Flow-based Scatterplots for Sensitivity Analysis

Yu-Hsuan Chan * Carlos D. Correa † Kwan-Liu Ma ‡

University of California at Davis

ABSTRACT

Visualization of multi-dimensional data is challenging due to the number of complex correlations that may be present in the data but that are difficult to be visually identified. One of the main causes for this problem is the inherent loss of information that occurs when high-dimensional data is projected into 2D or 3D. Although 2D scatterplots are ubiquitous due to their simplicity and familiarity, there are not a lot of variations on their basic metaphor.

In this paper, we present a new way of visualizing multi-dimensional data using scatterplots. We extend 2D scatterplots using sensitivity coefficients to highlight local variation of one variable with respect to another. When applied to a scatterplot, these sensitivities can be understood as velocities, and the resulting visualization resembles a flow field. We also present a number of operations, based on flow-field analysis, that help users navigate, select and cluster points in an efficient manner. We show the flexibility and generality of this approach using a number of multidimensional data sets across different domains.

Keywords: Uncertainty, Data Transformations, Principal Component Analysis, Model Fitting

Index Terms: K.6.1 [Management of Computing and Information Systems]: Project and People Management—Life Cycle; K.7.m [The Computing Profession]: Miscellaneous—Ethics

1 INTRODUCTION

Incorporating uncertainty and sensitivity analysis in visual analytics tools is essential to improve the decision-making process. On one hand, it provides the analysts a means to assign confidence levels to the insight gained through the analysis. On the other hand, it gives tool makers a methodology for measuring and comparing the robustness of data and visual transformations.

To gain insight from complex multi-dimensional data, a number of data analysis approaches have been proposed, such as multi-dimensional scaling, projections and sampling, which reduce either the number of observations or the number of variables in a large data set [31]. With the advent of interactive graphics, a number of techniques have been made possible that alleviate the issues of complexity, large size and multi-dimensionality, such as interactive PCA [23], multi-dimensional navigation [14], among others.

The purpose of visual analytics, however, remains the same: to gain insight on possible correlations and trends in a complex data set. In this paper, we focus on a general strategy, *sensitivity analysis* (SA), which is a common approach to understand the relationships between variables and outputs.

Sensitivity analysis is the analysis of changes in the output of a transformation as we vary the inputs. When we study pairwise correlations, sensitivity analysis tells us the rate of change of one variable Y with respect to another variable X. The variables can be input random variables, in which case sensitivity indicates the variational relationship between the two, or one of them could be a derived (dependent) variable, in which case sensitivity indicates the sensitivity of the data transformation used to derive that variable.

Therefore, sensitivity analysis is essential for discovering the factors that most contribute to output variability, finding stability regions of various transformations over the data, and understanding the interaction between variables, outputs and transformations. Although numerous approaches have been proposed to find the sensitivity coefficients of transformations, we focus on differential analysis, where sensitivities are defined as the partial derivatives of a group of variables with respect to another group of variables. Differential analysis is attractive when the data and visual transformation can be defined in closed form. When this is not possible, approximating them by exploring the parameter space becomes computationally expensive. For this reason, approximations based on sampling approaches are more appropriate.

In this paper, we present a novel augmentation of traditional scatterplots, which are useful for sensitivity analysis and general exploration of multidimensional data. The key idea behind our augmentation is the analogy of scatterplots with flow. In a XY scatterplot, if the position of a data point is given by the coordinates (x,y), then the derivative $\partial y/\partial x$ is analogous to a velocity measure at that point. Therefore, one can understand a scatterplot as a scattered collection of position and velocity measures. Based on these derivatives, one can predict the positions of interpolated points in the XY space and extract a global sense of *flow*. This analogy has a number of applications for visual analysis, which we explore in this paper: (1) the explicit representation of sensitivity parameters as tangent lines helps analysts discover local and global trends in a 2D projection. (2) sensitivity parameters can be quantified to measure the complexity of a given 2D projection and find pair-wise correlations between variables. For example, one can augment an xy scatter plot with the derivatives of a third variable z with respect to, say x. When the flow appears smooth, one can safely re-project the scatterplot in the axes zy and expect a smooth transition, which helps understand how different variables are related. (3) One can cluster and select data points based on the similarity of the flow properties around each point.

To this end, we propose certain key operations on flow-based scatterplots that are not possible using traditional means: (1) Simultaneous visualization of tri-variate correlations, using the derivative of a third variable, (2) smooth transitions and navigation of multi-dimensional scatterplots, and (3) selection and clustering by streamline, which groups data points together that lie closer to the *streamlines* generated by a data point. We demonstrate the feasibility and potential uses of our approach through a number of examples in a variety of domains.

2 RELATED WORK

2.1 Multivariate Analysis

Multivariate analysis is at the core of visual analytics. Approaches can be categorized as data-centered approaches, such as regression [13], generalized additive models [19] and response surface analysis [5], or visual-centered approaches. Since data is often large and complex, data-driven approaches often employ simplifi-

*chany@cs.ucdavis.edu

†correac@cs.ucdavis.edu

‡ma@cs.ucdavis.edu

cation techniques, which either reduce the number of observations, such as binning, sampling [32] or clustering [4], or reduce the number of dimensions in the data, such as projections [27] and multidimensional scaling. Visual-centered approaches follow a different strategy, where correlations and trends emerge as salient structures in the human visual system. Often times, these approaches are coupled with interactive manipulation. For example, Jeong et al. propose to augment traditional data analysis tools such as Principal Component Analysis with interactive manipulation for a better understanding of the transformation and the data itself [23]. Yang et al. integrate analysis tools with visual exploration of multivariate data [35] using the Nugget Management System, which incorporates user interest to guide the analysis. In this paper, we present a combination of analysis and visualization tools that exploit sensitivity analysis for effective exploration and navigation of multidimensional data.

