Interactive Dimensionality Reduction Through User-defined Combinations of Quality Metrics

Sara Johansson and Jimmy Johansson

Abstract—Multivariate data sets including hundreds of variables are increasingly common in many application areas. Most multivariate visualization techniques are unable to display such data effectively, and a common approach is to employ dimensionality reduction prior to visualization. Most existing dimensionality reduction systems focus on preserving one or a few significant structures in data. For many analysis tasks, however, several types of structures can be of high significance and the importance of a certain structure compared to the importance of another is often task-dependent. This paper introduces a system for dimensionality reduction by combining user-defined quality metrics using weight functions to preserve as many important structures as possible. The system aims at effective visualization and exploration of structures within large multivariate data sets and provides enhancement of diverse structures by supplying a range of automatic variable orderings. Furthermore it enables a quality-guided reduction of variables through an interactive display facilitating investigation of trade-offs between loss of structure and the number of variables to keep. The generality and interactivity of the system is demonstrated through a case scenario.

Index Terms—Dimensionality reduction, interactivity, quality metrics, variable ordering.

---

1 INTRODUCTION

Multivariate data sets with hundreds of variables are increasingly common in several application areas, such as surveys, product development and simulations. One example scenario is the analysis of questionnaire data where hundreds of questions result in equally many variables. For this data traditional visualization techniques for multivariate data, including parallel coordinates [9, 22], scatter plot matrix [5] and table lens [19], are usually inefficient.

A common way of addressing the difficulties involved in visualization of data with a large number of variables is to employ dimensionality reduction prior to visualization. Numerous dimensionality reduction methods are available, preserving different structures within the data set. The importance of a certain structure, such as similarity, correlation or variance, is highly dependent on the task of the data analysis. For exploratory analysis of data, as well as for many other analysis tasks, several structures or quality metrics may be of high importance.

This paper introduces a dimensionality reduction system for exploratory analysis of multivariate data sets with large numbers of variables, providing interactive analysis of the whole data set based on quality metrics selected by the user. The importance of different metrics is controlled by user-defined weight functions and the importance of the variables is based on the combination of selected quality metrics and their weight functions.

By definition all dimensionality reduction methods result in some loss of information since data is removed from the data set. Hence it is of great importance to reduce the data set in a way that preserves the important structures within the original data set. The amount of information lost in a dimensionality reduction, here referring to the loss of structure due to removal of variables, depends not only on the number of variables removed, but also on the structures within both retained and removed variables. The system presented in this paper provides a visual representation enabling investigation of trade-offs between the number of variables removed and the loss of information. This guides the selection of the number of variables to keep, and provides an understanding of the information loss.

The order of variables in displays has a large impact on the effectiveness of the visualization and on our ability to perceive structures in the data [3]. Finding one appropriate variable ordering enhancing several interesting structures at once may be unrealistic, hence the proposed system offers a selection of automatic orderings, each enhancing the structures of an individual quality metric. These orderings are performed subsequent to the dimensionality reduction and have no influence on the amount of information that is lost.

To summarize, the contribution of this paper is a dimensionality reduction system that enables:

- User defined combination of quality metrics for dimensionality reduction, using weight functions.

- Automatic ordering of variables to enhance perception of patterns selected by the user.

- Interactive quality guided reduction of variables and selection of the number of variables to keep.

The remainder of the paper is organized as follows. Section 2 presents the background and related work. In section 3 the features of the dimensionality reduction system are described in detail. Section 4 describes the performance of the system and presents a case scenario and in section 5 conclusions and future work are presented.

2 BACKGROUND AND RELATED WORK

Visualization of multivariate data sets with a large number of variables is a major challenge in information visualization. This section presents previous research related to visualization of this kind of data.

2.1 Visual Representations

Data sets with hundreds or even thousands of variables are nothing unusual, but few visualization techniques are able to simultaneously represent all variables of such data sets effectively. Commonly used visualization techniques for multivariate data, such as parallel coordinates [9, 22], scatter plot matrix [5] and table lens [19], are unable to effectively represent much more than twenty variables simultaneously.

However, some visualization techniques able to visually represent a large number of variables exist. An example is pixel displays [12] using diverse pixel layout schemes for different analysis tasks. These techniques represent each data value with one pixel, and although they are able to represent large data sets, they are still limited by screen space. Furthermore, the effectiveness of visualization techniques is

- *Sara Johansson is with Norrköping Visualization and Interaction Studio (NVIS), Linköping University, Sweden, E-mail: sara.johansson@itn.liu.se.*
- *Jimmy Johansson is with Norrköping Visualization and Interaction Studio (NVIS), Linköping University, Sweden, E-mail: jimmy.johansson@itn.liu.se.*

not only limited by the size of the monitor, but also by human visual capabilities, hardware and interactivity [7]. The Value and Relation (VaR) display [25] shows a large number of variables by representing each variable in a two-dimensional space using glyphs arranged to express relationships among variables and applying a pixel-oriented approach to represent data items within the glyphs. The major limitation of VaR is that highly correlated glyphs overlap, making it difficult to identify multivariate patterns such as clusters, although interaction partly solves this problem. In [17] a hierarchical point placement technique (HiPP) is presented. HiPP aims at providing exploration of large data sets in a two-dimensional display by using multiple levels of detail through hierarchical clustering.

2.2 Dimensionality Reduction

A common approach to visualization of data sets with large numbers of variables is to precede the visualization with some dimensionality reduction, visually representing the data set in a lower dimensional space preserving some of the structures of the original data set. Dimensionality reduction has the benefit of not having any limitation in number of variables in the original data set, although a higher number of variables often results in increasing computation time.

Some of the more popular dimensionality reduction methods are Principal Components Analysis (PCA) [11], which preserves variance by extracting a number of principal components, Multidimensional Scaling (MDS) [15], preserving dissimilarities, and Self Organizing Maps (SOM) [13], preserving topological and metric relationships. In [14] a family of dimensionality reduction techniques are presented, which simultaneously take several structures within the data into consideration. These methods are able to transform data sets of hundreds of variables onto low dimensional representations, where the variables often are linear combinations of the original variables. However, the user has little or no influence on the result of the dimensionality reduction. In [24] an interactive system is presented where the user can guide the MDS process during computation and select local regions of interest to focus the computational effort on.

