INTERPRETATION, INTERACTION, AND SCALABILITY FOR STRUCTURAL DECOMPOSITION TREES

René Rosenbaum[1], Daniel Engel[2], James Mouradian[1], Hans Hagen[2] and Bernd Hamann[1,2]

[1]*Institute for Data Analysis and Visualization (IDAV), Department of Computer Science, University of California, Davis, CA 95616-8562, U.S.A.*

[2]*International Research Training Group (IRTG) 1131, University of Kaiserslautern, Kaiserslautern, Germany*

Keywords: High-dimensional Data, Projections, Value Visualization, Relation Visualization, Interaction, Scalability.

Abstract: Structural Decomposition Trees (SDTs) have been proposed as a completely novel display approach to tackling the research problem of visualizing high-dimensional data. SDTs merge the two distinct classes of relation and value visualizations into a single integrated strategy. The method is promising; however, statements regarding its meaningful application are still missing, constraining its broad adoption. This paper introduces solutions for still-existing issues in the application of SDTs with regard to interpretation, interaction, and scalability. SDTs provide a well-designed initial projection of the data to meaningfully represent its properties, but not much is known about how to interpret this projection. We are able to derive the data's properties from their initial representation. The provided methods are valid not only for SDTs, but also for projections based on principal components analysis, addressing a frequent problem when applying this technology. We further show how interactive exploration based on SDTs can be applied to visual cluster analysis as one of its application domains. To address the urgent need to analyze vast and complex amounts of data, we also introduce means for scalable processing and representation. Given the importance and broader relevance of the discussed problem domains, this paper justifies and further motivates the usefulness and wide applicability of SDTs as a novel visualization approach for high-dimensional data.

1 INTRODUCTION

The visualization of high-dimensional data sources is a common but still unsolved problem. Structural decomposition trees (SDTs) (Engel et al., 2011) represent a novel approach to this challenge. SDTs combine value and relation visualizations into one approach and thus provide a variety of benefits not available in existing visualization technology. Means for interaction based on the novel representation add unique options for data exploration. Research concerning SDTs, however, is constrained to the introduction and description of the construction and fundamental aspects of this novel displaying approach. Statements concerned with its utilization and application to concrete problems have not yet been published. This results in long learning efforts and may also lead to misinterpretations and failure when the technique is applied.

This paper aims to provide guidance for different aspects important for the successful application of SDTs in data visualization. After reviewing related work in the area of high-dimensional data visualization (Section 2), the first part of this paper (Section 3) is particularly concerned with the interpretation of an SDT. Thereby, we focus on alignment and length of the different dimensional anchors and derive their meaning and mathematical properties for gaining first insight quickly. Practical implications of these statements provide the users with an understanding of the capabilities of SDTs. Introducing guidelines for visual cluster analysis, the second part (Section 4) is concerned with appropriate interactive data exploration. Based on a common data browsing approach, strategies for gaining further insight into the data are provided. Addressing the need for visualization technology able to deal with large and high-dimensional data sets, the third part (Section 5) of the paper discusses different means to reduce the complexity of the data and representation. Empirical results show that much time can be saved when methods to introduce scalability with regard to the number of data points and dimensions are applied. We show that ambiguities in the association of data points to data clusters

Rosenbaum R., Engel D., Mouradian J., Hagen H. and Hamann B..
INTERPRETATION, INTERACTION, AND SCALABILITY FOR STRUCTURAL DECOMPOSITION TREES.
DOI: 10.5220/0003841106360647
In *Proceedings of the International Conference on Computer Graphics Theory and Applications* (IVAPP-2012), pages 636-647
ISBN: 978-989-8565-02-0

can be significantly reduced by a scalable visual representation.

The main contributions of this paper to the current state of research are the following

- We provide statements and practical implications to the geometric interpretation of SDTs and related projections.
- We introduce means for an interactive exploration of a high-dimensional data set in the context of visual cluster analysis.
- We introduce and justify means for a scalable processing and representation of large and complex data sets.

We conclude (Section 6) that SDTs are a valid means for the visualization of high-dimensional data, but must be understood and applied in the right way in order to gain insight quickly and without misinterpretation and wrong conclusions. This paper aims to provide the necessary information to accomplish this objective successfully.

2 RELATED WORK

2.1 Visualization of High-dimensional Data

As a result of most data acquisition tasks today, high-dimensional data has always been of strong interest to the visualization community. Many different approaches and techniques have been proposed. According to (Engel et al., 2011) they can be categorized into *value* or *relation visualizations*.

By focussing on the conveyance of data coordinate values for every data point, **value visualizations** allow for a detailed analysis of the data. The parallel coordinates plot (see Figure 1, top/left) is a typical representative of this category. Due to their focus on value representation for each data point, a common problem with all associated techniques is that they are often not scalable with regard to the amount and dimensionality of the data. As a result this usually leads to clutter and long processing times as the number of dimensions and amount of data points increase. In order to overcome these issues, cluster-based approaches (Johansson et al., 2005; Zhou et al., 2008; Artero et al., 2004), appropriate means for interaction (Elmqvist et al., 2008; Hauser et al., 2002), and better utilization of the available screen space (McDonnell and Mueller, 2008) have been proposed. Clutter reduction is also achieved by dimension ordering arranging the dimensions within the visual representation based on correlations within the data. The research conducted by ANKERST ET AL.(Ankerst et al., 1998) was the first to formally state this problem and has later been successfully expanded in (Yang et al., 2003a) and (Peng et al., 2004). Although these methods are great improvements to reduce clutter, the displayed information is often too detailed and a meaningful representation can generally not be obtained for large data sets.

Instead of aiming at communicating individual values, **relation visualizations** are designed to convey relationships within the data. They are mostly point mappings projecting the m-dimensional (m-D) data into the low-dimensional presentation space. As relations within the data may be too complex to be completely conveyed in presentation space, projections are usually ambiguous. A prominent and widely applied point projection approach is principal components analysis (PCA) conveying distance relations in m-D space by projecting into a plane that is aligned to capture the greatest variance data space without distorting the data (see Figure 1, top/right). Contrary, multi-dimensional scaling (MDS) commonly uses general similarity measures to represent the data, but leads to distortion and a visualization that may be difficult to interpret. Interpretation of the representation is a general issue with point projections as long as no means to comprehend the parameters used for the projection are available. One such option are dimensional anchor (DA) visualizations (Hoffman et al., 1999) projecting and displaying the basis vectors along the data points (see Figure 1, top/right). These DAs are also an appropriate means to adjust the projection interactively (Kandogan, 2001). Relation visualizations usually lead to a meaningful overview of the data. Their effectiveness, however, strongly depends on the quality of the initial projection and the means provided to interpret and interact with it. Current research mainly focuses on improved representation of specific data structures, e.g., scientific point cloud data (Oesterling et al., 2010), a better incorporation of domain-appropriate analysis techniques, e.g., brushing and filtering (Jänicke et al., 2008), or computational speed gains (Ingram et al., 2009).