2.2 Sensitivity Analysis

Sensitivity analysis refers in general to the analysis of the variation of the outputs in a model to small perturbation of their inputs. Numerous approaches have been proposed to this end. A number of methods fall into the class of local analysis, such as adjoint analysis [6] and automated differentiation [17], where the sensitivity parameters are found by simply taking the derivatives of the output with respect to the input, $s_{ij} = \partial Y_i / \partial X_j$. Because this is usually done in a small neighborhood of the data, they are usually called local methods. Others have proposed global estimates of sensitivity, which use sampling or statistical techniques. The most common statistical method is based on variance, which provides an estimate of the sensitivity in terms of the probability distribution of the inputs [1, 7, 20, 22, 29]. Other approaches directly introduce perturbation on the input data by manipulating certain parameters and compute the ensuing variation on the output. Since it is computationally expensive to try the entire parameter space, numerous approaches use sampling-based methods as extensively surveyed by Helton et al [20]. Different simulation strategies have been applied, including random, importance and Latin hypercube sampling [21].

Frey and Patil also reviewed a number of sensitivity analysis methods [16]. Tanaka surveyed the sensitivity analysis in the scope of multivariate data analysis [30]. Specific analyses for certain common data analysis tools have been proposed. Chan et al. presented a sensitivity analysis for variance-based methods in general [7]. Cormode et al. [10], Chau et al. [8] and Ngai et al. [26] proposed extensions to perform k-means clustering on uncertain data. Similar studies have been carried out to quantify the sensitivity and uncertainty of the principal components of multi-variate data [33, 34]. Kurowicka and Cooke extended the issue of uncertainty analysis with high dimensional dependence modeling, combining both analytical tools with graphic representations [25].

Barlowe et al. [3] proposed the use of histograms and scatterplot matrices to visualize the partial derivatives of the dependent variable over the independent variables and to reveal the positive or the negative correlations between the output and the factors in a multivariate visual analysis. Correa et al. [11] used sensitivity analysis to propagate the uncertainty in a series of data transformations and propose a number of extensions to show this uncertainty in 2D scatter plots. In this paper, we generalize the idea of sensitivity visualization as flow-based scatterplots. Bachthaler et al. [2] presented the continuous scatterplot, which generates a continuous density function for a scatterplot and alleviates the issues with missing data. Our idea of flow-based scatterplots has a similar concept, which attempts to find a continuous representation of the density that explains the 2D plot. However, we use a local analysis based on derivatives to find local trends in a scattered manner.

Projection is a commonly used dimension reduction technique for multi-variate data sets, useful when visualizing high dimensional data in 2D or 3D spaces. Scatter plots are intuitive to understand when studying the relationship between two variables. However, projected points may result in clutter and overlap for large and high dimensional data sets. To solve this problem, Keim et al. [24] proposed generalized scatter plots to augment the degree of overlap and the distortion. Other augmentations have been proposed by Collins et al. [9], that enhance the spatial layout of plots with clustering information, and Shneiderman et al., [28], that link multiple substrate plots to superimpose the cross-substrate relationships.

Another issue of scatter plots is that we can only see a limited number of variables after projection. It is common to show a scatterplot matrix to enumerate all possible combinations of projections of variables, but we need an effective navigation between these different projections. Scatter dice [14] is an alternative that exploits interactive capabilities to navigate a large scatter matrix and help visual analytics. However, the evaluation and the effectiveness of a projection is a topic often overlooked in visualization. In our paper, we propose a novel mechanism for navigating the dimensions of a multidimensional dataset, based on the sensitivity of variables to one another.

3 Flow-based Scatterplots

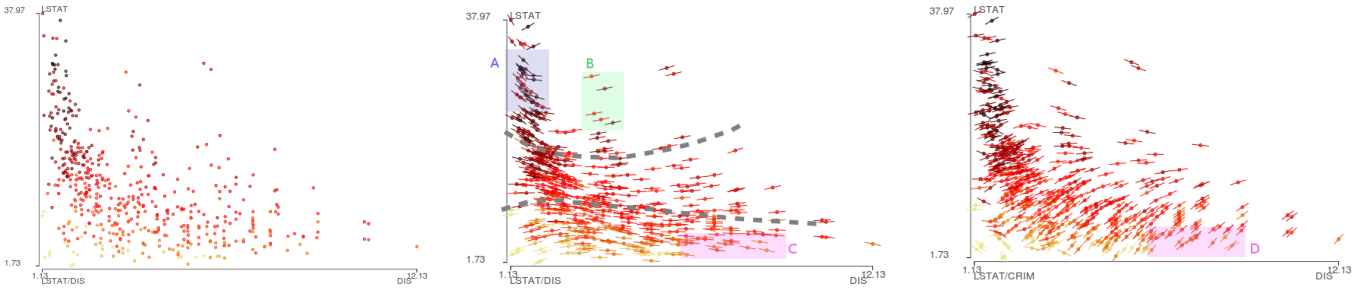
2D scatterplots are a commonly used visual representation that help see the relationship between two variables in a multidimensional data set. As shown in Figure 1(a), a 2D scatter plot is only able to show a limited number of variables as the number of visual attributes, such as position, size, color and transparency, can only be used sparingly. In certain cases, the overuse of these attributes makes it difficult to understand correlations between variables due to visual clutter.

In this paper, we propose a new type of scatterplot, called flow-based scatterplot, which augments the traditional metaphor using sensitivity information. Sensitivity refers to the change in an output variable in terms of a change in the input. In the case of a 2D scatterplot, the simplest representation of sensitivity is through an explicit depiction of the derivative of the variable in the y axis with respect to the derivative of the variable in the x axis.

To illustrate our technique, let us consider the Boston housing price data set. This data set is a collection of environmental, geographic, economic and social variables to predict the median value of housing in the Boston metropolitan area [18], which contains 506 records and fifteen continuous variables. Some variables include geographic information, such as DIS, the weighted distances to five Boston employment centers, LSTAT, the percentage of the lower status of the population, CRIM, per capita crime rate by town, and RM, the average number of rooms per dwelling, among others.

Figure 1 shows a scatterplot of two variables named DIS and LSTAT, with color encoding the median housing price. Without augmentation, this scatterplot shows the same information as in Figure 1(a). After adding the sensitivity information, we obtain a sense of the *flow* of the data. This can be seen in Figure 1(b) as a collection of line segments. The slope of that segment indicates how sensitive is the Y variable in a local neighborhood and whether that sensitivity is positive or negative. For example, we can clearly see global trends indicated by dotted lines. Moreover, it gives us an idea of more localized trends. For example, points in regions A, B and C exhibit different behavior in the LSTAT variable as we increase the DIS variable. For data points in region A, LSTAT decreases rapidly as DIS increases, while data points in region C do not change dramatically. Conversely, data points in B increase in LSTAT as DIS increases. Therefore, such sensitivity visualization not only helps users understand how variables behave toward changes in another, but also recognizes whether data points have differed locally in terms of sensitivity.

