FSelector: Variable Selection Using Visual Features

Tommy Dang*
Texas Tech University

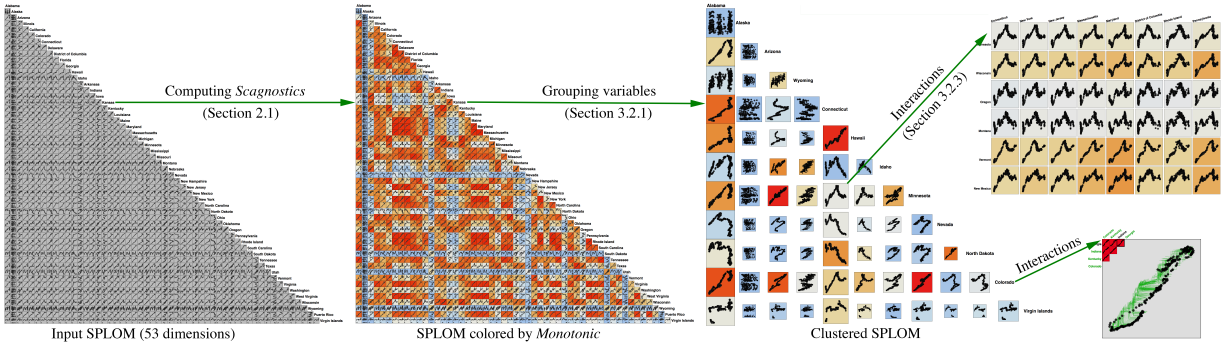

Figure 1: Major processing stages in *FSelector* visualization for an input data with 53 variables. The data were retrieved from the U.S. Bureau of Labor Statistics (BLS) website. Red scatterplots are high *Monotonic* while blue scatterplots are low *Monotonic*.

Abstract

Visual representation of large datasets should allow us to focus on essential dimensions when restricted to limited visual space. This paper presents an approach for abstracting multi-dimensional data with a focus on grouping the individual attributes based on visual features (or *Scagnostics*) such as density, skewness, shape, outliers, and texture. Working directly with these visual characterizations, we propose a prototype, called *FSelector*, to guide users when interactively exploring high dimensional datasets. In particular, selected (leading) variables are organized in a grid layout, allowing users to rapidly identify interesting pairs of variables and to focus on analyzing the original variables directly.

Index Terms: Human-centered computing—Information Visualization—Visualization techniques—Scagnostics

1 Introduction

Visual analysis can support the use of meaningful features or variables when developing a model. But visual analysis of high-dimensional data is still a challenging task, often condensed in the catchphrase "curse of dimensionality." One aspect of this problem is: distances between points tend toward a constant as the number of dimensions grows and becomes infinite in the limit. Faced with this problem, we need to find some relevant subset of variables that convey interesting patterns and structures in the high-dimensional data. Our approach is to detect, rank, and group relevant dimensions of a high-dimensional dataset, when pairs of dimensions are considered as more relevant if they share similar data distribution patterns w.r.t other variables in the data. Our method is based on nine characterizations of the 2D distributions of pairwise projections on a set of points in multidimensional space. These characterizations include measures such as density, skewness, shape, outliers, and texture. Our application is designed to handle the types of multivariate data series that are often found in security, financial, social, and other sectors.

Variable selection is one of the most fundamental research areas in multidimensional data analysis. Many variable selection algorithms include variable ranking as a selection mechanism because of its simplicity, scalability, and empirical success [17]. Several publications use variable ranking as a baseline method [10, 16]. Variable correlation is the basic method of this class. Figure 2 outlines the limitations of ranking criteria for individual variables and highlights the benefits of considering pairwise projections. In Figure 2(a), observations on two variables, identically and bimodally distributed, are displayed in a scatterplot. From a marginal perspective, the two variables appear to be redundant. However, an important class separation (between orange and blue) is revealed in the 2D projection. The histograms on the top and right display the class distributions on each dimension.

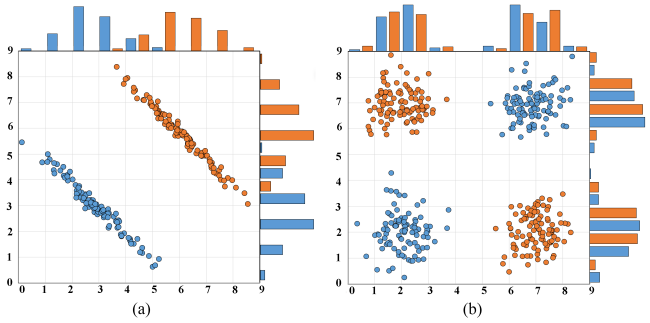

Figure 2: Two identically distributed variables: (a) Class separation is revealed by using two variables instead of one, (b) Nonlinear class separation is evident in a scatterplot, but it is not detectable in the marginal distributions.

Variables uncorrelated with a class variable can nevertheless predict a class variable perfectly when considered jointly [17]. Figure 2(b) shows one such example. The two classes consist of disjoint clusters whose densities overlap in the margins, as depicted in the histograms along each axes. When taken together, however, the two classes (highlighted in orange and blue) are completely separated.

The two above examples highlight why we need to consider joint distributions in multidimensional data analysis. (This paper focuses on 2D axis-parallel projections, but the same principles apply to higher-dimensional projections.) Furthermore, focusing only on

*Tommy Dang is with Department of Computer Science, Texas Tech University. E-mail: tommy.dang@ttu.edu

simple correlations risks concealing other structures that could be found in lower-dimensional projections. The variables in Figure 2(a) are negatively correlated, but this statistic does not characterize the real story in the plot, namely, that there are two well-separated clusters. Therefore in this paper, we consider a set of visual features which characterize different possible distributions of data points in a scatterplot [33]. The set of visual features (hereafter, *Scagnostics*) works on unlabelled data.

To examine 2D scatterplots, our first step is to extract visual features of data points distributions for each pair of variables. We then group relevant variables by the similarity of their features w.r.t other variables in the data. Finally, we present our clustering results in a scatterplot matrix (SPLOM) with multiple layers [14]. Our contributions in this paper are:

- We present a new approach for abstracting multidimensional data by grouping the individual attributes based on unusual distributional features such as convexity, compactness, or outliers.
- We propose a prototype, *FSelector*, to guide users on interactively exploring high dimensional datasets. The visual interface supports a full range of non-blocking interactions, such as zooming, brushing and linking, ordering, and filtering.
- We highlight the benefits of our approach by using *FSelector* on real-world datasets.