Due to the rather diverse properties of value and relation visualizations, they each have distinct application domains. Thus, they are often used simultaneously in exploratory multi-view systems, such as in (Paulovich et al., 2007). Few publications exist that tackle the problem of combining both classes into a single approach. Most of them have been proposed for value visualizations, such as the technology described in (Yang et al., 2003a; Yang et al., 2003b;

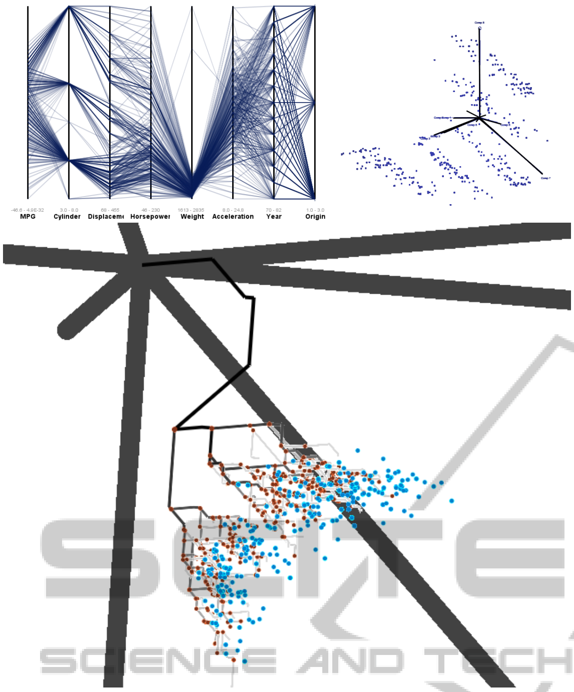


Figure 1: The two classes of visualizations for high-dimensional data (value (top/left) and relation (top/right) visualizations) are brought together by SDTs (bottom). All visualizations represent the well-known "cars" data set. The SDT highlights the five distinct clusters by its branch structure also conveying the respective differences in the data values.

Yang et al., 2004), (Johansson and Johansson, 2009), or (Yuan et al., 2009). SDTs represent a completely different approach that promises to bridge the existing gap between both classes.

2.2 Structural Decomposition Trees

SDTs are founded on a sophisticated data projection, but provide additional means to represent the dimension contributions for each data point (see Figure 1, bottom). This is mainly achieved by introducing a tree structure showing the projection path for each displayed data point and thus its individual dimension values. The projection paths also allow for an unambiguous identification and interpretation of data points that reside at different locations in m-D space, but have been projected in close proximity in the projection space.

A main problem in showing the different projection paths is the introduced clutter. SDTs overcome this issue by introducing a multi-stage processing pipeline. Hierarchical clustering is used to identify, aggregate, and bundle common line segments of m-D data points. The resulting tree has minimal overall branch length reducing the redundancies considerably. Appropriate representation of the individual dimension contributions is accomplished by a well-designed drawing order. The tree itself is represented by colored lines whereby the number of elements within this subtree is encoded by branch thickness (see Figure 1, bottom). The initial SDT projection thereby maximizes the space between tree paths and thus allows for a better interpretation of the visualization (Engel et al., 2011).

Different means for interaction either on individual or groups of data points or the whole representation make possible for further exploration of the data (see Figure 2). The projection can be significantly changed by a re-arrangement of the end points of the DAs. These dimension vectors can be independently modified in their lengths and angles relative to each other. Thereby, so-called variance points are placed along the unit circle in order to indicate angles that lead to other promising projections. Means for highlighting dimensions and projection paths allow one to emphasize all line segments corresponding to the coordinates of a dimension or to investigate the structural decomposition of the data. Data filtering can be achieved by collapsing and hiding subtrees. After collapsing, the main value contributions of all associated data points are still visible and can be used and interpreted, e.g., for comparison with other subtrees.

Published research concerned with SDTs mainly focusses on the technical foundations of the approach. Semantic aspects, interpretation, and the use of SDTs to visualize large data sets are not provided.

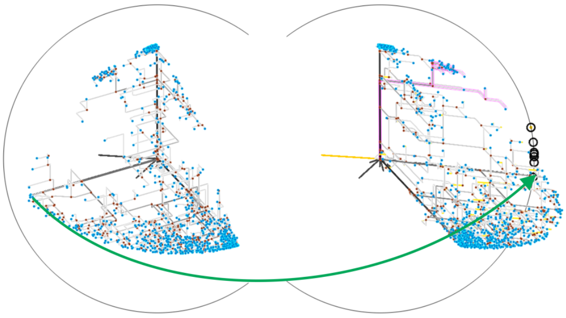


Figure 2: The main interactions provided by SDTs: *Repositioning of DAs* (green arrow) allows for an intuitive adjustment of the projection. *Dimension highlighting* (yellow DA) conveys the individual contributions of a dimension in the data (yellow tree segments). *Path highlighting* (purple tree branch) is intended to emphasize interesting tree branches and substructures.

3 INTERPRETATION OF THE INITIAL LAYOUT

Projections are a powerful means to convey relations in high-dimensional data. Due to the characteristics of dimension reduction, however, they are often difficult to interpret. In previous work it was shown that SDTs are specifically suited to depict data coordinates in a way that aims at intuitive interpretation. Experimental studies of PCA-based projections showed that the projection conveys properties of the data by the length and relation of the DAs to each other. This, however, has never been explicitly quantified.

In this section, we investigate in full detail how the initial arrangement of DAs in SDTs relates to the corresponding variables in the data and how the user can interpret this arrangement to infer knowledge about the data. Due to the use of a PCA-related projection method for SDTs, the given statements apply to all PCA-based projections. We shortly recall *dimensional anchors and their arrangement*, after which we are *linking the properties of DAs to those of the data*. We first outline why DAs are used to reflect a PCA projection and how their initial arrangement is defined. This is expressed by latent features in the data, i.e., the eigenvectors and eigenvalues of the data's covariance matrix indicating the information content within the different data dimensions. In order to understand which data properties are visually encoded in a projection, we investigate how the projection is defined by these features and what information is thereby depicted. This is expressed by a derivation of the *spectral decomposition of the covariance matrix*. After these steps, we show that the specific DA arrangement allows one to derive *conclusions* and data properties that are of keen interest to the user but not depicted by the common plotting of principal components. Finally, statements to *implications* of these properties aim for a better understanding of an arbitrary PCA-based projection avoiding its misinterpretation.