One of the advantages of plotting sensitivity is that now we can represent more dimensions in a single plot. Therefore, we are less



(a) (b) (c)

Figure 1: (a) Traditional scatter plot between two variables (b) Sensitivity visualization of the same two variables, where data points are augmented with derivatives. (c) Sensitivity visualization for a third variable, useful for analyzing tri-variate correlations.

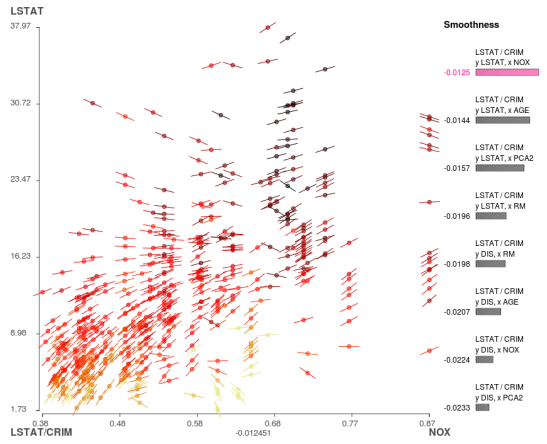


Figure 2: Smoothness ranking view. Next to each augmented scatterplot, we show the smoothness ranking of other variables.

bound by the inherent loss of information that occurs when projecting high dimensional data into a 2D space. In our case, we can plot the sensitivity of another variable with respect to one of the scatterplot axes. If we show the derivatives of a variable with respect to another variable (different from the ones used in projection), then we can begin making queries and formulating hypotheses about tri-variate correlations, instead of bi-variate queries that are typical of 2D scatterplots. An example is shown in Figure 1(c), where we show the same data points as before, using two variables named LSTAT and DIS, but we plot the sensitivity of the variable in Y (LSTAT) with respect to another variable named CRIM. Note that, although the data points have the same location in the X-Y plane, the sensitivities differ. We immediately have a different sense of flow, which changes the way we begin to formulate hypotheses about the three variables. For example, we see that, for points in region D in Figure 1(c), variable CRIM increases as DIS increases, but the same cannot be said about LSTAT, which only seems to increase when LSTAT is larger and decreases when LSTAT is low. Therefore, we may regard sensitivity derivatives as another attribute of nodes that represents relationships between two particular variables. Sensitivity derivatives of U with respect to V shows the relationship between U and V for each data point, and the projection variables (X, Y) decide where to locate these nodes of such derivative attribute. Some particular projections might place these nodes in a way that show global trends and correlations between variables U and V, which helps us understand the relationship between both U and V, and X and Y. In this paper, we show a number of operations, based on flow analysis, to help us identify these relationships.

3.1 Computing Sensitivities

As described before, there are different ways to compute the sensitivity of one variable with respect to another. In this paper, we follow a variational approach, where the sensitivity can be approximated by the partial derivative of one variable with respect to another. Since we do not know the analytic closed form of the function between two variables in the general case, we approximate the partial derivatives using linear regression. Because we do this in different neighborhoods around each point, we employ the method of moving least squares. We obtain the partial derivatives of a variable y with respect to x considering the Taylor approximation of y around a given point (x_0, y_0):

$$y_i = y_0 + \frac{\partial y}{\partial x}(x_i - x_0) \tag{1}$$

Then, we approximate the partial derivatives for point (x_0, y_0) in a neighborhood of N points, as:

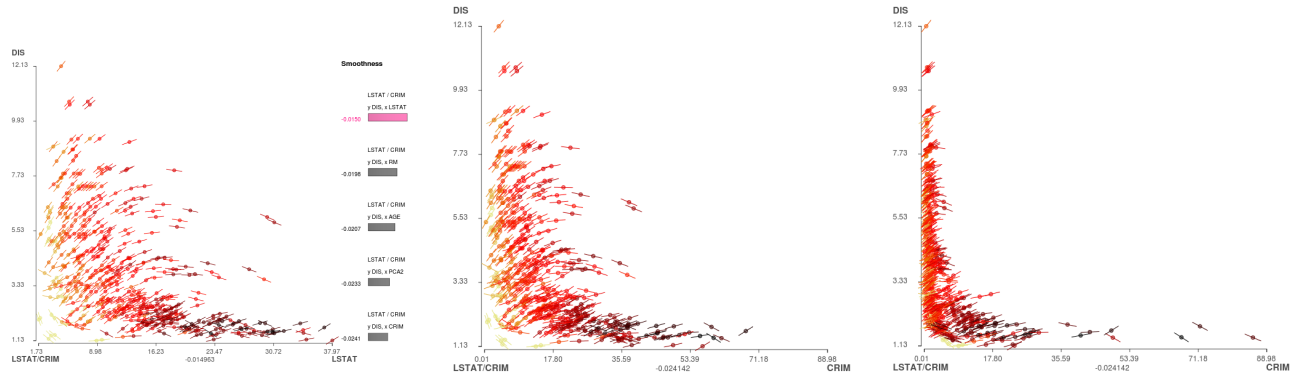
$$\frac{\partial y}{\partial x} \approx \frac{\sum_{i=0}^{N}(y_i - y_0)(x_i - x_0)}{\sum_{i=0}^{N}(x_i - x_0)^2} \tag{2}$$

With this information, we augment the scatterplot using tangent line segments on each data point. Each tangent line is computed as follows. For a given point (x_0, y_0), we trace a line between points $(x_0 - \delta vx, y_0 - \delta vy)$ and $(x_0 + \delta vx, y_0 + \delta vy)$, where $(vx, vy) = normalize(1, \frac{\partial y}{\partial x})$ and δ is a parameter that controls the length of the tangent lines.