2 Related Work

2.1 Visual Features

Scagnostics (Scatterplot Diagnostics) were developed by Wilkinson et al. [32], based on an unpublished idea of Paul and John Tukey, in order to discern meaningful patterns in large collections of scatterplots. The Tukeys' original idea was intended to overcome the impediments involved in examining large scatterplot matrices (multiplicity of plots and lack of detail). Wilkinson's implementation enabled for the first time *Scagnostics* computations on many points as well as many plots. *Scagnostics* computations depend on proximity graphs that are all subsets of the Delaunay triangulation: the minimum spanning tree (MST), the alpha complex, and the convex hull. Figure 3 shows an example of the three geometric graphs generated on the same set of data points. The nine *Scagnostics* measures are named *Outlying*, *Skewed*, *Clumpy*, *Dense*, *Striated*, *Convex*, *Skinny*, *Stringy*, and *Monotonic*.

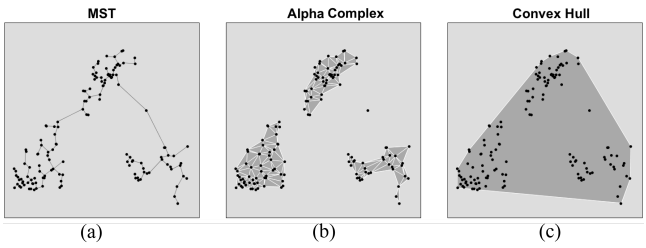

Figure 3: Three geometric graphs for computing *Scagnostics* measures: Minimum Spanning Tree, alpha complex, and convex hull.

Since introduced, *Scagnostics* has been used in many different application domains. TimeSeer [13] uses *Scagnostics* for detecting unusual distributions within multivariate time series data [12]. The application was demonstrated to find abnormalities in the US unemployment data and to detect the correlation between meteorological measurements from the Gulf of Maine in 2008. *ScagExplorer* [15] clusters similar scatterplots and provides a comprehensive summary of the 2D relations of variables in a dataset. Following in the *Scagnostics* spirit, Behrisch et al. have recently introduced Magnostics [4] for retrieving potentially interesting matrix views to support the exploration of networks. This approach ranks matrix views according to the appearance of specific visual patterns, such as blocks and lines, indicating the existence of topological motifs in the data, such as clusters, bi-graphs, or central nodes.

2.2 Variable Ranking and Selection Methods

High-dimensional data analysis has attracted a lot of attention from both the information visualization and machine learning communities. Liu et al. [24] recently provided a thorough review of the recent developments in visualizing high-dimensional data. Dimension reduction, subspace clustering, and topological features are interesting research directions in this field [23].

With the main focus on variable correlations, Yang et al. [34] propose a *Value and Relation* (VaR) display, together with a rich set of navigation and selection tools, for interactive exploration of datasets. VaR uses pixel-oriented representations to reveal data patterns, which allows VaR to handle a large number of variables in limited screen space. The VaR navigation tools allow users to identify hidden patterns. The VaR selection tools enable users to further explore dimensions of interest on demand. The Rank-by-feature framework [27] computes statistical summaries (means, standard deviations, correlations, etc.) on univariate and bivariate distributions and then ranks them to identify similar distributions. The tool supports the effectiveness of characterizing scatterplots in order to navigate a large corpus of statistical data. A review of quality metrics in high-dimensional data visualization can be found in more recent publications [1, 6, 26].

Turkay et al. [29] introduce the interactive visual identification of a manageable number of factors to facilitate the interactive visual analysis of high-dimensional datasets. Each selective factor represents a subgroup of dimensions. These factors can be iteratively refined to provide a better representation of the relations between the original dimensions. Kandogan [21] proposes just-in-time descriptive analytics, where attributes are mapped into different groups into the visual space. This analytics tool lets a user view data operations such as scaling and rotation. Joia et al. [20] present an interesting multi-dimensional projection-based visualization technique to detect clusters in visual space using a deterministic sampling scheme sensitive to unbalanced data.

Correlation analysis has been the focus of many other researchers in multidimensional visual analysis [17]. This analysis can help to reveal the higher-dimensional relationships that can exist in multivariate data. Zhang et al. [36] propose the use of a correlation map for both numerical and categorical variables. This approach visualizes data relations within the sub-spaces spanned by correlated variables by projecting the data into a corresponding tessellation of the map. Peng et al. [25] identify outliers in two dimensions using popular clustering algorithms, such as K-means. Zhao and Kaufman [37] use sorting to optimize the ordering of variables in parallel coordinates; by doing so, they hope to reveal trend and correlation information through polylines. A technique is presented by Silva et al. [11] to visually explain 2D scatterplots created by multidimensional projections. The authors propose a visual approach to detecting which dimension contributes most to similarity relationships over the projection. There are several feature selection metrics used in machine learning that can be used to rank the importance of dimensions [16, 17].

3 Design Decisions for the *FSelector* Visualization

FSelector uses *Scagnostics* to characterize the 2D data distributions of pairwise projections in multidimensional space. Variables are compared on their feature spaces w.r.t other variables in the data. Similar variables are grouped together and presented in a grid layout,

which allows users to identify interesting pairs of leading variables and to compare their sets of following variables.

The next section starts with the design motivations behind selecting and using all *Scagnostics* measures for cluster analysis. In the second part, we describe different components in the *FSelector* visualization, the overview grid layout, and possible interactions.

3.1 Motivations

The Pearson product-moment correlation coefficient is a measure of the linear dependence between two variables. Correlation criteria are often used for microarray data analysis [31] in drug discovery and development [9]. The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables. While Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). Variable correlation is a basic method for feature ranking and selection [10, 16, 34, 36]. The central premise of feature selection is that the data contains many variables that are either redundant or irrelevant, and can thus be removed without incurring much loss of information [5]. Redundant or irrelevant features are two distinct notions: one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated [17].

The *Correlation Feature Selection* measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other" [18, 35]. According to the *Correlation Feature Selection* measure, *Variable 1* and *Variable 2* in the example in Figure 4 are a good feature subset since they are highly correlated (Spearman correlation coefficient is 0.94 as depicted in the greenish last bar in the top left chart) while they are both uncorrelated to *Variable 3* (Spearman correlation coefficients are 0.34 and 0.44 respectively). However, grouping *Variable 1* and *Variable 2* creates information loss since neither *Variable 1* nor *Variable 2* is a good representative feature of this group. In other words, the pairwise projections of *Variable 1* vs. *Variable 3* and *Variable 2* vs. *Variable 3* have distinct patterns and hence their bar charts (at the bottom) of nine *Scagnostics* are very different. Especially, the *Clumpy* measure (in green) for *Variable 1* vs. *Variable 3* is much lower compared to that of *Variable 2* vs. *Variable 3* scatterplot.