DAs and their Arrangement. Since SDTs can be computed and visualized both in 2D or 3D space, the following considerations are made for an arbitrary display dimensionality p. We assume that n m-dimensional data points are stored row-wise in $X \in \mathbb{R}^{(n\times m)}$. By the projection, X is mapped to display point coordinates $\widetilde{X} \in \mathbb{R}^{(n\times p)}$.

The linear mapping of an m-D data point X_i in p-D display coordinates $\widetilde{X}_i$ is computed by the linear combination of DAs $a_j \in \mathbb{R}^p$, $1 \leq j \leq m$, with the corresponding coordinate of X_i:

$$\widetilde{X}_i = \sum_{1\leq j\leq m} a_j X_{i,j}. \qquad (1)$$

This technique, the mapping in star coordinates (Kandogan, 2001), can be understood as a generalization of drawing 3D objects on paper to arbitrary dimensions. In the original work, however, the DAs are initially arranged in a uniform distribution along a unit circle. In general, this leads to a non-orthogonal projection. This can be misleading because the distance in display space does not reflect distance in $\mathbb{R}^m$. To avoid this, a projection is designed to minimize this mapping error. This error is commonly expressed as the sum of squared pairwise distance differences arising from the mapping from m to p dimensions, $\sum_{1\leq i,j\leq n}(D(X_i,X_j) - d_2(\widetilde{X}_i,\widetilde{X}_j))^2$, where d_2 is the Euclidean distance metric and D is an appropriate distance metric of the application domain. This error can be minimized, for example, by PCA in the case $D = d_2$. Instead of expressing the data by the original unit vectors, PCA computes new orthogonal directions (principal components) in which the data has maximal variance and re-expresses all data points in coordinates of these principal components. The projection is then defined by the p principal components that capture the highest variance in the data. Although distance relations between data points are captured well in this projection, the interpretation of principal components is not intuitive. In almost all applications, the link to the original data is essential for analysis. Therefore, the depiction of the original data coordinates and relations between the original data dimensions is an important aspect for a projection.

Linking Properties of DAs to those of the Data. In previous work (Engel et al., 2011), both approaches have been combined and the initial arrangement of DAs has been defined to reflect a (weighted) PCA projection into p-dimensional display coordinates. We utilize DAs to make possible a better interpretation and more intuitive understanding of the underlying projection without losing any of the underlying projection's benefit. In this research, we investigate the properties of this DA projection in more detail and deduct which properties of the DAs link to which properties in the data. The following considerations are based on the data's covariance matrix. Without loss of generality, we assume X to be centered and, since the used weighting scheme in previous work changes the covariance matrix (to be weighted) a priori, we can neglect the weighting in the following. We also neglect the global scaling by n^{-1} that does not influence relations in the data.

The underlying PCA projection $\widetilde{X}$ of X is defined as $\widetilde{X} = X\,\widehat{\Gamma}$, with $\widehat{\Gamma} = (\gamma^{(1)}, ..., \gamma^{(p)}) \in \mathbb{R}^{(m\times p)}$ being the matrix storing column-wise the eigenvectors of the corresponding p largest eigenvalues of the covariance matrix S of X. Equation (1) implies that the linear mapping of DAs $A = (a_1, ..., a_m)^T \in \mathbb{R}^{(m\times p)}$ is defined as $\widetilde{X_i} = X\ A$. In order to initially arrange the DAs such that their mapping is equivalent to that of the PCA, we define each DA as a row vector of $\widehat{\Gamma}$:

$$a_i = \left(\gamma_i^{(1)}, ..., \gamma_i^{(p)}\right)^T. \qquad (2)$$

This step is equivalent to the projection of the original unit vectors $\mathbf{1}_i \in \mathbb{R}^m$ to $\mathbb{R}^p$ subject to the same rotation, i.e., $a_i^T = \mathbf{1}_i^T\widehat{\Gamma}$.

It is important to note that PCA projects X by reducing its dimensionality to p in an optimal variance-preserving way. Thus, the information that is actually displayed by this projection is that of the inherently defined best rank-p approximation $\widehat{X}$ of X.

Spectral Decomposition of the Covariance Matrix. The process of dimensionality reduction by maximizing variance becomes clear when considering the spectral decomposition of S. That is the decomposition of the combined variances of all elements in X into successive contributions of decreasing variance: $S = \lambda_1\gamma^{(1)}\gamma^{(1)^T} + ... + \lambda_r\gamma^{(r)}\gamma^{(r)^T}$, with λ_k being the k highest eigenvalue of S and $\gamma^{(k)}$ the corresponding eigenvector for $1 \le k \le r = rank(X)$.

Each contribution $S^{(k)} = \lambda_k\gamma^{(k)}\gamma^{(k)^T}$ thereby increases the rank of the matrix summation by one. λ_k holds the variance of the contribution, whereas $\gamma^{(k)}\gamma^{(k)^T}$ defines the mixing of this variance, i.e., how this contributes to S. Consequently, the covariance matrix of the PCA's p-dimensional best rank-p approximation $\widehat{X}$ of X equals the sum over the first p contributions, where usually $p \ll rank(X)$. The covariance between dimensions i and j of the projected data $\widehat{X}$ is therefore

$$\widehat{S}_{i,j} = \sum_{1\le k\le p} \lambda_k\gamma_i^{(k)}\gamma_j^{(k)}. \qquad (3)$$

Similarly, $\widehat{X}$ can be defined by $\widehat{X} = X\ \widehat{\Gamma}\widehat{\Gamma}^T$. For the dimensions (columns) in $\widehat{X}$ the following equation holds: $\widehat{X}_{\bullet,i} = \sum_{1\le j\le m} X_{\bullet,j}(\widehat{\Gamma}\widehat{\Gamma}^T)_{i,j}$. $\widehat{X}_{\bullet,i}$ is built from X by the linear combination of all $X_{\bullet,j}$ with coefficients $(\widehat{\Gamma}\widehat{\Gamma}^T)_{i,j} = \sum_{1\le k\le p}\gamma_i^{(k)}\gamma_j^{(k)}$. Consequently, these coefficients define the orthogonal projection of the data and account for the similarities between columns in $\widehat{X}$, i.e., for $rank(\widehat{X})$.