In our experiments, we compute the neighborhood of N points as an isotropic region around each point of a radius W. This radius controls how local or global is the flow. When W is small, the derivatives capture the local variability of data and reveal localized trends. On the other hand, when W is large, the flow represents the global trend in the data. An example is shown in Section 5.2. The variable width helps us reveal local trends where the global correlation is low. Instead of making an automatic decision in terms of correlation, flow-based scatterplots offer the option to the analyst to explore the spectrum of trends and correlations interactively.

3.2 The Smoothness of a Flow Scatterplot

As can be seen from Figures 1(b-c), sensitivities provide a sense of *flow* of the data points in the projection space. This flow helps reveal overall trends. For some projections, these sensitivities show certain critical regions, where linear trends coincide at some point but then diverge. This means that data points in that region can either go up or down, possibly depending on other variables. This suggests that this particular projection is hiding a lot of complexity that may be identified through a different projection. To measure the complexity of a flow-based scatterplot, we turn to second derivatives, which tell us how fast the tangent lines change in a



(a) (b) (c)

Figure 3: Navigation between a projection LSTAT-DIS (a) to a projection CRIM-DIS (c). Since the derivative of LSTAT with respect to CRIM in the projection is detected by the system as one of the smoothest, the user changes the projection according to this derivative. We see a smooth transition in the x-axis from LSTAT to CRIM (b) which helps users maintaining the context of the transformation. Most of the data points move forward left in this linear transformation.

given neighborhood. If the tangent lines do not change much, the scatterplot has a maximum smoothness, or a small second derivative. On the other hand, if the tangent lines vary drastically in small neighborhoods, the scatterplot has low smoothness, or a high second derivative.

To compute the second derivative in a neighborhood around a data point (x_i, y_i), we follow the same moving least squares approach. Let C_i denote the local complexity (or unsmoothness) around a point, computed as the linear regression of the first derivatives or sensitivities:

$$C_i \approx \frac{\sum_{j=0}^{Neighborhood} (\frac{\partial y}{\partial x}|_{x_j,y_j} - \frac{\partial y}{\partial x}|_{x_i,y_i})(x_j - x_i)}{\sum_{j=0}^{Neighborhood} (x_j - x_i)^2} \quad (3)$$

Then, the total smoothness of the plot can be computed as an inverse sum of the local complexity for all N data points:

$$S = \frac{1}{\sum_{i=0}^{N} C_i} \quad (4)$$

Note that the unsmoothness of $node_i$ in Equation 3 has the same form as the sensitivity derivative in Equation 2. This Equation 3 measures how this node differs from its neighbors in terms of its sensitivity $\frac{\partial y}{\partial x}$. The larger the difference from it to its neighboring nodes,the less smooth the trend around this neighborhood is, since this node has drastic change in slope from others. In Section 4, we show that this measure is useful for ranking the variables in a high-dimensional data set and provide an intelligent way to navigate between different dimensions.

3.3 Scatterplot Streamlines

One of the contributions of our paper is a different way of looking at scatterplots that uses flow as a metaphor. According to this metaphor, if the location of a data point in an XY plot represents position, then the sensitivity coefficient, shown as a small tangent line, can be understood as a *velocity*. In 2D flow visualization, it is common to represent the stationary directions of flow using streamlines, which integrate the velocities to simulate the path that a particle would take if placed in this flow. To show more global trends, we employ the same scheme. Since we have a scattered collection of points, we use a *scattered integration* scheme, which computes new directions based on the local velocity (or derivative), in terms of the neighboring elements. To integrate the derivatives along the streamline, we used second order Runge-Kutta. A streamline spanned by a point p_0 in the 2D domain is a series of connected points found using the following recursive method. For a point p_k, the next point in the streamline p_{k+1} is found as:

$$p'_k = p_k + hv(p_k) \quad (5)$$

$$p_{k+1} = p_k + hv(p'_k) \quad (6)$$

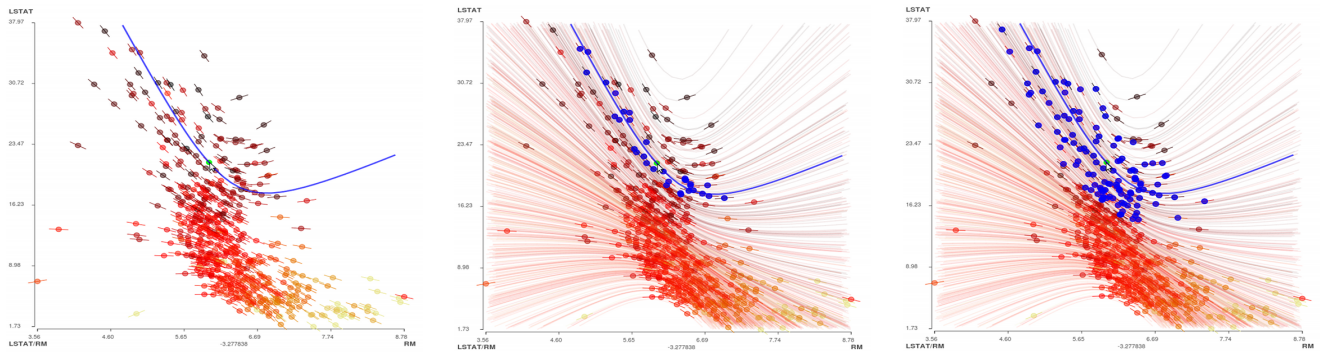
where h is the discretization distance between consecutive points in the streamline and $v(p)$ is the derivative evaluated at point p. We apply this mechanism forwards and backwards in time (with positive and negative h, respectively) and stop a line at the boundaries of the scatterplot. Note that we only compute streamlines for the derivatives of the Y variable with respect to X. For a third variable, using sensitivity as segments is useful, but the use of streamlines may be misleading, as the evolution of data points is no longer defined in the 2D space. Computing and drawing a single streamline of a data node of interest is real-time and highly interactive, as the dark blue line of the selected node in green in Figure 4(a). Therefore we can hover over different data points to examine the streamline of the node, and augment the streamline with different length and selection distance. Computing streamlines takes in the worst case O(N) time, but with the use of spatial data structures, it can be done in O(logN) time. During the computation of all streamlines of a plot, users can interact with the smoothness ranking view, such as changing to another projection, hovering over the node of interest to show its streamline, and selecting node by the selecting streamline.