A common drawback of the techniques described above is that the relationship between the original and reduced sets of variables is usually not intuitive. Approaches presenting more intuitive relationships are to group similar variables, using one representative variable per group, or to select a representative subset of variables, using quality metrics. An example of variable grouping is the Principal Component Variable Grouping (PCVG) presented in [10]. PCVG is based on PCA but uses the principal components to group the original variables.

Several interactive systems for dimensionality reduction exist, focusing on different structures within the data. In [27] the Visual Hierarchical Dimension Reduction (VHDR) system is presented. VHDR creates a variable hierarchy based on similarity between variable pairs and provides possibilities to modify the hierarchy and to select interesting variables. The Dimension Ordering Spacing and Filtering Approach (DOSFA) [26] has evolved from VHDR and provides dimensionality reduction based on a combination of similarity and of a given importance measure, such as variance. DOSFA is also aiming at finding a visual layout that facilitate the detection of structures within the data. A similar approach is presented in [4] where a system for reduction and ordering of variables based on variable pair similarity is described. Further approaches to variable ordering are presented in [18], where clutter reduction is achieved through variable ordering using individual clutter measures for different visual representations. In [21] the rank-by-feature framework focusing on interactive detection of interesting structures for pairs of variables is presented. Ranking criteria are selected from a list of available criteria, and interesting pairs can be interactively selected for further analysis. A similar approach is presented in [8] where pairs of variables are ranked based on a 'goodness of clustering' measure. These systems combine algorithmic analysis of data with interactivity to find a dimensionality reduction suitable for the users need and for the task of the analysis.

2.3 Summary and Contributions

As presented in this section, a large number of dimensionality reduction techniques and systems exist. For several, the relationship between the original and reduced data set is not intuitive and the user has little or no control over the reduction. Moreover, many techniques are limited either to focusing on pairs of variables or to only taking a single quality metric into consideration. For many analysis tasks several structures are of high significance and for some patterns, such as clusters, multivariate structures are of more importance. DOSFA is a bit more flexible, but does not provide any possibility to adjust the impact of the two measures used.

The contribution of this paper is a dimensionality reduction system that combines user-defined and weighted quality metrics and allows interactive control of parameters affecting the reduction. Furthermore, the system enables investigation of the trade-off between the number of variables and loss of information, guiding the selection of the number of variables to keep, which is not supported by previous methods.

3 SYSTEM

The system presented in this paper is an interactive dimensionality reduction system where algorithmic analysis is combined with user influence and control over the analysis process. The structure of the system provides interactive modifications of the result and thus enables efficient exploration of multivariate data sets containing a large number of variables. The work flow of the system is presented in figure 1 and is as follows:

1. A data set is loaded into the system and the user selects quality metrics to use and sets the parameters for the quality metric analysis.

2. The system performs quality analysis for the selected metrics individually and determines a quality value for each variable and metric.

3. The relationship between the number of variables to keep and loss of information is presented to the user in an interactive display. At this point the user can also modify the importance of individual quality metrics, updating the display accordingly.

4. The user decides on the number of variables to keep in the reduced data set and the system selects the most important variables from the original data set based on quality values and metric importance.

5. In the final step before the reduced data set is displayed, the user selects which visual representations to use and which quality metric the variable ordering should enhance.

6. The reduced data set is displayed using the selected representations and orderings. From here any of the previous steps can be repeated to modify the reduced data set.

Due to the diversity of possible analysis tasks succeeding a dimensionality reduction, a range of selectable visual representations for multivariate data is available within the system. Providing an adaptable environment for visual and exploratory analysis based on user preferences and the analysis task. The steps of the system work flow will be described in more detail in the following sections. For illustration a synthetic data set containing 1320 data items and 100 variables, of which 14 contain significant structures, is used. Further details on the data set can be found in section 4.

3.1 Quality Metrics

The importance of a data structure or quality metric differs between different analysis tasks. Accordingly, a general system for analysis and determination of the quality of a data set and of the variables within a data set cannot be limited to a few quality metrics of equal importance. The system presented in this paper enables user-defined
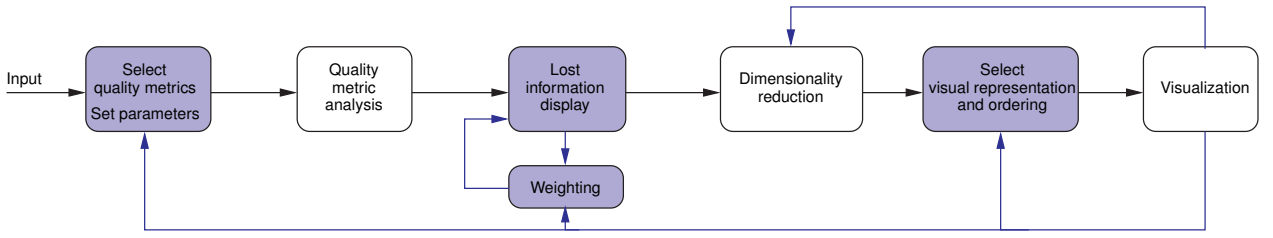
Fig. 1. The work flow of the dimensionality reduction system, blocks and arrows in blue represent sections initialized and controlled by the user. The user decides on which quality metrics to use, data analysis is performed and the relationship between information loss and number of variables to keep is presented, also providing a possibility of quality metric weighting. The user decides on number of variables to keep, dimensionality reduction is performed and the reduced data set is displayed using representations selected by the user.

and weighted combinations of quality metrics to determine the most important variables in data sets with large numbers of variables.

Any data analysis method where a quality value can be extracted for individual variables can be used as a quality metric within the system. In this paper the focus will be on three specific data structures: correlation, outliers and clusters. These are chosen since they are common structures with clear differences. Suggestions of other possible metrics are variance, different distribution metrics and many of the metrics based on convex and non-convex hulls, presented in [23].

Throughout this section the following notation is used: a data set X, includes M variables and N items, $\vec{x}_i$ is an item where $i = 1,...,N$, $\vec{x}_j$ is a variable where $j = 1,...,M$ and $x_{i,j}$ is the data value for item i in variable j. The normalized version of a value a_b is represented by $\tilde{a}_b$.