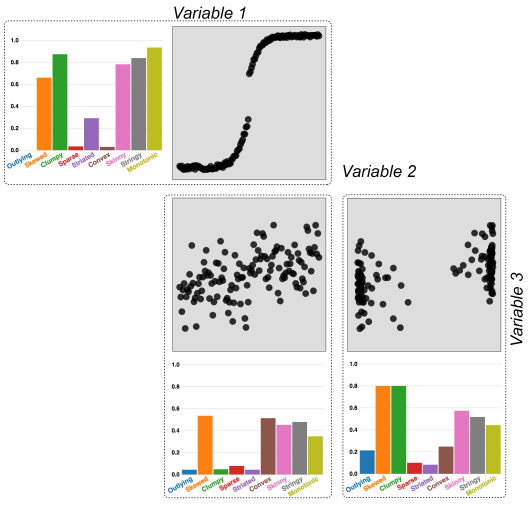

Figure 4: An example of *Correlation Feature Selection* measure containing three variables: *Variable 1*, *Variable 2*, and *Variable 3*. The *Scagnostics* signature for each pairwise projection is color-coded and displayed next to the plot in the same dashed box.

While the data in the previous example (Figure 4) were synthesized, the data in this example (Figure 5) are from the National Research Council for university rankings in 2006 (see more details in Section 4.1). The three variables examined here are *Program Size Quartile*, *Number of Students Enrolled*, and *Awards Per Allocated Faculty Member*. As depicted, *Program Size Quartile* and *Number of Students Enrolled* have a high covariance (Spearman correlation coefficient is 0.9) and high *Striated* (which is not captured by simply using correlation). On the other hand, both of these variables are relatively uncorrelated with *Awards Per Allocated Faculty Member* (Spearman correlation coefficients are 0.17 and 0.11 respectively). More importantly, the data patterns in two scatterplots (at the bottom) are distinct, and hence their *Scagnostics* signatures (bar charts at the bottom) are completely different (since *Program Size Quartile* is categorical while *Number of Students Enrolled* is continuous).

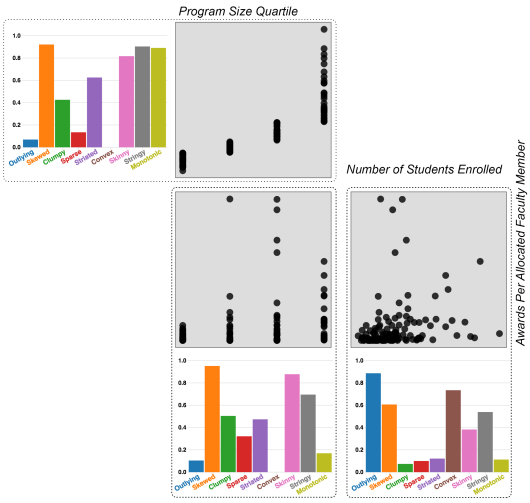

Figure 5: Second example of *Correlation Feature Selection* measure containing three variables: *Program Size Quartile*, *Number of Students Enrolled*, and *Awards Per Allocated Faculty Member* in the university ranking data.

From the above examples, we can see that even for variables that are highly correlated, their pairwise projections w.r.t a third variable present completely different data patterns. Therefore, they should not be grouped together in a feature subset (according to the *Correlation Feature Selection* criteria) to avoid information loss. In other words, correlation might not always be a good criterion for variable selection. In contrast, our approach compares two variables within the context of other variables. Moreover, our approach considers nine *Scagnostics* which convey many possible distributions of data points in a scatterplot [33]. This reduces information loss by focusing on a single visual feature, such as *Monotonicity*.

Let $Scag_s(i,k)$ be the visual measure s of scatterplot of variable i vs. variable k where s is one of the nine *Scagnostics*. For each pair of variables, we sum up the differences of their *Scagnostics* signatures w.r.t a third variable (and we repeat for all other variables in the data). The dissimilarity between two variables i and j is computed by the following equation:

$$Dissim(i,j) = \frac{\sum_{k=1,k\neq i,k\neq j}^{v} \sum_{s=1}^{S} |Scag_s(i,k) - Scag_s(j,k)|}{(v-2)\times S} \quad (1)$$

where k is the third variable, v is the number of variables in the dataset, and S is the number of visual features ($S = 9$). Equation 1 excludes the scagnostic values of variables i and j ($k \neq i, k \neq j$) since when $k = i$, $Scag_s(i,i)$ is not valuable to compare to $Scag_s(j,i)$.

Equation 1 allows for interpreting scatterplots from the attribute point of view. In particular, we propose a Scagnostics-based measure to assess the similarity between attributes. Two variables are considered similar if their pairwise projections w.r.t other variables in the dataset present similar distributions (or contain redundant information) and therefore should be grouped together.

One might argues that the choice of averaging differences of *Scagnostics* measures over all other variables does not sound very appealing and destroys the information on where and how two variables are dissimilar, due to Manhattan distance used in Equation 1. We believe the reserve and clarify this in the example in Figure 6. Consider four variables *v1*, *v2*, *v3*, and *v4* where (*v1*,*v4*) pair has a *Scagnostics* signature of [0, 0, 0, 0.5, 0.5, 0.5, 1, 1, 1], (*v2*,*v4*) pair has a *Scagnostics* signature of [0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5], and (*v2*,*v4*) pair has a *Scagnostics* signature of [1, 1, 1, 0.5, 0.5, 0.5, 0, 0, 0]. While (*v1*,*v4*) signature is vastly different from (*v3*,*v4*) signature (as shown the the two red boxes), they have the same dissimilarity scores when compared to (*v2*,*v4*) (the lost of information on individual *Scagnostics*). However, this does not lead to the conclusion that *v1* and *v3* are similar since based on Equation 1, $Dissim(v1, v3)$ is computed based on the comparisons of signatures of the two red plots and signatures of the two yellow plots.

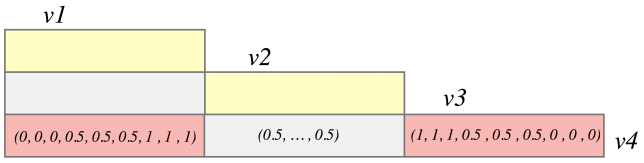

Figure 6: Averaging differences of *Scagnostics* measures: While (*v1*,*v4*) signature is different from (*v3*,*v4*) signature as shown the the two red scatteplots, they give the same dissimilarity score when compared to (*v2*,*v4*) signature.

3.2 *FSelector* Components

Figure 1 shows a schematic overview of *FSelector*:

1. **Processing:** Our approach computes nine *Scagnostics* measures of each pairwise projection of variables in the input data. Then variables are clustered based on their *Scagnostics* features vs. each of other variables. The dissimilarity between two variables is computed using Equation 1. At the end of the Processing stage, we have a list of representative variables and their followers (children in each group).

2. **Visualization:** The representative variables for each cluster are displayed in a summary SPLOM. The number of representative variables is smaller than the number of original variables.