4 Operations on Flow-Based Scatterplots

In the previous section, we have shown a number of metaphors based on 2D flow visualization, such as tangent lines and streamlines, that help us augment traditional scatterplots to highlight the sensitivity and possible correlations in multi-dimensional data. In this section, we describe how we can exploit this metaphor to enable novel operations on 2D scatterplots.

4.1 Multi-dimensional navigation

As mentioned in the previous subsection, flow-based scatterplots allow us to depict information of other variables than the ones used in the two main axes of the plot, in the form of sensitivity lines. However, trial-and-error exploration of all combinations of variables in the search for insightful correlations proves impractical. For a data set of M variables, the number of combinations of augmented plots can be up to M^4. Although scatterplot matrices are common to depict all pairwise correlations in a multidimensional data set, a similar matrix for augmented scatterplots proves impractical. Therefore, we need a systematic way to evaluate the flow-



(a) (b) (c)

Figure 4: Selection by a streamline allows us to group neighboring data points by a particular trend pattern of interest, with different selection distance d. (a) A streamline in blue of a selected data pointed in green (b) Selection of a small d = 0.03 (c) Selection of a larger d = 0.10. Note that the line segments of sensitivity and the background streamlines are computed with a larger neighborhood parameter W=10.0, and thus they depict the vertical patterrn of the overall data distribution.

based scatterplots, in the search for the *salient* combinations that help reveal correlations and patterns.

Here, we present a method that evaluates a scatterplot to find the variables that are interesting, in terms of the smoothness of the sensitivity derivatives in the projection, and that should be used in subsequent projections and enable intelligent navigation of the multi-dimensional space.

Our approach is as follows: for a given projection XY, we measure the smoothness of the flow-based scatterplot with derivatives dZ/dX with respect to every variable Z. We then rank the variables Z in terms of their smoothness. The rationale for the ranking is quite simple. Let us assume that a projection XY with derivatives dZ/dX has a maximal smoothness, i.e., its second derivative is 0. Therefore, if $\partial^2 Z/\partial X^2 = 0$, then the derivative $\partial Z/\partial X$ is a constant and variables Z and X are related linearly. Therefore, if we reproject the points in axes ZY, we expect data points to move smoothly in the new x-direction, by a linear factor, which is easy to follow and comprehend while doing the re-projection. Now, as the complexity of the scatter plot increases (i.e., it becomes less smooth), the relationship between the transition variables Z and X increases in complexity and is more difficult to understand when we reproject along these variables. For this reason, the ranking provides a mechanism to intelligently navigate the data.

Figure 2 shows the visual design of the ranking view using the Boston housing price data set. On the left, we show a list of different projections of the data. In this case, we show a multidimensional data set of Boston housing prices and the variables that affect them. In this view, we plot a variable called NOX (pollution) vs. LSTAT (A socio-economical variable). On the left, we see a ranking of other candidate variables that the user can explore, namely AGE, DIS, PCA2, etc. Hovering over each of these variables shows the corresponding derivatives on the XY projection. Once a variable is selected, the scatterplot smoothly changes the projection to ZY, where Z is the new selected variable. The ranking provides a way for the analyst to pick the variables that have the smoothest re-projection. We believe this method of guided navigation is more intuitive than arbitrary re-projection, even if we alleviate the issues of re-projection with rotation and 3D projection, as obtained in systems like scatter dice [14]. An example is shown in Figure 3. In this ranking view, we found that the projection (LSTAT,DIS) with derivatives $\partial LSTAT/\partial CRIM$ has the maximal smoothness, and the derivative indicates that variables CRIM and LSTAT are closely related. The second figure shows an example of the reprojection halfway between LSTAT and CRIM. We can see the smooth linear transition at this time. The third figure shows the newly reprojected image in the LSTAT-CRIM plane.

4.2 Selection by streamline

Another issue with scatterplots is the selection of meaningful groups. A number of ways have been proposed to select groups of data points. The simplest one is by dragging a rectangular region in the scatterplot and selecting all points that fall inside that region. Unfortunately, this method often groups data points that may not be related or that are projected together in that particular view. Another mechanism is *brushing*, which allows the user to select an arbitrary region by "painting" the regions in the 2D plot. In this paper, we present another method, which uses streamlines. Streamlines, as described before, are lines that represent the imaginary flow of a particle in a given XY plot, by following the sensitivities of the corresponding points along that path. In a sense, a streamline represents the predicted change in the variable Y as we increase the variable X. Therefore, it is intuitive to select elements that are *near* a streamline, since those points locally exhibit a similar trend along the selected streamline. This trend, in turn, may help discover interesting correlations between the variables in the scatterplot that cannot be identified from the projection itself.

To select points based on a streamline, we allow the user to explore the streamlines interactively. When the user hovers over a data point, we show the streamline that emanates from this point, both forward and backwards in time (i.e., the streamline represents the trend for values of X before and after the data point). Then, we can select points in the 2D plot that have a distance d to the streamline less than a given threshold. An example is shown in Figure 4 for the Boston data set. Here, we show a scatterplot of LSTAT vs RM (average number of rooms per dwelling). As we increase the distance d of selection, we can pick data points that locally share a similar trend to the one selected. This is a good alternative to brushing and rectangular selection that highlights the linear relationship between the two variables.