3.1.1 Correlation Analysis

As correlation quality metric the Pearson correlation coefficient [20], r, is computed for each pair of variables within the data set. Equation 1 shows the computation of r for variables $\vec{x}_j$ and $\vec{x}_k$ where $k = 1,...,M$.

$$r(\vec{x}_j, \vec{x}_k) = \frac{N \sum_{i=1}^{N} x_{i,j} x_{i,k} - \sum_{i=1}^{N} x_{i,j} \sum_{i=1}^{N} x_{i,k}}{\left(N \sum_{i=1}^{N} x_{i,j}^2 - \left(\sum_{i=1}^{N} x_{i,j}\right)^2\right) \left(N \sum_{i=1}^{N} x_{i,k}^2 - \left(\sum_{i=1}^{N} x_{i,k}\right)^2\right)} \tag{1}$$

Using the correlation values for pairs of variables, individual correlation quality values are computed for all variables of the data set. This value is defined as $I_{\text{corr}}(\vec{x}_j) = \sum_{k=1}^{M} |r(\vec{x}_j, \vec{x}_k)|$ for $k \neq j$ and $|r(\vec{x}_j, \vec{x}_k)| > \varepsilon$ where ε is a user defined value in the range 0.05 to 0.5. The absolute value is used since only the magnitude of the correlation is of interest and small correlation values are ignored to avoid large numbers of insignificant correlations adding up to a total which appears to be significant.

3.1.2 Outlier Detection

For efficient detection of multivariate outliers a density and grid based approach is taken where outliers for every pair of variables are detected initially. For an item, $\vec{x}_i$, to be defined as an outlier in a 2D space the number of neighbour items, ρ, within a given radius, ϕ, around $\vec{x}_i$ should not exceed a threshold, ζ. Given ϕ the items of variables $\vec{x}_a$ and $\vec{x}_b$ are binned into a 2D grid where the length and width of the cells equal $\sqrt{\frac{\phi^2}{2}}$. Hence all items within a cell are within the distance ϕ from each other. Items in grid cells containing less than ζ items are considered candidate outliers. The Euclidean distance between candidate outliers and items of surrounding grid cells within the distance ϕ is calculated, and items with less than ζ neighbouring items are defined as two-dimensional outliers.

Higher order outliers are identified as items which are outliers for a set of variable pairs where the number of unique variables exceeds a threshold, η, for instance the variable pairs $\vec{x}_a$, $\vec{x}_b$ and $\vec{x}_a$, $\vec{x}_c$ contain three unique variables, $\vec{x}_a$, $\vec{x}_b$ and $\vec{x}_c$. For each higher order outlier item a quality value, $o_i = \sum_{j=1}^{K} \frac{1}{\rho+1}$, is computed, where K is the set of variable pairs for which the item is a two-dimensional outlier. Hence, outliers of high dimensionality and outliers with a small number of

neighbour items are given high quality values. When outlier items are identified, importance values for individual variables, $I_{\text{out}}(\vec{x}_j)$, is computed. This is done by summing o_i for all variables belonging to that variable where $o_i > \varsigma$, where ς is a small user defined value, typically 1, used to avoid large numbers of insignificant outliers adding up to what appears to be a significant outlier value.

Three default settings for ϕ and ζ are available within the system, defined as 'high', 'medium' and 'low' constraints on outliers. The default values are identified based on the average minimum distance within a randomly distributed data set and can be modified by the user.

3.1.3 Cluster Detection

The proposed system uses the Mafia clustering algorithm [16] to identify low dimensional sub-clusters, which are then the base of computing a cluster quality value for every variable. However, any clustering algorithm able to identify subspace clusters could be used for this. Mafia is a density based clustering algorithm for identification of subspace clusters that has evolved from the Clique clustering algorithm [1], which combines density and grid based clustering with Apriori reasoning [2]. Using this approach, a cluster can be defined as a region with higher density of data items than the surrounding regions. The cluster algorithm uses the definition of a unit, which is a rectangular cell in a subspace, and starts with a set of one-dimensional units. Only units with density above a given threshold are retained as dense units. Higher dimensional units are identified using a bottom-up approach similar to Apriori reasoning [2].

Given a collection of data items, S, that is a cluster in a k-dimensional space, S can also be defined as a cluster in every $(k-1)$-dimensional projection of the space. That is, if S is a cluster in the three-dimensional space $(\vec{x}_a, \vec{x}_b, \vec{x}_c)$, then it is also a cluster in the two-dimensional spaces $(\vec{x}_a, \vec{x}_b)$, $(\vec{x}_a, \vec{x}_c)$ and $(\vec{x}_b, \vec{x}_c)$. Based on this the clustering algorithm extracts k-dimensional dense unit candidates by joining $(k-1)$-dimensional dense units that share $(k-2)$ dimensions. Hence $(\vec{x}_a, \vec{x}_b, \vec{x}_c)$ can be extracted by combining $(\vec{x}_a, \vec{x}_b)$ and $(\vec{x}_a, \vec{x}_c)$ which share the dimension $\vec{x}_a$. Candidates with a density above the density threshold are retained as k-dimensional dense units. The algorithm iterates until no more candidates can be joined.

The computation time of this kind of cluster algorithms is highly dependent on the number of dense units that can be joined into candidate units and by the highest dimensionality of the dense units. Since any dense unit with k dimensions are also a dense unit in all projections in a subset of k, the time complexity is $O(c^k)$, where c is a constant [16, 1], and the running time is hence exponential to the highest dimensionality of any cluster. Due to this any major speed-up of the algorithm is related to either reducing the number of dense units or limiting the maximum dimensionality of the dense units.