3. **Interaction:** Users can select any variables in the summary SPLOM to inspect all pairwise projections within a cluster or any cells (scatterplots) in the summary SPLOM to compare the relationships between variables across different clusters.

The *FSelector* implements four analysis tasks:

- **T1**: Provide a summary view of multidimensional data [22]. *FSelector* clusters input variables based on their *Scagnostics* features (see Section 3.2.1) and then presents the clustering results using a SPLOM (see Section 3.2.2).

- **T2**: Select a representative variable or a representative scatterplot to expand all member variable in a cluster (see Section 3.2.2 and Section 3.2.3).

- **T3**: Sort variables based on their relevance to the representative variables (see Section 3.2.2).

- **T4**: Filter variables using individual *Scagnostics* or combinations of muiltiple measures [15] (see Section 3.2.3).

Our approach first groups the variables around leaders that are then visualized with a SPLOM.

3.2.1 Grouping Variables

Original variables from the input data are clustered together (visualization task **T1**) using the leader algorithm [19]. An initial *Scagnostics* threshold, called $sThreshold$ within the range from 0 to 1, is provided as an input of the algorithm [15]:

1. We initialize the leader list as an empty set ($L = \emptyset$).

2. For each variable V_i, we find the nearest leader (representative variable) in L which has the distance (in their *Scagnostics* space) to $V_i \leq sThreshold$.

   (a) If we could not find any leader variables satisfying this condition, we make V_i as a new leader and push V_i into the leader list L.

   (b) Otherwise, we insert V_i into the follower list of the nearest leader.

We might have used other popular clustering algorithms, such as K-Means or hierarchical clustering. There are several reasons we chose Hartigan's Leader. First, the number of iterations for step 2 of the above algorithm is linear $O(v)$ where v is the number of variables. In principle, this algorithm can handle larger datasets. In contrast, other well-known clustering algorithms require at least polynomial time complexity [19]. K-means may require exponentially many iterations even in the plane [2]. Vattani [30] provides a construction in two-dimensional space for which k-means requires an exponential number of iterations to stabilize. Second, this algorithm is not sensitive to the actual number of clusters in the data. Using this algorithm, we expect to get a smaller number of dimensions which represent high-dimensional data. We limit the number of leader variables (or $|L|$) from $log_2 v$ to $3 * log_2 v$. The main constraint (for using *log* functions) is about how large could be a SPLOM to allow users to read the content in each cell within the limited screen estate.

For a smaller dataset of 40 dimensions, we expect from 6 to 18 representative variables. For a larger dataset of 1,000 dimensions, we expect from 10 to 30 leader variables. This means that *FSelector* might need to adjust the *Scagnostics* threshold $sThreshold$ and repeat the above clustering algorithm a few times to get the expected number of leaders. In fact, we use *binary search* to quickly achieve the right $sThreshold$ for $log_2 v$ to $3 * log_2 v$ leading variables. We initialize $sThreshold = 0.5$.

3.2.2 Clustered Matrix

Orthogonal pairwise projections of leader variables are now presented in a SPLOM metaphor. Figure 7 shows an example for the NRC university rankings in Mathematics in 2006. Red plots are high *Monotonic* while blue plots are low *Monotonic*. As depicted in Figure 7(a), there are only a few variables which are highly correlated (highlighted in red). Therefore, grouping variables on their correlations will end up with many singleton clusters as depicted in Figure 7(b). In particular, variables are grouped if the Spearman correlation coefficients of their pairwise projections are at least 0.5 (it would not make much sense to group variables with correlation coefficients less than 0.5). With this setting, we obtain only two larger clusters (other are singleton clusters): the first cluster is leaded by *Program Size Quartile*) which has 7 member variables (including the leading variable) and the second cluster is leaded by *R rankings* which has 5 member variables as depicted in Box T of Figure 7(b).

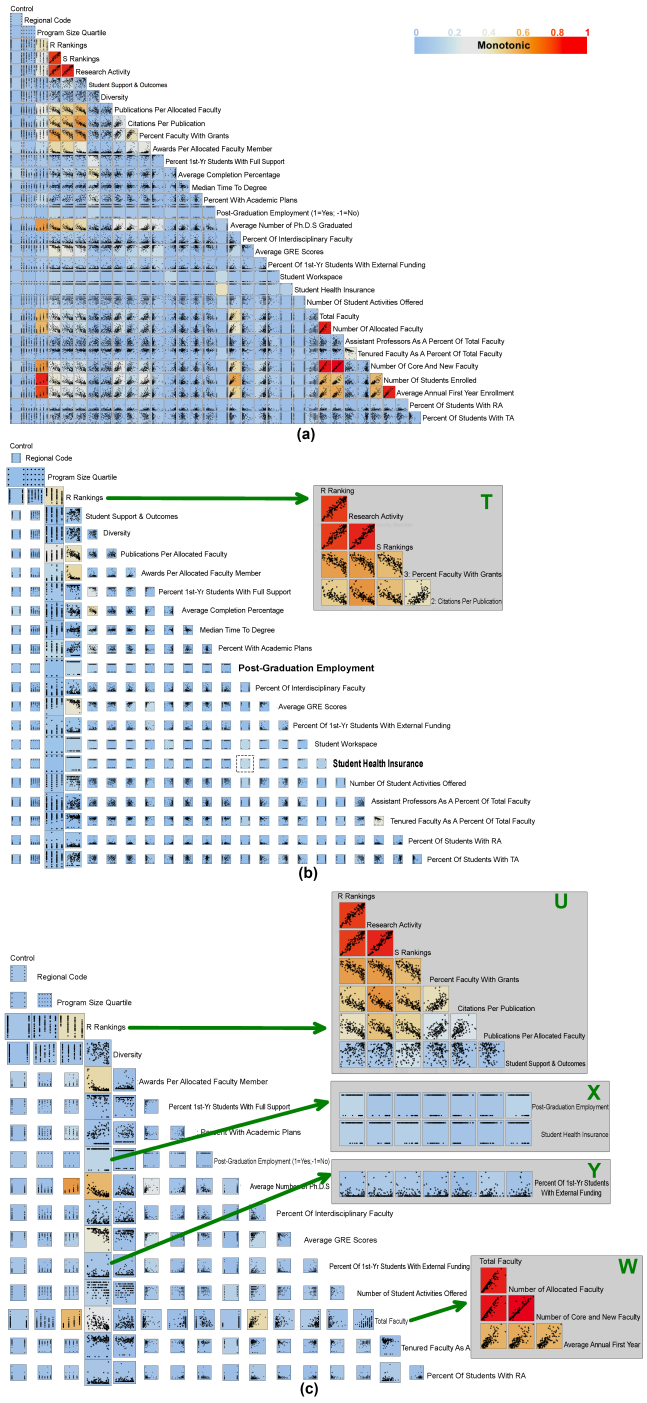

Figure 7: *FSelector* visualization for the NRC university rankings in 2006: (a) Scatterplot matrix of 33 variables in the input dataset. (b) The resulted SPLOM of clustering variables based on Spearman correlation coefficients (c) The resulted SPLOM of clustering variables based on the nine *Scagnostics*. By selecting a leader variable or a scatteplot, users can see members in its group as depicted in a popup window U, W, or X,Y.