4.3 Clustering by streamline

Finally, we see that, depending on the streamline we select, we get different groupings of data, all of which highlight certain degree of similarity between the local trends of each data point. This prompted us to try a more automatic approach, which clusters data together based on streamline. This clustering is in fact a classification of data points by their local trend and vicinity. If two points produce similar streamlines, we expect them to be related, since they predict a similar behavior as we vary the variable in the x-axis. Once we identify clusters, the classification may hint at different critical regions, where the local trends change dramatically. Since the streamlines represent trends, we hypothesize that the cluster-

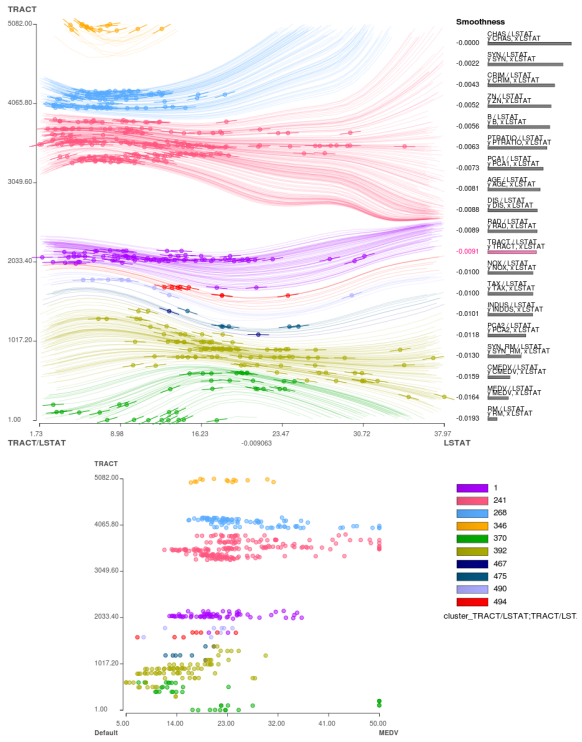


Figure 5: Clustering by streamlines. By applying streamlines to the flow, we can now classify data points based on similar trends. When we do that, we obtain a classification that may help identify salient data points or critical regions. In this case, we classify the Boston housing dataset in terms of the sensitivity of variable TRACT with respect to LSTAT. These ten groups are not linearly separable in this (LSTAT, TRACT) projection. In the bottom figure we show that these groups are in another projection (MEDV, TRACT).

ing based on streamlines provides a more robust classification of data. To classify streamlines, we follow a bottom-up hierarchical approach where each streamline is considered as its own cluster initially, and then clusters are merged hierarchically based on their similarity. We use Euclidean distance to compute the similarity of two streamlines.

An example is shown in Figure 5. In this projection (LSTAT,TRACT), we cluster those streamlines constructed from sensitivity measured by the derivative of TRACT with respect to LSTAT, which results in six main clusters (numbered as No.346, 268, 241, 1, 392, and 370 respectively) whose centroids lie diagonally in the plot, suggesting an inverse relationship between LSTAT and TRACT. Nodes are colored by such streamline clustering results. This clustering helps identify groups when used in another projection, such as when looking at classes in the median housing value (MEDV) in the scatterplot below.

5 Results

In addition to the Boston housing price dataset, we explored different aspects on three multidimensional datasets.

5.1 Iris

The iris data set consists of 150 records and 4 variables regarding the classification of a number of species of the iris plant, including the length and width of the sepal and petal of the plant [15]. One of the challenges of this data set is that two of the classes, namely Iris-virginica and Iris-versicolor, cannot be linearly separated. In Figure 6 we illustrate the use of clustering by streamline to find the variables involved in this classification. On top, we see the three classes in terms of the Petal and Sepal length of the iris plant. One of the two classes is clearly separated. In the middle, we show the result of applying clustering by streamline using three clusters. We see that this clustering gets close to the actual classification of the plants, except for five points, highlighted in circles. Compare to the image at the bottom, where we classify the data using k-means based on the 2D proximity in this projection. Clearly, clustering by streamline behaves better, which indicates that the classification can be explained not in terms of the variables themselves, but in terms of their derivatives.

5.2 Forest Fires

The forest fires data set comprises 517 records of forest fires in Portugal, to predict the occurrence and size of forest fires in terms of environmental and meteorological properties, such as temperature and wind [12]. We use this example to illustrate the effects of neighborhood size when computing the derivative to reveal local and global trends. In Figure 7, we show a scatterplot of two variables, DMC (Duff Moisture Code) and DC (Drought Code), which are both indexes used by the Fire Weather Index to measure the danger involved in a fire. We see some local linear trends in the midst of a more global linear trend. In Figures 7(b-d), the color shows the result of clustering by streamline for varying neighborhoods with $W = 0.1, 2.5$ and 10.0, respectively. When W is small, the streamlines follow individual linear trends. For $W = 2.5$, the clustering now reveals that data has a rather horizontal trend when data points are grouped together in larger neighborhoods, indicating increasingly large variance in the X dimension. When the variance is low, such as in the cluster in blue, the streamlines follow a similar trend to the local one. However, for the purple group, where the variance is large, a different trend emerges. For a large W, on the other hand, we are able to extract the global trend, which, in this case, happens to align to the local trends nicely. This is usually the case for correlated variables. Choosing a right neighborhood depends on the size of features the analyst wants to identify and whether correlations in the data can be explained locally or globally. In our experiments, W is a free parameter tuned by the user.

5.3 Wine

The wine data set comprises 13 variables of 178 observations of the chemical composition of wines growing in a particular region in Italy and the relationship to color intensity and hue. Figure 8 shows that we have found a smooth flow-based scatterplot of (Alcohol,Color) and examined this in the smoothness ranking. We see a rather simple distribution of sensitivities, as shown in Figure 8(a). From this view we would like to know which variable is a good reprojection to navigate next. By viewing the scatterplot of the same Y (Color) axis but different X axis, we found that AlcAsh variable may be a good candidate for x-axis in the new projection. Therefore, we examined the relationship between the previous x-axis Alcohol and the new x-axis candidate AlcAsh by computing the streamline-augmented scatterplots of each. Figure 8(b) is a scatterplot of (Alcohol, AlcAsh) with the derivative of AlcAsh over Alcohol, while Figure 8(c) it is a scatterplot of (AlcAsh, Alcohol) with the derivative of Alcohol over Alcash. Both the scatterplots have smooth streamlines, which means that the change in one of these two variables does not cause a dramatic change in the other, and suggests that they are interchangeable. Thus we changed the x axis from Alcohol to AlcAsh, as shown in Figure 8 (a) and (d). We see that this transition is indeed smooth as verified by the smooth animation. Also, we compare the clustering by streamlines for the two views (a) and (d), as shown in (f) and (d) respectively. We can see that the clustering results of these two views are very similar. They both contain a large main cluster at the bottom, a smaller cluster at the top of the main cluster, and the rest of data points at