The number of dense units can be reduced by employing an adaptive grid size approach [16], where one-dimensional regions with similar densities are identified as units of varying sizes. Since the range of these units differs, an individual density threshold, $\theta = \frac{\alpha \beta N}{D}$, is computed for every region where α is a cluster dominance factor, which indicates how much the density of a region must deviate from a uniform distribution to be considered a cluster region, β is the range of
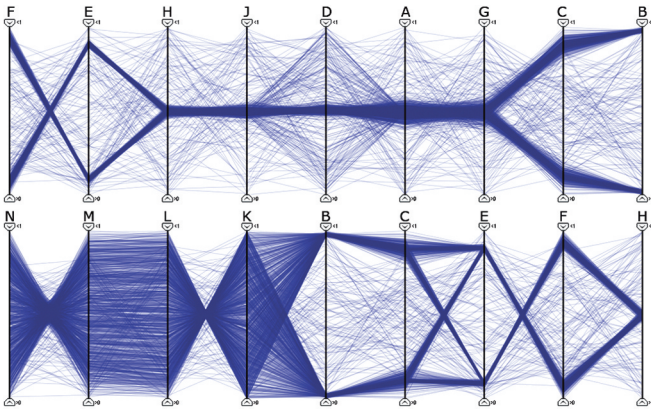
Fig. 2. The synthetic data set reduced to 9 variables using different quality metric weights and variable orders. In the top view clustering is assigned a large weight and the variables are ordered to enhance the cluster structures. In the bottom view a corresponding weighting and ordering is made for correlation structures.

the unit, N is the total number of data items in the data set and D is the range of the variable containing the one-dimensional unit. A k-dimensional unit is considered dense if its density is higher than the thresholds of all one-dimensional units of which it is composed.

Within the proposed system the clustering algorithm has been slightly modified to further speed-up the cluster detection by using a variable removal approach inspired by the retaining of 'interesting' subspaces in [1]. The goal of the cluster analysis in the system is to identify variables with high importance for cluster structures. Due to this only variables with high cluster coverage are of interest. Hence, units belonging to variables where only a small number of items are part of a cluster are removed from further analysis. Furthermore the maximum dimensionality of clusters is limited using a dimensionality threshold, ξ. The creation of candidate dense units iterates until no more candidate units can be extracted or until $k = \xi$. The maximum cluster dimensionality is defined by the user, who can also control the cluster dominance factor α. In [16] an α-value above 1.5 is said to be acceptable, and based on this, three default values are presented to the user, similarly to the outlier constraints of outlier detection.

A quality value, σ_c, is computed for each cluster based on its density, d_c, its dimensionality, k_c, and the fraction of the data set that it covers, f_c. A dense cluster with high coverage in a subspace with a large number of variables is considered to be of high quality. $\sigma_c = \tilde{d}_c \tilde{k}_c \tilde{f}_c$, where $\tilde{d}_c, \tilde{k}_c, \tilde{f}_c \in [0, 1]$. Individual cluster importance values, $I_{\text{clust}}(\vec{x}_j)$ for variables are computed by summing σ_c for all clusters where the variable is part of the subspace and where $f_c \geq \varphi$, where φ is a small user defined value, typically 0.02, used to avoid large numbers of insignificant clusters, containing only a small fraction of the data set, adding up to what appears to be a significant cluster value.

3.2 Weighting

As described in section 3.1 a variable quality value is computed for each quality metric and for each variable. To provide control of importance of quality metrics each of these values is normalized so that maximum variable importance for a metric equals 1 and minimum equals 0. The relative importance of the individual metrics is defined by assigning weight values, and a global importance value, $I(\vec{x}_j)$, is computed for each variable based on these. Through the weight values a single highly important metric can be given high impact on the dimensionality reduction, and in the same way metrics that are of no importance do not have to be considered at all. If a single metric is of interest to the user, the weight values of the others can be set to 0. Equation 2 shows how to compute $I(\vec{x}_j)$ when using correlation, outliers and clusters as quality metrics and where w_{corr}, w_{out} and w_{clust} are the weight values.
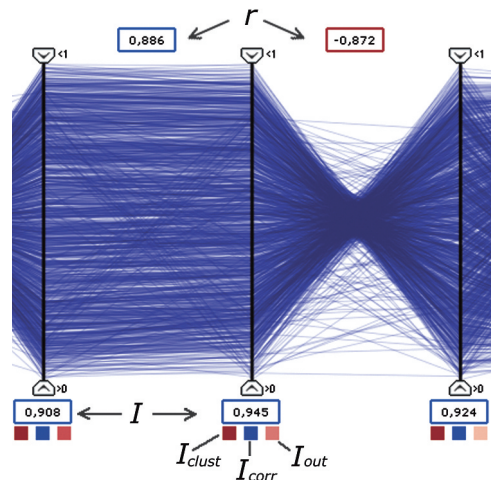

Fig. 3. The visual aids of a parallel coordinates display, facilitating understanding of the importance and quality metrics of individual variables. r is the correlation between adjacent axes, with negative correlation in red and positive in blue. I is the global importance value, and I_{clust}, I_{corr} and I_{out} are the cluster, correlation, and outlier quality values.

$$I(\vec{x}_j) = w_{\text{corr}} I_{\text{corr}}(\vec{x}_j) + w_{\text{out}} I_{\text{out}}(\vec{x}_j) + w_{\text{clust}} I_{\text{clust}}(\vec{x}_j) \qquad (2)$$

The part of the system work flow that is most computationally heavy is the quality metric analysis described in section 3.1, but this depends entirely on the computation time of the selected quality metrics. Once that analysis is performed it will not need to be re-performed unless any quality metric parameters have to be changed. Hence modification of weight values and re-computation of variable importance is performed in a few milliseconds even for data sets with hundreds of variables. In figure 2 two examples of the synthetic data set reduced to 9 variables, using different weight values, are displayed using parallel coordinates. In the top view clusters are given five times as high a weight as correlation and outliers and in the bottom view correlation is given five times as large weight value. As can be seen the highly correlated variables to the left in the bottom view are not part of the reduced data set when clustering is assigned a higher importance (top).

To facilitate the understanding of the importance and structures within the individual variables, the user is provided with visual aids (figure 3). Using colour, ranging from red (low importance) to blue (high importance), the global importance value as well as the value of each individual quality metric is displayed in an easily perceived manner.