Conclusions. With the above considerations in mind, we show in the following that the length of each DA and the angles between them reflect specific properties of the projection and of the projected data $\widehat{X}$. The mixing matrix $\widehat{\Gamma}\widehat{\Gamma}^T$ holds normalized contributions to $\widehat{S}$ and relates to the DA's arrangement in the sense that $(\widehat{\Gamma}\widehat{\Gamma}^T)_{i,j} = \sum_{1\le k\le p} S_{i,j}^{(k)}/\lambda_k = \widetilde{S_{i,j}}$, whereas $\widetilde{S_{i,j}} = \cos\angle(a_i, a_j)\ ||a_i||_2\ ||a_j||_2$. We can draw the following conclusions:

1. The length of DAs equals the standard deviation of the respective dimension in $\widehat{X}$, normalized for each contribution $\widehat{S}^{(k)}$ by its variance λ_k.

$$||a_i||_2 \overset{(2)}{=} \sqrt{\sum_{1\le k\le p}(\gamma_i^{(k)})^2} \overset{(3)}{=} \sqrt{\widetilde{S}_{i,i}} = \tilde{s}_i$$

2. The cosine of the angle between two DAs equals the correlation of the respective dimensions in $\widehat{X}$, where both covariance and standard deviation are normalized for each contribution $\widehat{S}^{(k)}$ by its variance λ_k.

$$\begin{aligned} \cos\angle(a_i, a_j) &= \frac{a_i^T a_j}{||a_i||_2||a_j||_2} \\ &\overset{(2)}{=} \frac{\sum_{1\le k\le p}\gamma_j^{(k)}\gamma_i^{(k)}}{\tilde{s}_i\ \tilde{s}_j} \\ &\overset{(3)}{=} \frac{\widetilde{S}_{i,j}^{(k)}}{\tilde{s}_i\ \tilde{s}_j} = \tilde{r}_{i,j} \end{aligned}$$

Figure 3 shows an SDT as an example for DA visualization highlighting these means for visualization.

Implications. It is important to emphasize that $\widehat{X}$ does not represent the whole data X but only its best rank-p approximation. That is, $\widehat{X}$ is the approximation of X that can be optimally depicted in p dimensions with regard to its variance. Therefore, $\widehat{X}$ is the orthogonally projected data on the subspace $\mathbb{R}^p$ which is spanned in a way that the projection reflects the dominant trends in X. However, $\mathbb{R}^p$ can only cover the most important information in the data. While other subspaces that are left out globally account for less variance in the data, relations therein may still be of importance for the user. Unfortunately, this information cannot be captured in a single projection and, consequently, parts of the relations between the original data dimensions in $\mathbb{R}^m$ are lost. The user has to be aware of this dilemma because it may lead to

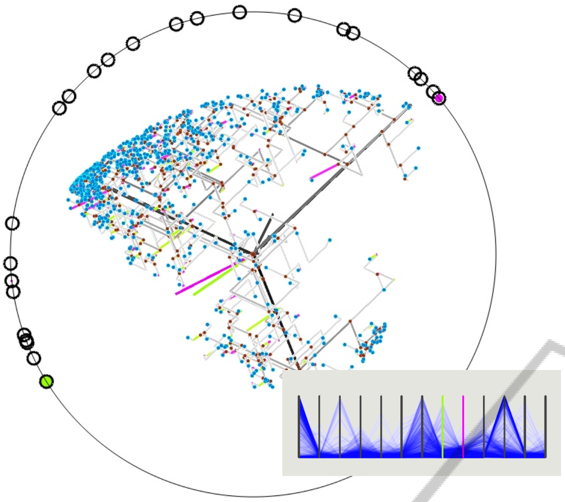


Figure 3: A 13-dimensional air quality data set visualized by the SDT approach: Dimensions exhibiting a large variance are represented by long DAs. Correlations within the data are indicated by DAs that are placed close to each other (colored axes). The variance points associated to these dimensions indicate that these correlations hold for the whole high-dimensional data space (parallel coordinates plot provided for illustration purposes only).

possible misinterpretations stemming from the visual assessment of the DAs' properties.

Because principal components are mutually orthogonal, it is possible that the depicted standard deviation of certain dimensions is lower in the initial projection than in other projections (globally ”less optimal” combinations of principal components). This depends on the overall information content of this dimension in the subspaces collapsed by dimensionality reduction. Thus, the knowledge derived from the DAs can only be a subset of the hidden information and usually represents a high-level view only. To avoid misinterpretation, they must be further evaluated.

Indicators for that information loss, e.g., the principal components, are often not part of the projection. The quality of the approximation $\widehat{X}$, with regard to one dimension, is directly reflected by the amount of lost variance in this dimension. A single projection cannot convey this information. To overcome this drawback, an SDT display provides variance points for each dimension. The number of variance points associated to a DA is equal to the dimensionality of the data set. Each variance point consists of two circles. The outer circle's radius equals s_i, the dimension's standard variation. The inner circle equals $\widehat{s}_i$, the part of the dimension's standard variation that is shown in the projection. The ratio between both circles reflects the actual variance of this dimension in the projection ($\widehat{X}$) in relation to the dimensions' total variance, i.e., $(\widehat{s}_i/s_i)^2$. and thus, allows one to infer knowledge about the importance of the data dimension. Variance points also provide guidance for interactive exploration. As their positions along the unit circle hold the position of this DA in other orthogonal projections, they can be found and explored quickly. The position of the i'th variance point thereby reflects the position of the corresponding DA in the orthogonal projection defined by $\widehat{\Gamma} = (\gamma^{(1)}, \gamma^{(i)})$, the combination of the first and i'th principal component. As shown in Figure 3 and 4, a wide range of the variance is usually captured by a few large variance points strongly reducing the data space that must be explored.

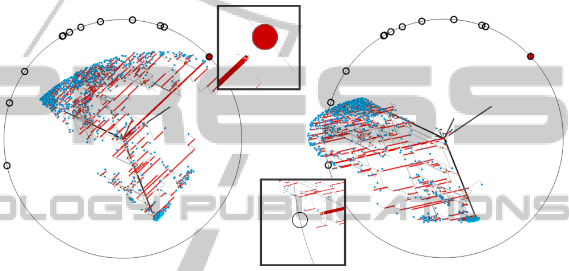


Figure 4: Variance points help to find other orthogonal projections of the data. Large variance points (left) indicate projections most suited to convey the variance in the data. Opposite variance points (right), even if not accounting for much variance, often lead to strongly different projections helping to identify unexpected data properties.

Commonly, the user is aware of the fact that projections have an inherent information loss. Projections that map different points in $\mathbb{R}^m$ to the same location in $\mathbb{R}^p$ make this fact clear. This ambiguity is often a severe problem and also stems from the principal illustration of “lost” subspaces. These points differ in those subspaces that are disregarded by dimensionality reduction and are therefore projected onto the same location. By visualizing the projection path of each data point, SDTs prevent possible misinterpretations by assuring the user that data points are only equal when they share the same path. The display of an SDT, however, requires additional processing power and introduces further graphical primitives into the data representation. This leads to occlusion problems, clutter, and long processing times when large data sets are to be displayed. How to solve these issues by proper interactive exploration and means for a scalable processing and representation is discussed in the following sections.