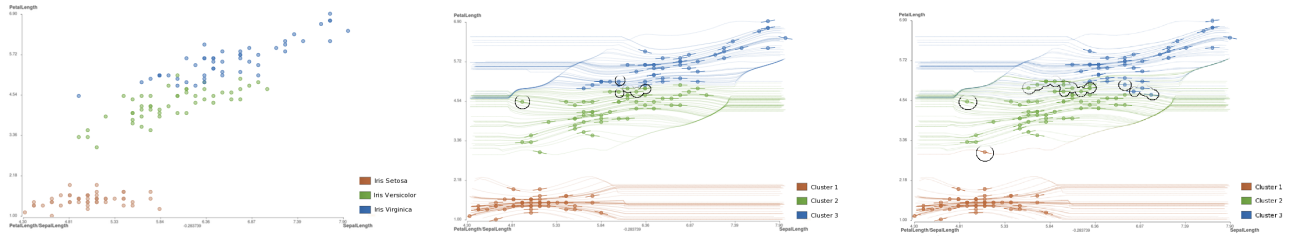
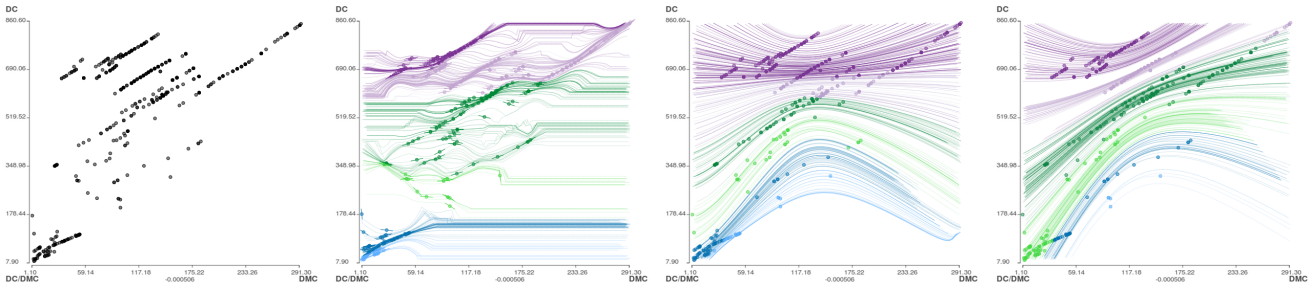
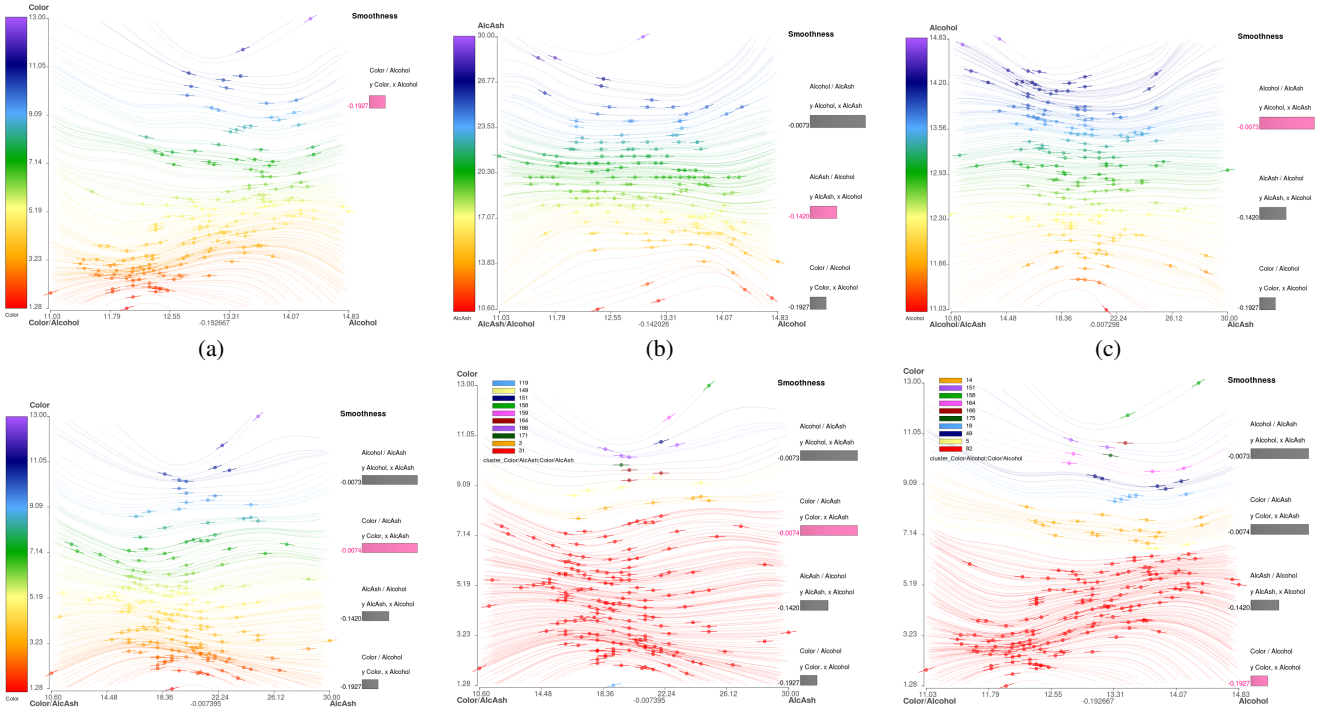


Figure 6: Visual exploration and streamline-based classification of the Iris data set. Although two of the classes are not linearly separable, the streamline classification helps identify the two groups visually in terms of their trend. Now the difference between the two classes can be explained in terms of, not only the two variables, but also the partial derivatives.



(a) Scatter plot (b) W = 0.1 (c) W = 2.5 (d) W = 10.0

Figure 7: Flow-based clustering for the Forest fire data set. Clustering helps identify groups regarding the relationship between the two indexes, DC and DMC, that are used by the FIre Weather Index to meature the danger involved in a fire. We change the neighborhood size parameter W to show how local and global trend can be revealed by streamlines and clusters.



(a) (b) (c)

(d) (e) (f)

Figure 8: Example navigation using a Wine dataset. The sensitivity ranking view suggests smooth transition between projections from Alcohol to Alcalinity of Ash (AlcAsh).

the top are classified to eight different clusters because these nodes have very different local trends.