3.3 Information Loss and Selection of Data Set Size

The appropriate size of a reduced data set that is to be visualized is dependent both on the structure of the data and on the task of analysis, as well as on the size of the display to use. Using user-defined quality metrics to analyse the structures within a data set and to extract individual importance values enables a task- and user-oriented approach to identification of important variables. Within the presented system a reduction from an M-dimensional to a K-dimensional data set is performed by retaining the K variables with highest global importance value, $I(\vec{x}_j)$, and by removing the remaining $(M - K)$ variables. When a weighting is performed, as described in section 3.2, the value of $I(\vec{x}_j)$ is recomputed for each variable resulting in a new set of variables being selected as the K most important.

To enable investigation of the trade-off between number of variables retained in the reduced data set and amount of information lost, and hence to facilitate the decision on appropriate size for the reduced data set, the system presented in this paper offers an interactive display presenting the relationship between number of variables to keep and the amount of information lost, $I_{lost} = \frac{I_{removed}}{I_{total}}$, where $I_{removed}$ is the
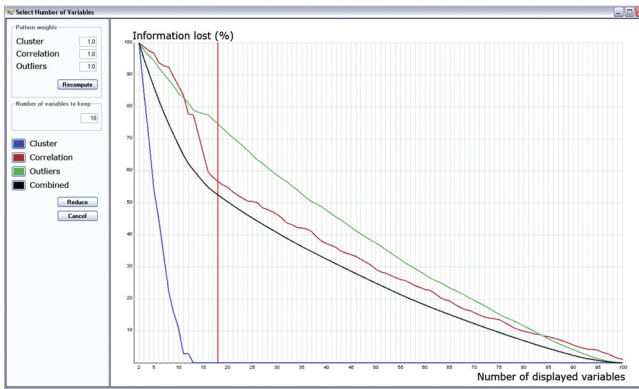
Fig. 4. Interactive display of the amount of information lost relative to number of variables to keep in the reduced data set. The black line represents the combined information loss for all quality metrics, the blue, red and green lines represent information loss in cluster, correlation and outlier structures respectively. The red vertical line corresponds to the number of variables currently selected.

Fig. 5. Example of variable ordering algorithm for cluster enhancement. Initially the clusters are ordered according to quality values. For each iteration the reordering is found that results in the longest sequence of connected variables being part of c_i, without traversing the borders of previous clusters (represented by red and pink rectangles)

sum of $I(\vec{x}_j)$ for the removed variables and I_{total} is the sum of $I(\vec{x}_j)$ for all variables in the data set.

The interactive display (figure 4) consists of a line graph and a graphical user interface for modification of weight values and selection of number of variables to keep. The line graph displays the relationship between I_{lost} (y-axis) and number of variables to keep in the reduced data set (x-axis), representing each quality metric individually by a line and using one line for the combined importance value of all metrics. A similar approach is taken in [6], where quality measures for data abstractions such as clustering and sampling are integrated into multivariate visualizations. A vertical line is used in the interactive display to facilitate identification of lost information for the selected number of variables. If retaining 18 variables, according to the position of the vertical line in figure 4, it can be seen from the display that some of the retained variables contain no cluster information at all. In figure 6 the corresponding 18 variable data set is visualized using parallel coordinates. As can be seen from the visual aids at the bottom of the axes, the five left variables are of low global importance and also have low cluster and correlation importance. By looking at the patterns of the lines it is also quite easily seen that these variables are rather noisy, hence more variables can be removed from the data set without losing much more information.

3.4 Variable Ordering

The order of variables in multivariate visualization has a large impact on how easily we can perceive different structures in the data. The proposed system combines several quality metrics to find a dimensionality reduction that can be regarded as a good representation of the original data set, focusing on the structures that are of interest for the particular analysis task at hand. Finding one appropriate variable ordering enhancing all interesting structures at once may, however, be unrealistic since enhancement of some structures obstructs enhancement of others. An aim of the system is to provide the user with good visual representations of the reduced data set by enhancing the existing structures within the data. The ordering of variables in the system has been approached by supplying a selection of automatic orderings, each enhancing the structures of an individual quality metric. The user can interactively switch between different variable orderings and to enable comparison of different variable orders multiple displays are used.

Within this paper three quality metrics are discussed, correlation, outliers and clusters. The variable ordering focusing on correlation is inspired by the variable ordering described in [4]. In connection to the correlation analysis in the initial phase of the dimensionality reduction, johormation on the correlation of variable pairs is computed. When the variables of the reduced data set are to be ordered, the variable pair, $\vec{x}_a\vec{x}_b$, with the highest correlation within this data set is identified.

This pair forms the basis of the ordering. Next the variable pair with the highest correlation containing $\vec{x}_a$ or $\vec{x}_b$ and a variable that is not yet ordered is identified. The unordered variable is positioned at the left or right border of the ordered variables, next to the variable with which it forms a highly correlated pair. This continues by identifying variable pairs with highest correlation containing one of the ordered variables positioned at the leftmost or rightmost position and one of the not yet ordered variables, until all variables are ordered.

The variable orderings enhancing cluster and outlier structures are based on the quality values calculated for each cluster or outlier in connection with the cluster and outlier detection, and are performed in the same way. An example of the ordering algorithm, based on cluster structures, is shown in figure 5, where $S_{reduced}$ is the set of variables retained after dimensionality reduction. The ordering algorithm is performed as follows:

1. Initially the clusters are sorted in descending order according to quality value, as shown in figure 5 where the ordering is to be based on three clusters, c_0, c_1 and c_2.

2. In the first iteration all variables in $S_{reduced}$ that are part of the first cluster, c_0, are positioned next to each other. In figure 5, c_0 includes variable 6. This variable is not part of $S_{reduced}$ and is hence not taken into consideration. The red rectangle in the figure represents the positions of the border variables of c_0.

3. In the subsequent iterations the variables in c_i that are part of $S_{reduced}$ and of any cluster, c_j where $j < i$, are identified. For c_1, for instance, variables 2, 3 and 10 are part of $S_{reduced}$ and variables 3 and 10 are also part of c_0.

4. The reordering of variables in $S_{reduced}$ that results in the longest sequence of connected variables being part of c_i, without any variables traversing the border positions of previous clusters (red rectangles) are found, and $S_{reduced}$ is reordered accordingly. For $i = 1$ this is achieved by switching the positions of variables 7 and 10, and positioning variable 2 next to variable 3.

5. The algorithm iterates from step 3 until no more re-orderings are possible, that is, if all variables in $S_{reduced}$ are separated by cluster borders, or until reordering is performed for all clusters.

