all correlations detected so far. After users finish the visual exploration, they click the *Create Model* button to automatically generate a model using information in the summary list. The users can then use the model in other statistics packages for further analysis.

Figure 7 shows a model construction example with the example dataset shown in Figure 1. In the first step (Figure 7b), it was detected that $\partial(y_0)/\partial(x_2)$ was a constant (the histogram with a yellow frame). The user right clicked on the histogram and sent x_2 to the *y = ax + f(other xs)* list, i.e., the information $y_0 = A_2x_2 + f_0(x_0,x_1)$ was recorded. Figure 7a shows the dialog after this operation. In the second step (Figure 7c), only x_0, x_1, and their derivatives were examined. Straight lines were detected from the x_0-$\partial(y_0)/\partial(x_1)$ and the x_1-$\partial(y_0)/\partial(x_0)$ scatterplots (highlighted by yellow frames). These scatterplots were clicked and the information $y_0 = B_0x_0x_1 + A_0x_0 + f_1(x_1,x_2)$ and $y_0 = B_0x_0x_1 + A_1x_1 + f_2(x_0,x_2)$ was recorded. Since the correlations between all independent variables and the dependent variable had been decided upon, the user clicked the *Create Model* button and the model: $y_0 = B_0x_0x_1 + A_0x_0 + A_1x_1 + A_2x_2 + C$ will be created and shown to the user.

6 User Study

6.1 Setup

A formal user study has been conducted to evaluate how our approach helps users understand the impact of independent variables on a dependent variable. We compared our first order partial derivative histogram view (derivative view to be short) with traditional scatterplot matrices without derivatives (scatterplot view to be short) since the latter is known for being good at revealing dimension relationships. Our assumption was that the derivative view could explicitly reveal correlations among multiple variables that were invisible from the scatterplot view.

The dataset we used was the Boston Neighborhood Housing Price (BNHP) dataset [2]. This dataset was selected since housing prices and their affecting parameters were so familiar to us that the correlations detected from this dataset could be justified. In order to eliminate the influence of users' prior knowledge of the housing price in the user study, we used x_0 to x_{11} to replace the true meaningful variable names, as shown in the BNHP variable name list. A screen capture of the BNHP dataset in the scatterplot view and a screen capture of the dataset in the derivative view, similar to the ones shown in Figure 5, were each printed out on A4 paper.

The user study was a within subjects, balanced user study. Eight graduate students participated in the user study, all of which majored in computer science. The subjects completed the study one by one with the same instructor. All work was done on paper with the screen captures since no interactions were evaluated in this study.

The tasks the subjects conducted were to classify the correlation of each independent variable to the dependent variable into one of three types, namely positive, negative, or ignorable/uncertain correlation. The subjects were also required to record their confidence in each decision they made using a 0 - 5 scale (0 - low confidence, 5 - high confidence).

6.2 Procedure

The user study was conducted as follows. Each subject worked two sections. Half of them worked with the scatterplot view first and the derivative view second. Half of them worked in a reversed order. Each section was conducted as follows. First, the instructor suggested to the subject how to find correlations from the view to be tested. A short question/answer time followed the instruction. Then the subject was given the screen capture of the view and conducted the tasks. At the end of the test, the participants were asked to complete a post-test questionnaire to rate their satisfaction on the two views based on the performed tasks.

6.3 Result

The correctness of the result was strongly in favor of the derivative view. The average correct answer rate for all variables was 99% for the derivative view and only 66% for the scatterplot view.

The variable by variable correct answer rates are shown in Figure 8a. This detailed view shows the difficulty the subjects encountered in identifying the influence of x_1, x_2, x_3, x_7, x_8, x_9, x_{10}, especially x_6, with the scatterplot view. No one got the correct answer for x_6 that weighted distances to five Boston employment centers had a negative effect on housing price in Boston. We observed that for these variables, their correlations to the dependent variable can only be observed if the effects of other variables were eliminated. Our approach showed its strength in revealing such hidden correlations.

The variable by variable average confidence rating is shown in Figure 8b. It is surprising to see that the subjects were fairly confident about their answers with the scatterplot view, even when the correctness of their answers was pretty low. This phenomenon reveals a crisis in existing multivariate visualization systems: users often perceive wrong correlations among multiple dimensions from the display, and they are fairly confident with the wrong insights. It

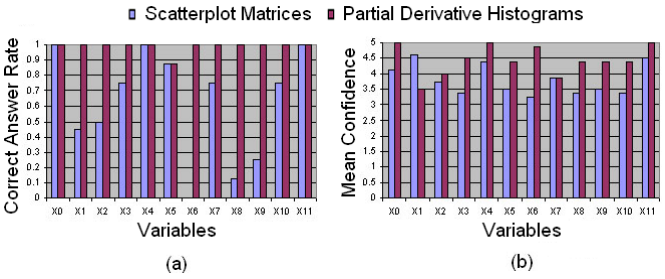


Figure 8: The user study result

seems that it is critical to introduce multivariate visual explanation techniques, such as the techniques presented in this paper, into existing multivariate visualization systems. The answers to the post-test questionnaire also showed a higher preference to the derivative view than the scatterplot view.

7 Conclusion

In this paper, we present a novel Multivariate Visual Explanation approach to support determination of correlations among multiple variables. This approach tightly integrates partial derivative calculation, inspection, and visualization into a multivariate visualization system. Our case studies and user study show how this approach was effectively used to facilitate interactive dimension reduction, multivariate model construction, and user understanding of multivariate relationships for high dimensional datasets.