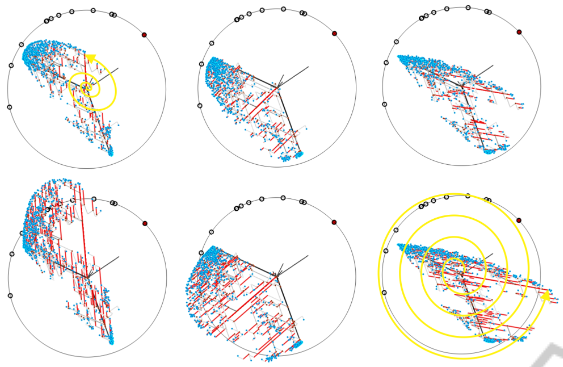


Figure 5: Moving a DA in circles activates the motion parallax effect of the human visual system letting the tree and the data points appear more ”plastic”. By providing many different coordinated projections, potential point clusters can be identified or verified. Best results are obtained by using a DA corresponding to a dimension with high variance.

4 INTERACTION FOR VISUAL CLUSTER ANALYSIS

Unconsidered subspaces, visual ambiguities, and occlusion issues within the initial representation can be addressed via interactive data exploration. This section introduces an exploration strategy for interactive visual cluster analysis as a common representative for the various goals in high-dimensional data visualization. The strategy was developed to cope with the large and complex data sets resulting from mass spectrography in air quality research (Bein et al., 2009). We follow the information visualization mantra (Shneiderman, 1996) starting with providing an *overview* to the data, then *filtering* data that are of minor interest, and eventually applying a drill-down step to uncover interesting *details*. As this involves many common exploration tasks, the given statements are broadly applicable to a variety of use cases.

A first **overview** about the data and its properties is provided by the initial projection of the data as described in the previous section. Subspaces hidden by dimension reduction, however, can contain further information important for the analyst. They are made available by a successive exploration of individual dimensions via their respective variance points. To achieve this, DAs are moved to other associated variance points leading to a different but still orthogonal projection of the data. To explore most important information first, it is meaningful to use large variance points indicating a strong inherent information content. We also propose to use variance points placed at opposite positions on the unit circle. Although position has no meaning regarding the amount of information content, this leads to strong changes in the projection and may reveal unexpected and important insight (see Figure 4). Switching between close points does not significantly change the projection and can usually be skipped even for large variance points.

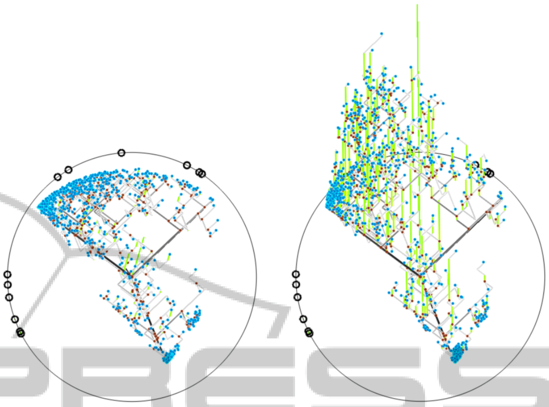


Figure 6: Cluster identification and verification by a stretching of a DA: While clutter hinders the detection of point clusters in the initial projection (left), this interaction reveals individual clusters and even sub-clusters in the data (right). Modification of a DA also allows the users to reveal the contribution and influence of the associated dimension to an examined cluster. The two clusters at the bottom of the representations are not affected by the interaction.

Even when subspace exploration is facilitated by variance points, usually there are still visual ambiguities resulting from dimension reduction and an overplotting of the points and the SDT. To overcome this, we introduce a novel exploration technique based on the motion parallax effect of the human visual system. By selecting and moving an appropriate DA, this effect creates a pseudo three-dimensional impression of the two-dimensional representation (see Figure 5) letting the points and the tree appear more ”plastic”. During this interaction, point clusters can be identified by their constant grouping. In our experiments, we revealed that continuously moving the DA in circles with varying diameter was particularly helpful to emphasize point clustering. By moving in circles, projections revealing point relations are displayed multiple times helping the human visual system to memorize this insight. Continuously changing diameter stretches or compresses potential clusters allowing for improved identification or verification. Not every dimension is equally suited to achieve this. We propose to select a dimension that does strongly contribute to higher tree branches, e.g., one that has a high variance in data values. As the movement strongly affects the top of the SDT leaving its stem nearly unchanged, it can increase motion parallax. Appropriate dimensions can easily be found by

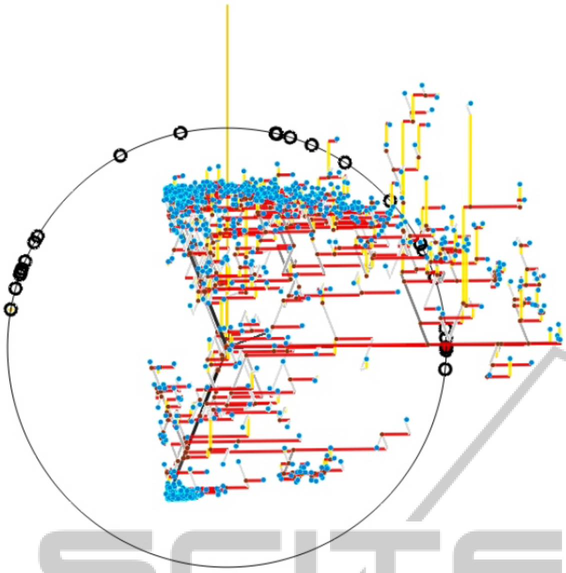


Figure 7: The orthogonal placement of two DAs (red and yellow color) in the presentation can simplify the evaluation of correlations between two-dimensions. Small contributions of points near the origin and large contributions at the top/right corner for both dimensions indicate a linear correlation for this pair of dimensions.

dimension highlighting emphasizing all line segments corresponding to the coordinates of a dimension and thus conveying their distribution. Sometimes only the orientation of the tree or of large branches is to be changed, e.g., to overcome visibility and occlusion issues. To support this, we propose to find and relocate a dimension with strong contribution to the stem of the SDT, e.g., a dimension with low variance. This leaves the initial crone structure of the SDT widely unaltered for further analysis.