In figure 2 an example of the synthetic data set reduced to 9 variables is displayed using parallel coordinates with two different variable orders. The top view is ordered according to cluster structures and the bottom view according to correlation.

4 RESULT

This system has been implemented using C# and DirectX. All tests have been run on a desktop computer with an Intel 2 GHz CPU, 2 GB of RAM and with an NVIDIA GeForce 7800 GTX graphics card.
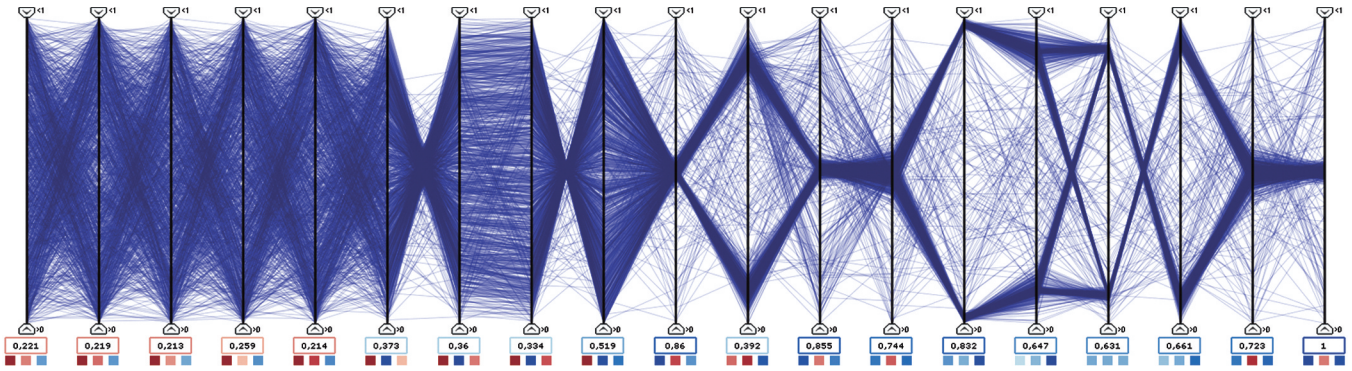
Fig. 6. The synthetic data set reduced to 18 variables, according to figure 4. The additional information at the bottom of the axes indicates by the red colour that the five leftmost variables, which mainly contain noise, have little importance in representing the structures of the data set.

Table 1. Computation time in seconds for quality metrics analysis in a synthetic data set containing 100 variables and 1320 items, with cluster dimensionality thresholds, ξ, and cluster dominance factors, α

	$\xi = 5$	$\xi = 7$	$\xi = 10$	$\xi = 12$
$\alpha = 1.5$	11	12	12	12
$\alpha = 1.7$	11	12	12	12
$\alpha = 1.9$	11	11	11	11

Table 2. Computation time in seconds for quality metrics analysis in the demographic data set for one year, containing 47 variables and 227 items. Due to more complex cluster structures the computation time is much higher for this data set than for the synthetic data set.

	$\xi = 5$	$\xi = 7$	$\xi = 10$	$\xi = 12$
$\alpha = 1.5$	115	356	364	386
$\alpha = 1.7$	83	220	245	245
$\alpha = 1.9$	18	51	58	58

To demonstrate the interactive features of the system and to provide understanding of how an analyst can work with it, a case scenario is described in this section. The analysis is performed on demographic data sets collected from the World Bank public database[1]. The data contains 47 variables for 227 countries during the years 2000 to 2007, and is divided into one data set per year.

4.1 Performance

The only computationally heavy part of the system is the quality metrics analysis for which the time complexity entirely depends on the quality metrics used. This analysis is performed initially and due to the structure of the system, where weighting, dimensionality reduction and variable ordering are all based on the quality values, all modifications within the system are performed in a few milliseconds once the quality values are computed.

The quality metrics used in this paper are correlation, outliers and clusters, as described in section 3.1. The total computation time for analysis with these metrics mainly depends on the cluster structures in the data set and on the maximum cluster dimensionality threshold, ξ. Tables 1 and 2 present the total computation time in seconds, using the synthetic data set presented in section 3 and the demographic data set of one year. Within the synthetic data set 10 variables contain strong cluster structures, 8 variables contain strong correlation structures and 11 contain strong outlier structures. Although smaller than the synthetic data set, the demographic data set contains more complex cluster structures, and as a result the quality analysis of this data set is more time consuming. However, the computation time can be greatly reduced by limiting the maximum cluster dimensionality, ξ, or by increasing the cluster density constraint, α, as shown in table 2.

4.2 Case Scenario

This section describes how a fictional person, Marie, who is a masters student in a social science programme writing an essay on the topic of development of countries, could use the presented system to identify subsets of interesting development indicators from the demographic data sets. Through previous work experience she is fairly familiar with interactive information systems and visual representations

of data. Prior to the analysis Marie does not know much about the existing structures within the data, but is interested in development indicators that are strongly related to each other and in groups of indicators for which countries follow a similar pattern. To avoid assigning significance to structures that exist only as a result of the number of years, she decides to analyse the individual years separately.

Initially Marie loads the data for 2001 into the system and sets the quality metric parameters. To identify the most important cluster structures but also to limit the computation time, she uses a maximum cluster dimensionality of 10 and the default 'medium' cluster dominance factor of 1.7. Marie is only interested in quite strong outlier structures and sets the outlier constraints slightly higher than the 'medium' default and sets the minimum number of variables to 10. The computation of importance values is performed and Marie is presented with the lost information display (figure 7). As can be seen the loss of cluster information is low compared to the loss of correlation and outlier
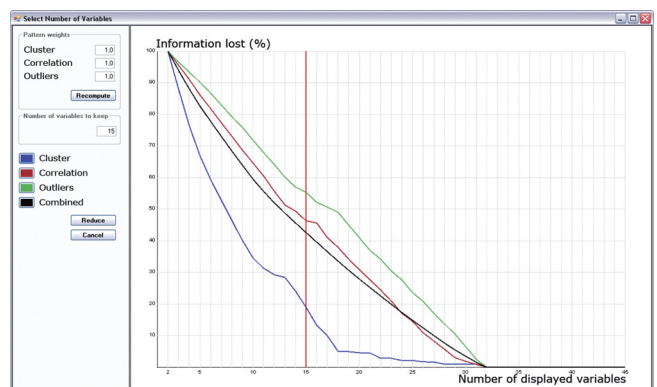