Figure 7(c) shows *FSelector* visualization for 17 leading variables when setting *Scagnostics* threshold $sThreshold = 0.3$ by the user. The size of each scatterplot in this matrix is decided by its dominating variable (the leader which has more follower variables in the two variable making a given scatterplot). For example on a standard screen resolution of 1440 by 900 (width by height), the largest scatterplot (for this example) in our *FSelector* SPLOM is about 52 pixels (or 900/17 for both plot height and width) while singleton clusters have sizes of 26 pixels (or half of the max plot size). Other clustered plot sizes are linearly scaled within this interval based on the number of member variables in their dominating cluster.

Box U depicts selecting a leading variable (visualization task **T2**), showing the details of this cluster on demand [28]. Variables in the selected group have been ordered by the *Monotonic* measure (visualization task **T3**) w.r.t the leader variable (*R rankings*) and displayed in a sub-matrix. As shown in Box U, *Research Activity* and *S rankings* have a high covariance with *R rankings* while *Percent Faculty with Grants*, *Citation per Publication*, and *Publication per Allocated Faculty* are not strongly correlated with *R rankings*. Notably, *Student Support & Outcomes* is uncorrelated to *R rankings*, but is still a member of this group (leaded by *R rankings*). Comparing Box T (produced by variable correlation) and Box U (produced by the *Scagnostics*-based measure), we can see that *FSelector* is able to group additional uncorrelated variables since it relies on their data patterns w.r.t to remaining variables in the input dataset (as shown in Equation 1).

We further explain this unique feature of *FSelector* by inspecting the pairwise projections of variables in Box U vs. other variables in Box X (2 variables) and Box Y (1 variable). As depicted in Box X, all 7 variables in Box U have similar distributions (w.r.t to *Post-Graduation Employment* and *Student Health Insurance* , which are also grouped in the same cluster). Hence, we reveal similar *Scagnostics* features. In brief, 14 scatterplots in box X are represented as one plot (at the origin of the green arrow to Box X) in the main SPLOM. In fact, *Post-Graduation Employment* and *Student Health Insurance* are uncorrelated but *FSelector* groups them based on the fact that they have almost identical projections w.r.t. other variables. Otherwise, by simply analyzing their pairwise projections, the variable correlation method in Figure 7(b) fails to capture their similar behaviors to other variables in the data.

Box Y shows 7 pairwise projections of variables in Box U vs. a singleton cluster (*Percent Of 1st-Yr Students with External Funding*). The similarity of these scatterplots supports clustering variables in Box U. Moreover, the dissimilarity of these scatterplots compared to those in Box X implicates their different *Scagnostics* measures. Therefore, variables in Box X and Y are separated into different clusters [18]. Box W presents another cluster of four variables where *Total Faculty* is the leading variable.

3.2.3 Interactions

Instead of showing all pairwise projections in a popup window, users can also examine the member scatterplots one-by-one via mouse clicks. An example cluster is depicted at the bottom right of Figure 1 for the US monthly employment in *Trade and Transportation*. In particular, 12 variables in the clustered SPLOM represent for 53 variables (states) in the data. The bottom right window shows the projection of *Georgia* and *Indiana* in the cluster led by *Colorado* (at the origin of the green arrow). Small thumbnails on the top left corner of the window allow users to navigate though different member scatterplots in smooth transitions. Current selected variables are displayed in black (while others are in green). Green trajectories keep track of how every data point moves between different member scatterplots in this group. As depicted in a close-up view in Figure 8, the green trajectories provide a summary picture of how data values extend in both dimensions. A smooth transition via these trajectories is triggered as users select any member scatterplots from the top left thumbnails. The trajectories become less useful as the number of member variables increases: Too many trajectory-crossings can occur.

Besides brushing and linking, the visual interface also supports

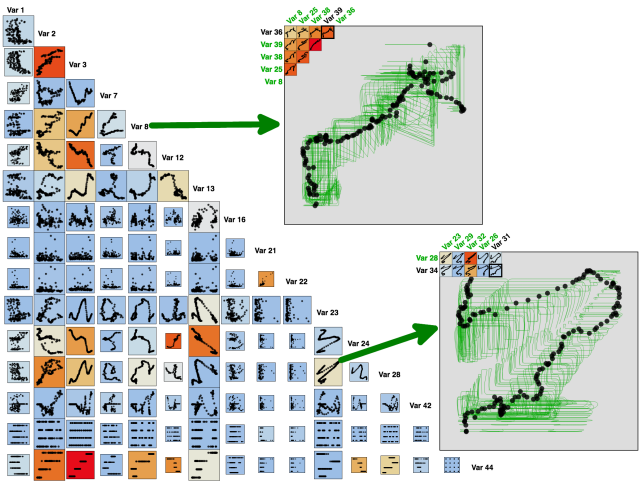

Figure 8: Our strategy to examine the member scatterplots on demand: Green trajectories track changes of data points along different pairwise projections in a cluster. Small thumbnails on the top left corner of the popup window animate data point movements on their trajectories when clicked.

other interactions, such as ordering, and filtering (visualization task **T4**). Figure 5 shows a filtering example for the NRC University Ranking data. In particular, we filter variables which are highly correlated (*Program Size Quartile* and *Number of Students Enrolled*) but their 2D projections w.r.t a third variable (*Awards Per Allocated Faculty Member*) have completely different *Scagnostics* signatures. The filtering process is done automatically and the returned results are ordered based on the filtering conditions.

4 Validation

4.1 Datasets

We will illustrate the features of *FSelector* mainly through examples. We use datasets retrieved from the UCI Repository [3] and other sources to demonstrate the performance of *FSelector*. Table 1 summarizes prominent aspects of these datasets ordered by the number of dimensions.

Table 1: Characteristics of datasets used for demonstrations and testings in the following sections. The datasets are listed by increasing order of the number of variables v. Datasets in gray background are demonstrated in Section 4.2. Datasets in yellow background are demonstrated in Section 4.3.