Besides the described circle movement, further insight into the structure of the data can be gained by modifying the lengths of DAs only. Dimensions causing a visual separation of data points are most likely to contribute to clustering. Enlarging the respective DA separates data points and can help identifying clusters or verifying assumptions (see Figure 6). All points of a potential cluster show a similar behavior during length changes. Path highlighting can be used for further verification. In case of a valid m-D cluster, all associated points must share the same projection path. Length modification is also particularly useful to visually emphasize value contributions in the tree. Large contributions can easily be identified by their strong response to length changes.

Potential correlations in the initial projection are indicated by DAs that are placed close together. As shown in the previous section, this is only true for the current rank-p approximation and must not hold for the whole data space. Correlations can be verified by an individual examination of the variance points for each involved dimension. If a dimension has only a single large and many small variance points, it can be concluded that most of its information is encoded in the current position of the associated DA. When true for all involved dimensions, they are correlated not only in the current presentation but also in the data space. In the contrary, when all variance points of a dimension are of same size, the variance is scattered and the current anchor position represents just a subset of this information. In this case, it cannot be concluded that the depicted correlation holds in data space and further examination is required. One means to achieve this is to limit further exploration to the involved dimensions using filtering.

In case correlation has to be evaluated for two dimensions only, they may also be placed orthogonally to each other. Their visual emphasis by dimension highlighting could then reveal typical patterns in the corresponding SDT branches (see Figure 7).

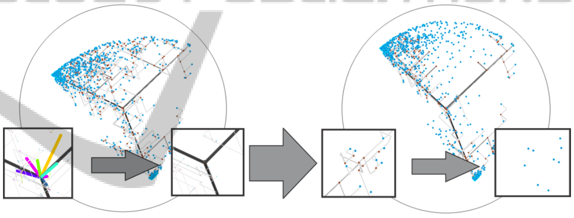


Figure 8: Interactive complexity and clutter reduction taking advantage of the capabilities of SDTs: dimension filtering (left) reduces the number of branch segments in the tree, node collapsing (right) the number of displayed subtrees.

Once an overview and first insights have been obtained, it is meaningful to **filter** out less interesting dimensions or subtrees to reduce clutter. Due to the fact that the visual contribution of a dimension is determined by the length of the corresponding DA, dimensions can be fully removed by placing their anchors at the center of the projection (see Figure 8, left). Appropriate candidates are dimensions that are correlated or show similar characteristics. They can be substituted by a single super-DA, whereby its angle is determined by the average and the length by the sum of all associated DAs. This changes the point projections only slightly. We further propose to remove dimensions having (1) very small variance points or (2) many, very small branches of similar length at high tree levels indicating little structure in the data. Less interesting subtrees can be removed from the visualization by node collapsing. The subtree is then represented by a single point (see Figure 8, right).

Once the data representation has been ”cleaned”

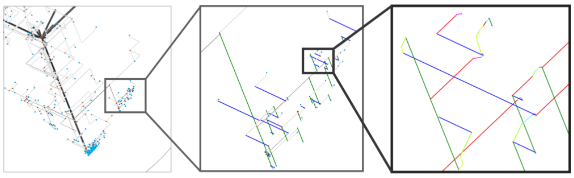


Figure 9: Means for zoom and pan interactions allow the users to drill-down into the presentation and data to obtain momentary detail.

be removing less interesting dimensions and data points, there is space for analysis on a more granular level. **Details** to identified clusters, such as their properties and potential sub-clusters, can be obtained by panning and zooming into the representation (see Figure 9).

For meaningful interaction, data display in real-time is mandatory. How to speed-up data processing and to improve the representation in case of many data points and dimensions is discussed in the next section.

5 SCALABLE PROCESSING AND REPRESENTATION

Large data sets with a high level of complexity lead to long processing times and visual clutter. This decreases the usability and general applicability and usefulness of a technology. These issues can be overcome by *scalable processing* and *representation* of the data discussed in this section.

5.1 Scalable Processing

5.1.1 Complexity

The processes required to compute an initial SDT representation are: hierarchy-building, configuring axes, and rendering. Complexity mainly depends on the number of data points, N, and dimensions, D.

Most of the processing power required to calculate and provide SDTs is used to establish the data hierarchy and the initial axis layout. Common hierarchical clustering techniques are quadratic in complexity with respect to the number of data points, because they require comparing $\binom{N}{2}$, or $O(N^2)$ point pairs; naive methods can be as complex as $O(N^3)$ because they perform each set of $O(N^2)$ comparisons every time two clusters are aggregated. Performing these comparisons usually involves computing a difference along each dimension, adding a complexity of $O(D)$.

The processing power needed to provide an appropriate initial data projection strongly depends on the applied approach. The PCA-based method used for SDTs has a complexity of $O(N^2)$ (Engel et al., 2011).

Rendering the SDT is the least expensive computation. Given the fact that the tree hierarchy consists of $2N$ or $O(N)$ nodes and rendering each node is linear in the number of dimensions, or $O(D)$, the rendering takes $O(ND)$ time.

Summarizing, the complexity of the processes required to compute an initial SDT representation is

$$O(N^2D) + O(N^2) + O(ND) = O(N^2D) \qquad (4)$$

5.1.2 Computation

In order to reduce the computation time of the initial SDT and axis layout, we propose N as well as D. Based on equation (4), reduction of N has the greatest impact on the complexity of the corresponding processes. This especially applies to the initial hierarchical clustering and the calculation of the projection.

Many different methods to meaningfully reduce the number of data points have been proposed (Ellis and Dix, 2007). A simple and often used strategy that can also be applied to SDTs is uniform sub-sampling (see Figure 10). More sophisticated methods can be used depending on known characteristics of the data, the purpose of the visualization, or current user demands. SDTs do not have any special requirements for the selection of an appropriate point reduction strategy.

While point reduction should have the greatest impact on initial computation times, data sets of high dimensionality can require significant computation and be difficult to visualize effectively. Similar to point reduction, numerous methods to meaningfully decrease D have been proposed. They range from simple methods, such as PCA, to more sophisticated strategies, such as the use of an importance hierarchy based on the dimensions (Yang et al., 2003c), to manual methods in case automated technology fails. In any case,

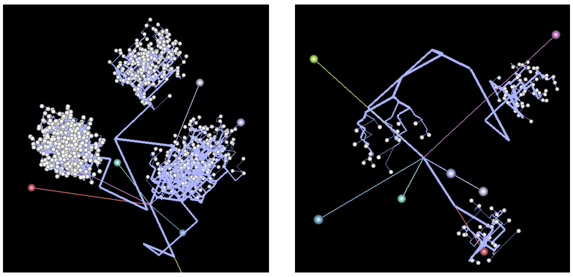


Figure 10: Example of an SDT for all data (left, 1000 data points) and a uniformly sub-sampled subset (right, 100 data points). Emphasized properties of the data are well conveyed in both representations. The large colored dots indicate end points of DAs.

such approaches reduce computation time proportionally to the fraction of dimensions eliminated. As long as the applied method does not remove or cover data properties the analyst might be interested in, SDTs do not demand any special dimension reduction strategy. The initial SDT layout before and after reduction, however, is usually different.