Fig. 7. Display of the information loss compared to number of variables in the reduced data set. The line representing loss of cluster information (blue line) is much lower than the loss of other structures. The red vertical selection line is positioned at the 15 variables position.
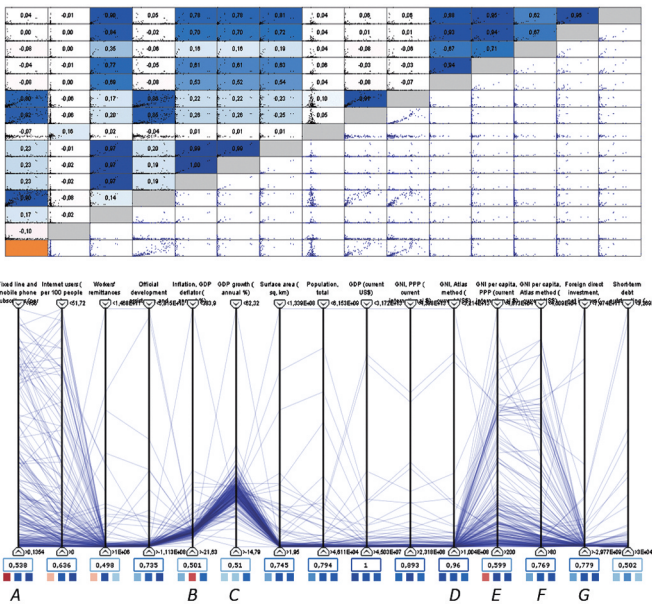
Fig. 8. The data set reduced to 15 variables, displayed using a scatter plot matrix and parallel coordinates. In the scatter plot matrix the correlation of variable pairs is displayed using text information and colour.
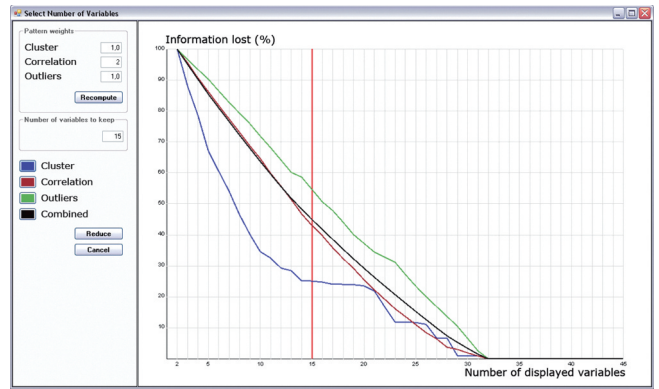

Fig. 9. The lost information display when correlation is given twice as much weight as cluster and outlier structures. The total amount of correlation lost when reducing the data set to 15 variables has decreased to below 45 percent.

information. Marie tries a reduction to 15 variables by dragging the vertical red line to corresponding position, where less than 20 percent of the cluster information and less than 45 percent of the combined information is lost. Marie is mainly interested in cluster and correlation structures, and examines the data set using a scatter plot matrix where the variables are ordered to enhance correlation structures, and parallel coordinates ordered to enhance clusters (figure 8). In the scatter plot matrix colour is used to enhance correlation, blue for positive and red for negative, where dark colours indicate a high correlation.

From the scatter plot matrix in figure 8 Marie realizes that there are no strong negative correlations between any variables in the reduced data set. Furthermore she is able to easily identify groups of variables that are strongly related to each other, which form islands of coloured cells. In the parallel coordinates one major cluster is easily perceived, stretching over most variables at the bottom of the axes. Within this cluster Marie finds the variable marked C, which is the annual gross domestic product (GDP) growth, especially interesting. This variable contains strong outlier structures, with several countries outside the cluster that have very high or very low annual GDP growth. Furthermore some lower dimensional cluster structures can be found, for instance in the axis sequence marked D, E, F and G, where D represents a measure of total gross national income (GNI), E and F represent two different measures for GNI per capita, and G represents foreign direct investment. Below the axes of the parallel coordinates the global importance values are shown together with colour information on the individual quality values. According to this information the variable marked A, representing the number of fixed line and mobile phone subscribers, have a low importance value on clusters, as shown by the dark red square, but with high value on both correlation and outliers. Furthermore she notices the inflation variable, marked B, which is the only variable with low correlation value.

To get more information on strongly related indicators, Marie decides to perform a reduction where correlation is given a higher importance than the other metrics, and re-opens the lost information display. Here she increases the correlation weight value. The variable importance values are instantly recomputed and the display is updated, as shown in figure 9. Now the amount of lost correlation information has decreased to below 45 percent for a reduction to 15 variables, whereas the amount of cluster information lost has increased to around 25 percent. The result of the weighted reduction to 15 variables is shown in figure 10. Within this reduced data set Marie immediately notices the

four variables marked A, B, C and D, which all have weak cluster importance but high correlation and outlier importance. Of these only A, representing number of fixed line and mobile phone subscribers, was part of the previously reduced data set. Hence Marie draws the conclusion that variables B, C and D, representing export of goods and services, merchandise trade and electric power consumption, might have some interesting countries not following the overall patterns although they are probably strongly related to other indicators.

Marie decides to make a comparison to another year, to get a clue to which relationships are consistent over time and which might indicate temporary structures for an individual year. She selects the data set for 2004 and performs quality metric analysis with the same parameter settings as before. The lines of the lost information display follow a similar pattern as for 2001 and, due to this, Marie reduces the data set to 15 variables to make a comparison with her previous findings. The reduced data set is displayed using parallel coordinates in figure 11, where the variables in the top view are ordered to enhance cluster structures and in the bottom view to enhance correlation.