Datasets	v	n	# scatterplots
Breast Cancer	32	569	496
NRC University Ranking	33	127	528
Trade and Transportation	53	324	1,378
Total nonfarm	53	324	1,378
Leisure and Hospitality	53	324	1,378
Government	53	324	1,378
NYC Subway Ridership	104	423	5,356

The National Research Council (NRC) ranking data comprise university rankings in Mathematics in 2006 according to different criteria (attibutes in the data). There are 33 variables represented in the dataset: *R-Rankings*, *S-Rankings*, ranking factors and information on 127 universities in the US. For *S-Rankings*, programs are ranked highly if they are strong in the criteria that scholars say are most important. For *R-Rankings*, programs are ranked highly if they have similar features to programs viewed by faculty as top-notch. Overall, we have totally 528 scatterplots with 127 data points (associated to 127 universities) in each scatterplot to compute *Scagnostic* measures.

The US employment datasets (highlighted in yellow in Table 1) [8] contain monthly employment of 53 states over 27 years from 1990 to 2016 across different economy sectors: *Trade and Transportation*, *Total nonfarm*, *Leisure and Hospitality*, and *Government*. The states are considered as attributes in high dimensional data. We demonstrate *FSelector* on these datasets in Figure 1 and Figure 10. Breast Cancer and NYC Subway Ridership datasets are retrieved from the UCI Repository [3].

4.2 Comparisons to Alternative Interfaces

Equation 1 allows interpreting variables from the attribute point of view. Moreover, the grid layout proposed in this paper allows us to focus on interesting variables, to compare them to sets of following variables, and to highlight pairs of leading variables (by size) instead of prototypical scatterplots as in *ScagExplorer* [15]. In other words in *ScagExplorer*, scatterplots are considered as individuals represented within a MDS; the focus is on visualizing similarity between prototypical scatterplots. In contrast, *FSelector* uses a SPLOM to put the focus on visualizing the relation between prototypical variables more directly.

Figure 9 shows comparisons of *ScagExplorer* and *FSelector* on two different datasets from the UCI Repository [3]: (a) Breast Cancer (32 variables) (b) New York City subway ridership (104 variables). These datasets are highlighted in darker backgrounds in Table 1. In particular, the *ScagExplorer* visualization (the left panels in each row of Figure 9) tries to convey typical pairwise projections to provide an overview of variable relations in the input data. The background colors of scatterplots encode their variable correlation (red for high *Monotonic*, blue for low *Monotonic*). Larger scatterplots represent more popular data patterns (more similar scatterplots in these clusters). *ScagExplorer* employs a free layout that disrespects the grid alignment of the original variables. *FSelector* overcomes this limitation by organizing leading variables in a grid layout where the size of each pairwise projection is computed based on sizes of the two corresponding groups of variables. By browsing by row/column, one can quickly identify important variables in clustered SPLOMs (the right panels in each row of Figure 9).

The right panels of Figure 9 also depict brushing a variable/scatterplot in *FSelector* visualizations. Particularly when users select a variable or a scatterplot in the main SPLOM, all pairwise projections in its group are displayed in a secondary matrix (in a popup window at the targets of green arrows). Figure 9(a) depicts selecting a leading variable (origins of green arrows: *Var 3*). The selection details are presented in half-square matrices to highlight the relationships between member variables within the cluster. As depicted, member variables in the selected cluster of the Breast Cancer dataset are strongly correlated while that might not be the case with other clusters/datasets.

Figure 9(b) depicts selecting a scatterplot (origins of green arrows). In this case, a full grid layout shows all possible 2D projections of two clusters (row and column) in the main SPLOM. This allows analysts to compare pairwise projections of variables across clusters. The first variables (of each row and column) in secondary grid layout are the leading variables from the row and column clusters of selected scatterplot (*Var 17* vs. *Var 92*). These leading variables are highlighted in bold. The consistent patterns of these pairwise projections explain why they should be grouped together. This is true from both row and column perspectives. Variables on rows or columns in the secondary layouts are considered redundant and therefore their details should be hidden in the main SPLOM (the main SPLOM contains only the representative variables). In other words, all scatterplots in this secondary grid layout are represented as a single cell (the origin of each green arrows) in the clustered SPLOM.