Point and dimension filtering also reduce the computation power needed to render and interact with the visualization in real time. Since rendering the SDT is linear with respect to both *N* and *D*, point reduction is not as superior a method with respect to reducing computation time as it is with respect to the initial computations. However, the fact that users can expect to have far more data points than dimensions underpins the significance of prioritizing point over dimension reduction. In order to justify the theoretical assumptions stated we conducted a number of experiments to assess the complexity of SDTs with regard to a changing number of *N* and *D*.

In the experiments we used a prototypical implementation of SDTs as described in (Engel et al., 2011) and a 255-dimensional air quality data set consisting of 70,000 points provided from our collaborators from the UC Davis Air Quality Research Center. For complexity reduction we applied uniform point subsampling and PCA. Figure 11 shows the initial SDT computation time for various subsets of the data with either 13, 50, and all 255 dimensions. As shown, the computation time is a non-linear polynomial function of *N*, and a linear function of *D*. The differences between the original and reduced data sets are significant. Specifically, computing the SDT for 1,000 points takes approximately 80% less time for a 50-dimensional subset of the data than for all 255 dimensions. Computations involving all 255 dimensions were 100 times faster for 200 points instead of 1,000 points, showing the tremendous growth in complexity as a function of *N*. This demonstrates the significance of point reduction. Clustering and initial projection, however, are computed in a pre-processing step and may be stored for multiple uses in cases where point reduction is not applicable or meaningful.

5.2 Scalable Representation

A major problem of all projection-based visualization techniques is the low-dimensionality of the presentation space. With too many dimensions, projection-based visualizations can become difficult to understand; SDTs in particular can become cluttered and busy, especially by rendering DAs. However, since SDTs aggregate common dimensional components of clusters into single branches, they are visually organized and can be easily interpretable for data sets with upwards of twenty dimensions. Depending on various properties of the data, such as the presence of distinct clusters, even higher-dimensional data sets containing thirty or even forty dimensions can be clearly interpretable. Beyond fifty dimensions, however, the sole presence of the shown DAs makes the display overcrowded and difficult to understand.

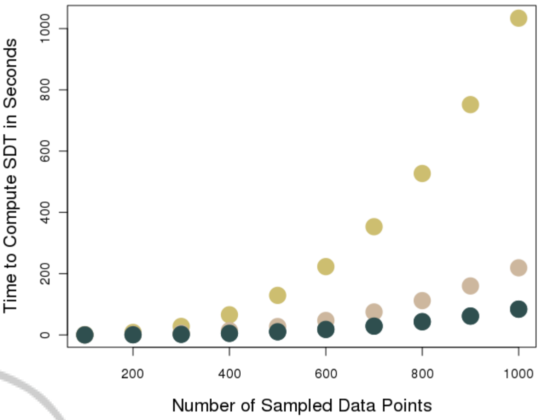
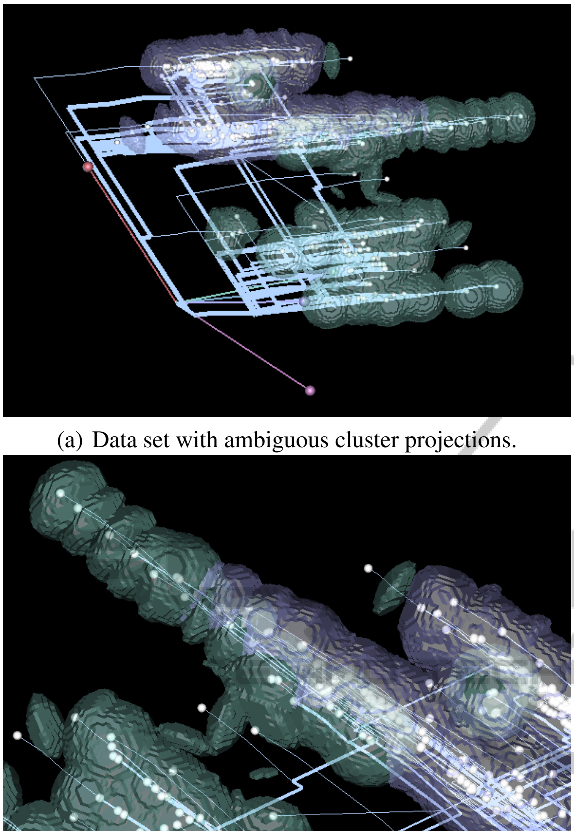


Figure 11: Initial SDT computation time as a function of *N* and *D*. The gold points represent 255-dimensional subsets of the data, the bisque points the 50-dimensional subsets, and the dark points the 13-dimensional subsets. When *N* is small, the computation time is almost identical for all examined subsets. The timing results reflect well the predicted theoretical behavior.

Another common issue of projection-based visualization techniques is that projecting data into a space with fewer dimensions can lead to a close positioning of data points that are much more distant in high-dimensional space. SDT's attempt to overcome this problem by means of a tree structure, but illustrating this structure might not remove all ambiguities in the representation for very large data sets. In order to overcome this, we propose to highlight and differentiate data points below certain tree levels. In order to achieve this, we enclose the data points and tree branches below a certain level in uniquely colored isosurfaces covering the associated points; the SDT itself remains unchanged. Any isosurface-rendering algorithm, such as the well-known Marching Cubes (MC) algorithm (Lorensen and Cline, 1987), is suitable for this task. As shown in Figure 12, this highlights the association and distribution of points belonging to the same cluster even when points associated with different clusters are located nearby or even at identical positions in presentation space.

When data sets are dense and many points occlude one another, the relative densities of clusters are ambiguous to the user. Isosurfaces can also be used to overcome this problem by conveying cluster den-



(a) Data set with ambiguous cluster projections.

(b) Closeup of ambiguous clusters.

Figure 12: Disambiguating point positions in SDTs: The green and purple isosurfaces correspond to two distinct subtrees of data. Their intersection indicates that these branches represent data points that reside at different locations in the high-dimensional data space but are positioned near each other in the presentation space.

sity. To achieve this, isosurfaces are rendered around sub-trees of the SDT in regions of high point density, quickly highlighting sub-trees that might require further analysis. Some isosurface rendering algorithms, including the MC algorithm, can directly or indirectly rely on point density in display space as a form of input and only draw isosurfaces where relevant. Isosurfaces can also be color- or opacity-coded to provide the users with more specific information on cluster density in a given region (see Figure 13).