6 Conclusions and Future Work

We have presented a novel visual representation of scatterplots useful for sensitivity analysis. Analogous to traditional scatterplots, which help elucidate pairwise correlations between two variables, flow-based scatterplots help us understand correlations between the change in one variable with the change in another. In addition, they help us understand tri-variate correlations, and we can formulate hypotheses of the relationship between two variables and the rate of change of another. In our proof of concept we have introduced visual analysis methods that help discover patterns in the data difficult to obtain through linear analysis. For example, we have shown that selection by streamline helps group points in a non-linear manner, often aligned with the boundaries of classes. The ranking of sensitivities is a novel way of navigating multidimensional data sets that combines automated analysis with visual and interactive control. We envision that, as data sets become larger and more complex, a combination of both analysis and visualization is critical. In our future work, we will explore the use of flow-based analysis to improve the classification of complex data, and also extend it to visualize the workings of parameterized models from machine learning. Moreover, we would like to conduct a thorough user study to verify whether users can interpret flow-based scatterplots and draw correct conclusions.

Acknowledgements

This research was supported in part by the U.S. National Science Foundation through grants CCF-0938114, CCF-0808896, CNS-0716691, and CCF-1025269, the U.S. Department of Energy through the SciDAC program with Agreement No. DE-FC02-06ER25777 and DE-FG02-08ER54956, and HP Labs and AT&T Labs Research.

References

[1] Leon M. Arriola and James M. Hyman. Being sensitive to uncertainty. *Computing in Science and Engg.*, 9(2):10–20, 2007.

[2] Sven Bachthaler and Daniel Weiskopf. Continuous scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1428–1435, 2008.

[3] S. Barlowe, Tianyi Zhang, Yujie Liu, Jing Yang, and D. Jacobs. Multivariate visual explanation for high dimensional datasets. pages 147–154, Oct. 2008.

[4] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

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

[6] D. Cacuci. *Sensitivity and Uncertainty Analysis: Theory Vol.1*. Chapman & Hall/CRC, 2003.

[7] Karen Chan, Andrea Saltelli, and Stefano Tarantola. Sensitivity analysis of model output: variance-based methods make the difference. In *WSC '97: Proceedings of the 29th conference on Winter simulation*, pages 261–268, Washington, DC, USA, 1997. IEEE Computer Society.

[8] Michael Chau, Reynold Cheng, Ben Kao, and Jackey Ng. Uncertain data mining: An example in clustering location data. In *Proc. of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006)*, pages 199–204, 2006.

[9] Christopher Collins, Gerald Penn, and Sheelagh Carpendale. Bubble sets: Revealing set relations over existing visualizations, 2009.

[10] Graham Cormode and Andrew McGregor. Approximation algorithms for clustering uncertain data. In *PODS '08: Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 191–200, New York, NY, USA, 2008. ACM.

[11] Carlos D. Correa, Yu-Hsuan Chan, and Kwan-Liu Ma. A framework for uncertainty-aware visual analytics. In *IEEE VAST 2009 Symposium*, pages 51–58, 2009.

[12] P. Cortez and A. Morais. A data mining approach to predict forest fires using meteorological data. In *J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, pages 512–523, 2007.

[13] Norman R. Draper and Harry Smith. *Applied Regression Analysis (Wiley Series in Probability and Statistics)*. John Wiley & Sons Inc, 2 sub edition, 1998.

[14] N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1539 –1148, nov.-dec. 2008.

[15] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179 – 188, 1936.

[16] H.C. Frey and S.R. Patil. Identification and review of sensitivity analysis methods. *Risk Analysis*, 22(3):553–578.

[17] Andreas Griewank. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[18] David Jr. Harrison and Daniel L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, March 1978.

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

[20] J.C. Helton, J.D. Johnson, C.J. Sallaberry, and C.B. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety*, 91(10-11):1175 – 1209, 2006. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004) - SAMO 2004.

[21] R. L. Iman and J. C. Helton. An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Analysis*, 1(8):71–90, 1988.

[22] Michiel J.W. Jansen. Analysis of variance designs for model output. *Computer Physics Communications*, 117(1-2):35 – 43, 1999.

[23] Dong Hyun Jeong, Caroline Ziemkiewicz, Brian Fisher, William Ribarsky, and Remco Chang. ipca: An interactive system for pca-based visual analytics. *Comput. Graph. Forum*, 28(3):767–774, 2009.

[24] Daniel A. Keim, Ming C. Hao, Umeshwar Dayal, Halldor Janetzko, and Peter Baka. Generalized scatter plots. *Information Visualization*, 2009.

[25] Dorota Kurowicka and Roger Cooke. *Uncertainty Analysis with High Dimensional Dependence Modelling*. John Wiley and Sons, 2006.

[26] Wang Kay Ngai, Ben Kao, Chun Kit Chui, R. Cheng, M. Chau, and K.Y. Yip. Efficient clustering of uncertain data. pages 436–445, Dec. 2006.

[27] Jonathon Shlens. A tutorial on principal component analysis, December 2005.

[28] Ben Shneiderman and Aleks Aris. Network visualization by semantic substrates. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):733 –740, sept.-oct. 2006.

[29] I. M. Sobolá. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Math. Comput. Simul.*, 55(1-3):271–280, 2001.

[30] Yutaka Tanaka. Recent advance in sensitivity analysis in multivariate statistical methods. *Journal of the Japanese Society of Computational Statistics*, 7(1):1–25, 1994.

[31] James J. Thomas and Kristin A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

[32] Steven Thompson. *Sampling*. John Wiley , Sons, Inc., 1992.

[33] Václav Šmídl and Anthony Quinn. On bayesian principal component analysis. *Comput. Stat. Data Anal.*, 51(9):4101–4123, 2007.

[34] Yoshihiro Yamanishi and Yutaka Tanaka. Sensitivity analysis in functional principal component analysis. *Computational Statistics*, 20(2):311–326, 2005.

[35] Di Yang, E. A. Rundensteiner, and M. O. Ward. Analysis guided visual exploration of multivariate data. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 83–90, 2007.