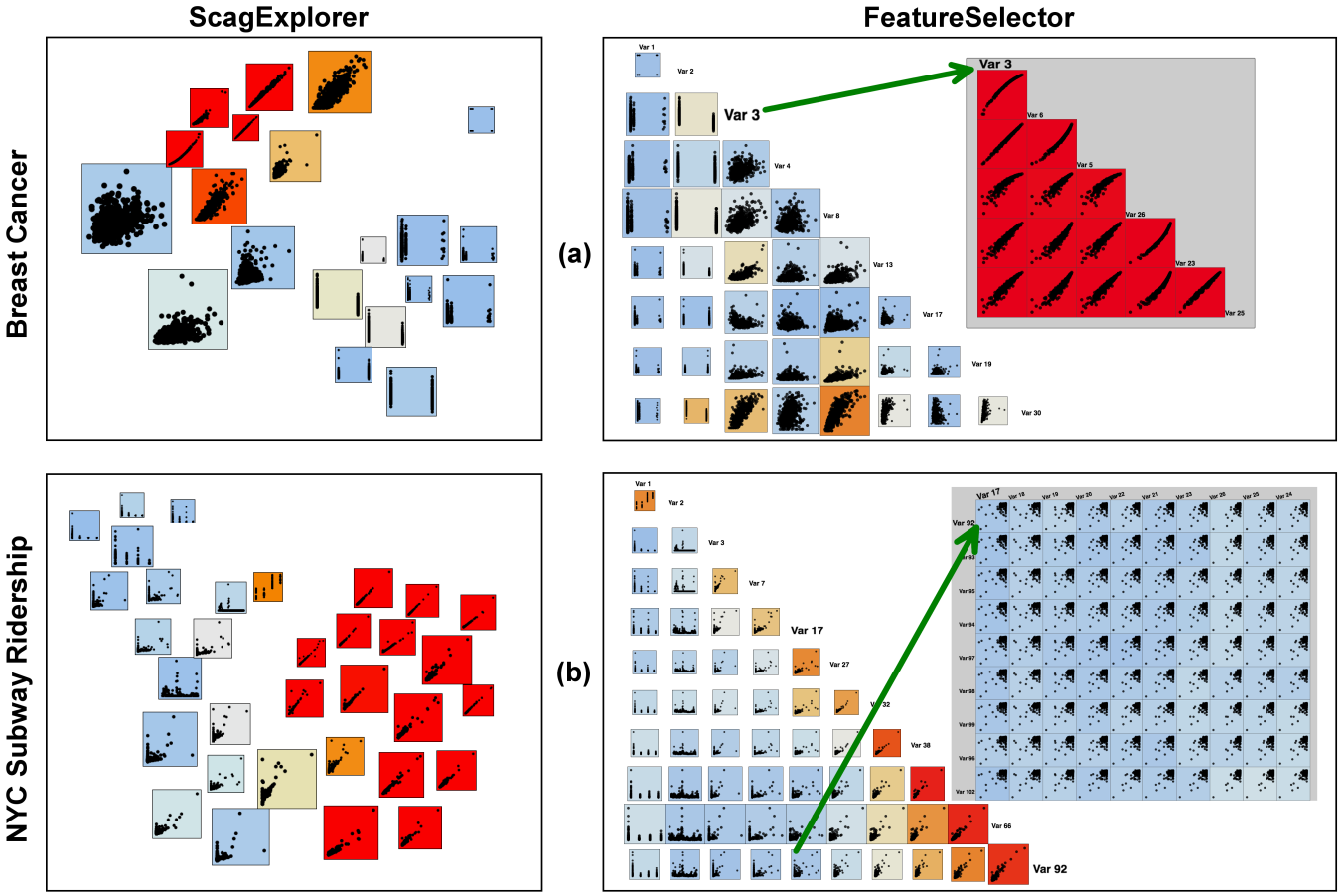

Figure 9: Comparison of *ScagExplorer* vs. *FSelector* on two datasets: (a) Breast Cancer (b) New York City Subway Ridership.

Scatterplots in the secondary (or "details on demand") views in *ScagExplorer* may come everywhere in the original SPLOM as depicted in Figure 8 of the *ScagExplorer* paper [15]. In contrast, scatterplots in the *FSelector* secondary views come from rows and columns in the input SPLOM (the first SPLOM in Figure 1). Consequently, *FSelector* allows users to discern the relation between prototypical variables and so put the focus on analyzing the original variables. This feature is beneficial in many application domains, such as chemistry or biology where working directly on original variables may be more valuable.

4.3 Use Case

In this section, we inspect *FSelector* using the US employment data from 1990 to 2016 retrieved from the BLS website (yellow datasets in Table 1). In particular, we consider different states as dimensions $v = 53$ and monthly employments as observations $n = 324$. Inspected economy sectors include *Total nonfarm*, *Leisure and Hospitality*, and *Government* as depicted in Figure 10(a), (b), and (c) respectively. States with similar employment 2D patterns (on their monthly employment) w.r.t. other states are grouped and encoded in the same color. There are about 6 or 7 non-singleton clusters automatically determined (using the leader algorithm in Section 3.2.2) in each example. *FSelector* displays a map on the right of summary SPLOM to add an additional layer of information specific to this data type. Blue is the biggest cluster while light gray is the states which do not share employment patterns with any other states on the given economic sector. Distinct regional patterns can be observed in these examples. Outliers can also be easily discerned, such as Nevada in Figure 10(b). Concretely, Nevada's monthly employment in *Leisure and Hospitality* over the past 27 years is not similar to any other states in the US except Hawaii. One can discover this after expanding Nevada's cluster. This is due to the fact that both Nevada and Hawaii have similar seasonal employment patterns (in *Leisure and Hospitality*) which are distinct from others.

In Figure 10(c), it is interesting that California, Oregon, and New Mexico are grouped together based on their employment patterns in the Government sector, which are independent of the fact that they consistently voted for Democrats since 1992. This is also true for many other states voting for Republicans. In Figure 10(d), we expand a cell: The similarity of these (high *Clumpy*) 2D projections helps to explain why Alabama, Arkansas, and other states in x axis are clustered together and why Louisiana is separated from other groups. Notice that these *Clumpy* patterns cannot be discerned in any orthogonal projections.

FSelector is implemented in javascript using the D3.js library [7]. The online prototype, demo video, and source code are available on our Github repository `https://featureselector.github.io/`.

5 Conclusion

This paper proposes a criterion for unsupervised feature selection for data analysis. Common methods for feature selection are based on the (pairwise) correlation of numeric features and identify a set of mutually uncorrelated features. The proposed new criterion is based on two ideas. First, the correlation between a pair of features is measured not directly between the pair, but between the correlations when each feature of the pair is combined with any other feature

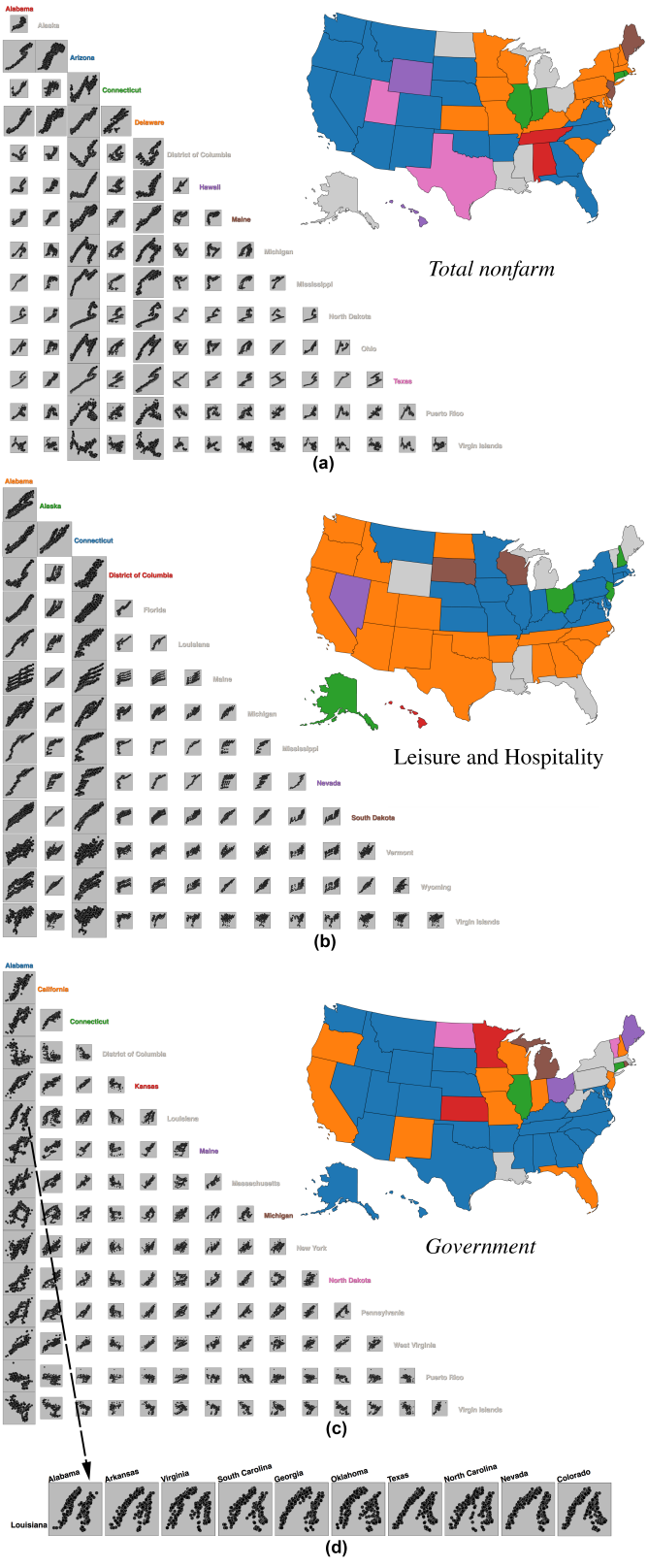

Figure 10: The clustered SPLOMs of *FSelector* for different BLS datasets: (a) *Total nonfarm* (b) *Leisure and Hospitality* (c) *Government*. An expansion of a cell in the clustered SPLOM (c) is presented in (d).

of the data set. And second, the correlation is based not on one of the standard statistical correlation measures, but on the difference between the nine Scagnostics measures observed for each pairing of variables in a scatterplot.