Multivariate visual explanation is a challenging topic and there is much more work to be completed. In the future, we will provide visual aid and automatic techniques to facilitate users in detecting more complex correlations in interactive model construction. We would like to integrate more analysis techniques, such as generalized additive models [11] and surface response analysis [3], into the MVE approaches and increase the scalability of our system in the number of dimensions and the types of data it supports. Integrating techniques to help users determine independent and dependent variables when no semantic information is provided is also an interesting extension to the system. More user studies will be conducted to evaluate the overall effectiveness of the MVE approaches.

Acknowledgements

This work is partially supported by UNCC internal faculty research grant 1-11436 and NIH grant 1R01GM 073082-0181. It is also partially supported by the National Visualization and Analytics Center (NVAC(TM)), a U.S. Department of Homeland Security Program, under the auspices of the Southeastern Regional Visualization and Analytics Center. NVAC is operated by the Pacific Northwest National Laboratory (PNNL), a U.S. Department of Energy Office of Science laboratory .

References

[1] R. Amar. and J. Stasko. Knowledge task-based framework for design and evaluation of information visualizations. *Proc. IEEE Symposium on Information Visualization*, pages 143–149, 2004.

[2] Boston neighborhood housing price dataset. http://lib.stat.cmu.edu/S/Harrell/data/descriptions/boston.html.

[3] G. Box and N. Draper. *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, 1987.

[4] S. Brahim-Belhouari, M. Kieffer, G. Fleury, L. Jaulin, and E. Walter. Model selection via worst-case criterion for nonlinear bounded-error estimation. *IEEE Transactions on Instrumentation and Measurement*, 49(3):653–658, 2000.

[5] W. Cleveland and M. McGill. *Dynamic Graphics for Statistics*. Wadsworth, Inc., 1988.

[6] N. Draper and H. Smith. *Applied Regression Analysis*. John Wiley and Sons, 1998.

[7] A. el-sallam, S. Kayhan, and A. Zoubir. Bootstrap and backward elimination based approaches for model selection. *Proc. 3rd International Symposium on Image and Signal Processing and Analysis*, pages 238–247, 2007.

[8] M. Friendly. Extending mosaic displays: Marginal, partial, and conditional views of categorical data. *Journal of Computational and Graphical Statistics*, pages 8:373–395, 1999.

[9] G.Cain and J. Herod. *Multivariable Calculus*. Georgia Tech, 1997.

[10] D. Harrison and D. Rubinfeld. Hedonic prices and the demand for clean air. *J. Environ. Economics & Management*, 5:81–102, 1978.

[11] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[12] T. Ho and T. Nguyen. Visualization support for user-centered model selection in knowledge discovery in databases. *Proc. 13th International Conference on Tools with Artificial Intelligence*, pages 228–235, 2001.

[13] A. Inselberg. The plane with parallel coordinates. *Special Issue on Computational Geometry, The Visual Computer*, 1:69–97, 1985.

[14] J. Mcclave and T. Sincich. *Statistics (10th Edition)*. Prentice Hall. Inc, 2003.

[15] J. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[16] D. Keim, H.-P. Kriegel, and M. Ankerst. Recursive pattern: a technique for visualizing very large amounts of data. *Proc. IEEE Visualization*, pages 279–286, 1995.

[17] G. Kimmel, P. Williams, T. Claggett, and C. Kimmel. Response-surface analysis of exposure-duration relationships: The effects of hyperthermia on embryonic development of the rat in vitro. *Toxicological Sciences*, pages 391–399, 2002.

[18] T. Kohonen. *Self Organizing Maps*. Springer Verlag, 1995.

[19] J. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978.

[20] J. Li. General explicit difference formulas for numerical differentiation. *Journal of Computational and Applied Mathematics*, pages 29–52, 2005.

[21] O. Linde and T. Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. *Proc. International Conference on Pattern Recognition*, 2:1–6, 2004.

[22] Q. Liu, X. Xu, and Z. Zhang. Applications of nonuniform fast transform algorithms in numerical solutions of differential and integral equations. *IEEE Transactions on geoscience and remote sensing*, 38(4):1551–1560, 2000.

[23] A. Martin and M. Ward. High dimensional brushing for interactive exploration of multivariate data. *Proc. IEEE Visualization*, pages 271–278, 1995.

[24] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. *Proc. IEEE Symposium on Information Visualization*, pages 65–72, 2004.

[25] J. Sible and J. Tyson. Mathematical modeling as a tool for investigating cell cycle control networks. *Method*, 41(2):238–247, 2007.

[26] J. Wang. Wavelet approach to numerical differentiation of noisy functions. *Communication on Pure and Applied Analysis*, 6(3):873–897, 2007.

[27] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. *Proc. IEEE Symposium on Information Visualization*, pages 157–174, 2005.

[28] Xmdvtool home page. http://davis.wpi.edu/~xmdv.

[29] J. Yang, A. Patro, S. Huang, N. Mehta, M. Ward, and E. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. *Proc. IEEE Symposium on Information Visualization*, pages 73–80, 2004.

[30] J. Yang, M. Ward, E. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. *Eurographics/IEEE TCVG Symposium on Visualization*, pages 19–28, 2003.