Isosurfaces are also a valid means to represent data points that have been filtered by one of the approaches described above or by interactive folding. Instead of the associated data points, only corresponding isosurface are shown. To visually de-emphasize such sub-clusters, however, they should be uniquely colored and highly transparent.

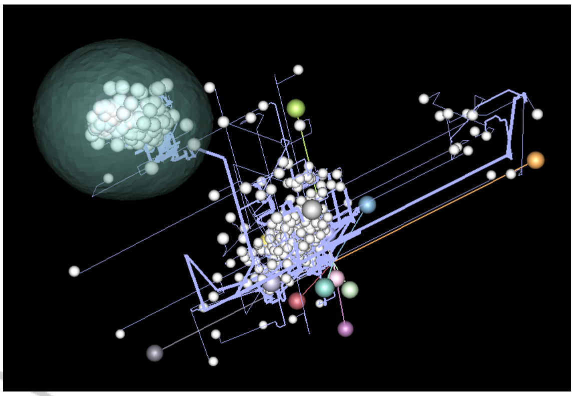


Figure 13: Disambiguating cluster density in SDTs: The front-most cluster in this visualization is significantly denser than the other clusters shown. This is conveyed by the green isosurface.

6 CONCLUSIONS

SDTs are a valid means to visualize and explore high-dimensional data. However, several questions important for a broad adoption still remain to be answered. Our paper addressed several of these questions. We were particularly interested in the insight that can be gained from an interpretation of the initial projection of the data. We delivered proof that the length and relation of DAs allow one to draw meaningful conclusions about the information content of a single and correlations between multiple dimensions of the data. We also provided means and guidelines for their interactive exploration in visual cluster analysis. In order to deal with large data sets, we proposed and empirically justified options for scalable processing and representation. All addressed problems and the introduced solutions showed that SDTs can be successfully used in a variety of application domains to cope with the challenging problem of high-dimensional data analysis, visualization, and interactive exploration.

The general abilities of SDTs have not yet been evaluated empirically. Future work can be directed at the design and implementation of a user study comparing SDTs to value visualizations, such as the parallel coordinates plot, and relation visualizations, such as PCA. When based on common real-world scenarios, this evaluation can also justify the usefulness of SDTs for a broad variety of application domains. The results we already obtained from our experiments and user feedback are promising and provide evidence for the true value of the novel approaches presented here.

REFERENCES

Ankerst, M., Berchtold, S., and Keim, D. A. (1998). Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *INFOVIS '98: Proceedings of the 1998 IEEE Symposium on Information Visualization*, pages 52–60, Washington, DC, USA. IEEE Computer Society.

Artero, A. O., de Oliveira, M. C. F., and Levkowitz, H. (2004). Uncovering clusters in crowded parallel coordinates visualizations. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 81–88, Washington, DC, USA. IEEE Computer Society.

Bein, K., Zhao, Y., and Wexler, A. (2009). Conditional sampling for source-oriented toxicological studies using a single particle mass spectrometer. *Environmental Science and Technology*, 43(24):9445–9452.

Ellis, G. and Dix, A. (2007). A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1216–1223.

Elmqvist, N., Dragicevic, P., and Fekete, J.-D. (2008). Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14:1141–1148.

Engel, D., Rosenbaum, R., Hamann, B., and Hagen, H. (2011). Structural decomposition trees. *Computer Graphics Forum*, 30(3):921–930.

Hauser, H., Ledermann, F., and Doleisch, H. (2002). Angular brushing of extended parallel coordinates. In *INFOVIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, pages 127–130, Washington, DC, USA. IEEE Computer Society.

Hoffman, P., Grinstein, G., and Pinkney, D. (1999). Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In *NPIVM '99: Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation*, pages 9–16, New York, NY, USA. ACM.

Ingram, S., Munzner, T., and Olano, M. (2009). Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics*, 15:249–261.

Jänicke, H., Böttinger, M., and Scheuermann, G. (2008). Brushing of attribute clouds for the visualization of multivariate data. *IEEE Transactions on Visualization and Computer Graphics*, 14:1459–1466.

Johansson, J., Ljung, P., Jern, M., and Cooper, M. (2005). Revealing structure within clustered parallel coordinates displays. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, pages 125–132, Washington, DC, USA. IEEE Computer Society.

Johansson, S. and Johansson, J. (2009). Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15:993–1000.

Kandogan, E. (2001). Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the ACM international conference on Knowledge discovery and data mining*, pages 107–116, New York, NY, USA. ACM.

Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. *Computer Graphics*, 21(4):163–169.

McDonnell, K. T. and Mueller, K. (2008). Illustrative parallel coordinates. *Computer Graphics Forum*, 27(3):1031–1038.

Oesterling, P., Heine, C., Jänicke, H., and Scheuermann, G. (2010). Visual analysis of high dimensional point clouds using topological landscapes. In *Pacific Visualization Symposium (PacificVis), 2010 IEEE*, pages 113 –120.

Paulovich, F. V., Oliveira, M. C. F., and Minghim, R. (2007). The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing*, pages 27–36, Washington, DC, USA. IEEE Computer Society.

Peng, W., Ward, M. O., and Rundensteiner, E. A. (2004). Clutter reduction in Multi-Dimensional data visualization using dimension reordering. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 89–96, Washington, DC, USA. IEEE Computer Society.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343.

Yang, J., Patro, A., Huang, S., Mehta, N., Ward, M. O., and Rundensteiner, E. A. (2004). Value and relation display for interactive exploration of high dimensional datasets. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 73–80, Washington, DC, USA. IEEE Computer Society.

Yang, J., Peng, W., Ward, M. O., and Rundensteiner, E. A. (2003a). Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proceedings of the Ninth annual IEEE conference on Information visualization*, pages 105–112, Washington, DC, USA. IEEE Computer Society.

Yang, J., Ward, M. O., Rundensteiner, E. A., and Huang, S. (2003b). Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the symposium on Data visualisation 2003*, VISSYM '03, pages 19–28, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

Yang, J., Ward, M. O., Rundensteiner, E. A., and Huang, S. (2003c). Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the Symposium on Data visualisation 2003*, VISSYM '03, pages 19–28, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

Yuan, X., Guo, P., Xiao, H., Zhou, H., and Qu, H. (2009). Scattering points in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 15:1001–1008.

Zhou, H., Yuan, X., Qu, H., Cui, W., and Chen, B. (2008). Visual Clustering in Parallel Coordinates. *Computer Graphics Forum*, 27(3):1047–1054.