In the future work, we would like to conduct formal evaluations of the two key ingredients of Equation 1 to better communicate why effectiveness of our approach: (1) why we compare all features with all other variables instead of the traditional approach that compares variables pairwise (2) why all nine *Scagnostics* are necessary to include in the clustering method. A systematical comparison isolating the role that each *Scagnostics* plays in the resulting clusters and in-depth analysis on the benefit of comparing all features with all other variables are valuable in this regard. Due to (1), computing *Scagnostics* is quadratically dependent on the number of variables v since we have $v*(v-1)/2$ scatterplots to compute these visual features. Parallel computing helps to make to technique scalable to the large v which needs further investigation in future research.

REFERENCES

[1] G. Albuquerque, M. Eisemann, and M. Magnor. Perception-based visual quality measures. In *IEEE VAST*, pp. 13–20, 2011.

[2] D. Arthur and S. Vassilvitskii. How slow is the k-means method? In *Proceedings of the Twenty-second Annual Symposium on Computational Geometry*, SCG '06, pp. 144–153. ACM, New York, NY, USA, 2006. doi: 10.1145/1137856.1137880

[3] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[4] M. Behrisch, B. Bach, M. Hund, M. Delz, L. V. Rden, J. D. Fekete, and T. Schreck. Magnostics: Image-based search of interesting matrix views for guided network exploration. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):31–40, Jan 2017.

[5] M. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. Wright, J. Wilson, F. Agakov, P. Navarro, and C. Haley. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports*, 5:10312, May 2015. doi: 10.1038/srep10312

[6] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, Dec. 2011. doi: 10.1109/TVCG.2011.229

[7] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–2309, 2011.

[8] Bureau of Labor Statistics. http://www.bls.gov/data/, March 2017.

[9] A. Butte. The use and analysis of microarray data. *Nature reviews. Drug discovery*, 1(12):951960, December 2002. doi: 10.1038/nrd961

[10] R. Caruana and V. R. de Sa. Benefitting from the variables that variable selection discards. *J. Mach. Learn. Res.*, 3:1245–1264, Mar. 2003.

[11] R. da Silva, P. E. Rauber, R. M. Martins, R. Minghim, and A. C. Telea. Attribute-based visual explanation of multidimensional projections. *Proc. EuroVA*, pp. 134–139, 2015.

[12] T. Dang and L. Wilkinson. TimeExplorer: Similarity search time series by their signatures. In *Proc. International Symp. on Visual Computing*, pp. 280–289, 2013.

[13] T. N. Dang, A. Anand, and L. Wilkinson. Timeseer: Scagnostics for high-dimensional time series. *Visualization and Computer Graphics, IEEE Transactions on*, 19(3):470–483, 2013. doi: 10.1109/TVCG.2012.128

[14] T. N. Dang, H. Cui, and A. G. Forbes. MultiLayerMatrix: Visualizing Large Taxonomic Datasets. In N. Andrienko and M. Sedlmair, eds., *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2016. doi: 10.2312/eurova.20161125

[15] T. N. Dang and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *2014 IEEE Pacific Visualization Symposium*, pp. 73–80, March 2014. doi: 10.1109/PacificVis.2014.42

[16] G. Forman. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305, Mar. 2003.

[17] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, Mar. 2003.

[18] M. A. Hall. Correlation-based feature selection for machine learning. Technical report, 1999.

[19] J. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.

[20] P. Joia, F. Petronetto, and L. G. Nonato. Uncovering representative groups in multidimensional projections. In *Computer Graphics Forum*, vol. 34, pp. 281–290. Wiley Online Library, 2015.

[21] E. Kandogan. Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pp. 73–82. IEEE, 2012.

[22] D. A. Keim, C. Panse, and M. Sips. Information visualization: Scope, techniques and opportunities for geovisualization. In J. Dykes, ed., *Exploring Geovisualization*, pp. 1–17. Elsevier, Oxford, 2004.

[23] J. Krause, A. Dasgupta, J. D. Fekete, and E. Bertini. Seekaview: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. In *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)*, pp. 11–19, Oct 2016. doi: 10.1109/LDAV.2016.7874305

[24] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. *IEEE transactions on visualization and computer graphics*, 23(3):1249–1268, 2017.

[25] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pp. 89–96. IEEE Computer Society, Washington, DC, USA, 2004. doi: 10.1109/INFOVIS.2004.15

[26] M. Sedlmair and M. Aupetit. Data-driven Evaluation of Visual Quality Measures. *Computer Graphics Forum*, 2015. doi: 10.1111/cgf.12632

[27] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, July 2005. doi: 10.1057/palgrave.ivs.9500091

[28] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pp. 336–. IEEE Computer Society, Washington, DC, USA, 1996.

[29] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser. Representative factor generation for the interactive visual analysis of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2621–2630, 2012.

[30] A. Vattani. K-means requires exponentially many iterations even in the plane. In *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry*, SCG '09, pp. 324–332. ACM, New York, NY, USA, 2009. doi: 10.1145/1542362.1542419

[31] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping. Use of the zero norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461, Mar. 2003.

[32] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings of the IEEE Information Visualization 2005*, pp. 157–164. IEEE Computer Society Press, 2005.

[33] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372, 2006.

[34] J. Yang, A. Patro, S. Huang, N. Mehta, M. O. Ward, and E. A. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pp. 73–80. IEEE Computer Society, Washington, DC, USA, 2004. doi: 10.1109/INFOVIS.2004.71

[35] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. pp. 856–863, 2003.

[36] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller. Visual correlation analysis of numerical and categorical data on the correlation map. *IEEE transactions on visualization and computer graphics*, 21(2):289–303, 2015.

[37] X. Zhao and A. Kaufman. Structure revealing techniques based on parallel coordinates plot. *Vis. Comput.*, 28(6-8):541–551, June 2012. doi: 10.1007/s00371-012-0713-0