This reduction contains most of the same variables as the 2001 data set, with the difference that for 2004 total external debt and merchandise trade are two of the 15 most important variables. Due to the different variable ordering in the views Marie is able to detect some varying patterns within the reduced data set. For instance the variables marked B, C, D, E and F contain some low dimensional cluster structures that are not as visible in the bottom view, and the bottom view displays the almost perfect positive correlation between variables G, H and I which represent GDP, GNI and the inflows of foreign direct investments. This pattern is hard to perceive in the top view where A is the foreign direct investment and F is the corresponding GDP variable.

Marie continues her analysis in a similar fashion for all years between 2000 and 2007, and identifies variables containing interesting structures and relationships. Through this she achieves an understand-
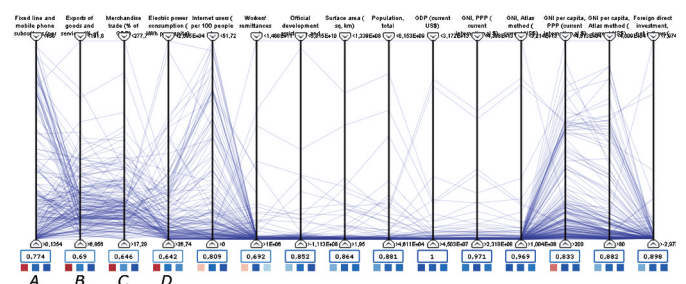

Fig. 10. The 2001 data set reduced to 15 variables when correlation is given twice the weight of cluster and outlier structures.
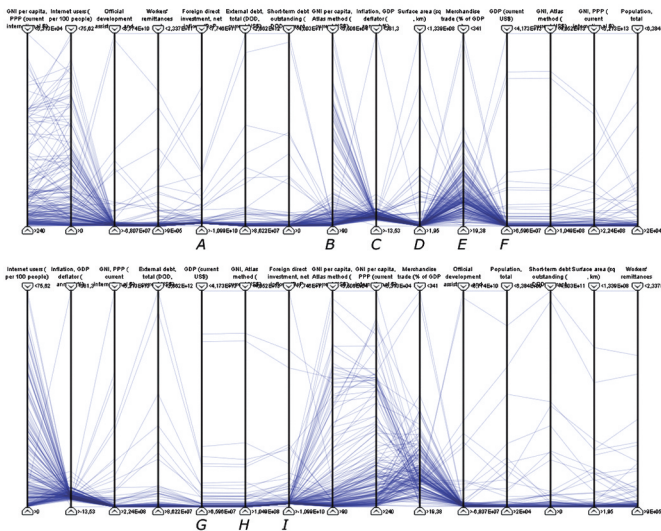
Fig. 11. The data set for 2004 reduced to 15 variables, displayed using parallel coordinates with variables ordered to enhance clusters (top) and correlation (bottom).

ing of some of the structures and relationships between different indicators and is aware of structures to examine more closely and pay attention to in her essay.

5 CONCLUSIONS AND FUTURE WORK

This paper introduces a dimensionality reduction system, aiming at effective exploration and visualization of multivariate data sets with hundreds of variables. The benefits of the system lie in its ability to combine user-defined and weighted quality metrics to preserve as many important structures as possible, and in enabling a quality-guided reduction of variables, hence providing a flexible, task-dependent and user-controlled dimensionality reduction and analysis environment.

The generality and interactivity of the system is presented through a case scenario, where the features of the system are used to identify significant variables and structures in a demographic data set including 47 variables. This scenario describes how an analyst can work with the system and demonstrates that the selection of number of variables to keep is facilitated through visual exploration of the trade-off between loss of information and number of variables, and that diverse quality metrics can be combined and assigned different importance using weight values, supplying flexible dimensionality reductions.

Next step is to perform a thorough evaluation to validate the effectiveness and usability of the system and to compare it with other available dimensionality reduction methods. Future work also includes implementation of additional quality metrics and improvements of the efficiency of the system and of existing quality metrics.

ACKNOWLEDGMENTS

REFERENCES

[1] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *ACM SIGMOD International Conference on Management of Data*, pages 94–105, 1998.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, pages 487–499, 1994.

[3] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering for an enhanced visualization of multidimensional data. In *Proceedings of IEEE Symposium on Information Visualization, InfoVis '98*, pages 52–60, 1998.

[4] A. O. Artero, M. C. F. de Olivera, and H. Levkowitz. Enhanced high dimensional data visualization through dimension reduction and attribute arrangement. In *Proceedings of the conference on Infomation Visualization IV'06*, pages 707–712. IEEE, 2006.

[5] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, May 1987.

[6] Q. Cui, M. O. Ward, E. A. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):709–716, Oct. 2006.

[7] S. G. Eick and A. F. Karr. Visual scalability. *Journal of Computational & Graphical Statistics*, 11(1):22–43, 2002.

[8] D. Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2(4):232–246, 2003.

[9] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(4):69–91, 1985.

[10] G. Ivosev, L. Burton, and R. Bonner. Dimensionality reduction and visualization in principal component analysis. *Analytical Chemistry*, 80(13):4933–4944, 2008.

[11] I. T. Jolliffe. *Principal Component Analysis, 2 ed.* Springer-Verlag, 2002.

[12] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.

[13] T. Kohonen. The self-organizing map. *Neurocomputing*, 21(1–3):1–6, 1998.

[14] Y. Koren and L. Carmel. Robust linear dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 10(4):459–470, 2004.

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

[16] H. Nagesh, S. Goil, and A. Choudhary. Adaptive grids for clustering massive data sets. In *First Siam International Conference on Data Mining*, 2001.

[17] F. Paulovich and R. Minghim. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1229–1236, 2008.

[18] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of IEEE Symposium on Information Visualization*, pages 89–96, Oct. 2004.

[19] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, pages 318–322. ACM, 1994.

[20] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, Feb. 1988.

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

[22] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of American Statistics Association*, 85(411):664–675, 1990.

[23] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings of IEEE Symposium on Information Visualization, InfoVis '05*, pages 157–164, 2005.

[24] M. Williams and T. Munzner. Steerable, progressive multidimensional scaling. In *Proceedings of IEEE Symposium on Information Visualization*, pages 57–64, Oct. 2004.

[25] 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 IEEE Symposium on Information Visualization 2004, INFOVIS 2004*, pages 73–80, October 2004.

[26] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proceedings of IEEE Symposium on Information Visualization*, pages 105–112, 2003.

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