Structural Decomposition Trees: Semantic and Practical Implications

Daniel Engel[1], Hans Hagen[1], Bernd Hamann[2], and René Rosenbaum[3]

[1]Department of Computer Science, University of Kaiserslautern, Germany,
[2]Institute for Data Analysis and Visualization (IDAV), Department of Computer Science,
University of California, Davis, USA,
[3]Institute for Computer Science, University of Rostock, Germany
{engel,hagen}@cs.uni-kl.de
hamann@cs.ucdavis.edu
rrosen@informatik.uni-rostock.de

Abstract. The visualization of high-dimensional data is a challenging research topic. Existing approaches can usually be assigned to either relation or value visualizations. Merging approaches from both classes into a single integrated strategy, Structural Decomposition Trees (SDTs) represent a completely novel visualization approach for high-dimensional data. Although this method is new and promising, statements on how to use and apply the technique in the context of real-world applications are still missing. This paper discusses how SDTs can be interpreted and interacted with to gain insights about the data more effectively. First, it is discussed what properties about the data can be obtained by an interpretation of the initial projection. These statements are also valid for other projections based on principal components analysis, addressing a frequent problem when applying this technique. Further, a detailed and task-oriented interaction guideline shows how provided interaction methods can be utilized effectively for data exploration. The results obtained by an application of these guidelines in air quality research indicate that much insight can be gained even for large and complex data sets. This justifies and further motivates the usefulness and wide applicability of SDTs as a novel visualization approach for high-dimensional data.

Keywords: High-Dimensional Data Visualization, Projections, Interaction

1 Introduction

The visualization of high-dimensional data is a common but still unsolved problem. Structural decomposition trees (SDTs) [1] 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 related visualization technology. Research concerning SDTs, however, is constrained to the introduction and description of fundamental aspects of this novel displaying approach only. Although, research concerned with main utilization strategies have recently been published [2], the eligibility of SDTs when extensively applied in complex real-world visual data analysis has not been discussed in literature so far.

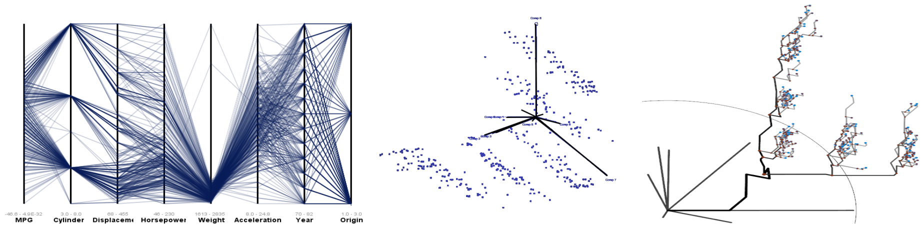

Fig. 1: The two classes of visualizations for high-dimensional data (value (left) and relation (center) visualizations) are brought together by SDTs (right). All visualizations represent the well-known "cars" data set. The SDT highlights the five distinct clusters by its branch structure also conveying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

This paper provides guidance and an example 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 provide practical implications supporting the users in their understanding of high-dimensional data projections. Classifying and discussing means for interaction provided by SDTs from a functional point of view, the second part (Section 4) is concerned with appropriate data exploration. For each listed interaction, distinct aims and guidelines for its appropriate application are stated. The third part (Section 5), discusses results we obtained from an application of SDTs in the visual analysis of air quality data. It shows that by taking advantage of the introduced methods much insight can be gained even for complex and large data sets. We conclude (Section 6) that SDTs are a valid means for visualizing and gaining insight into high-dimensional data, but must be understood and applied in the right way in order to avoid misinterpretation and wrong conclusions. Here, we 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 are of strong interest to the visualization community. Many different approaches and techniques have been proposed. According to [1], they can be categorized into *value* or *relation visualizations*. By focusing 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, 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 [3–5], appropriate means for interaction [6, 7], and better utilization of the available screen space [8] have been proposed. Clutter reduction is also achieved by dimension ordering arranging the dimensions within the visual representation based on correlations within the data. After the initial formal problem statement [9], this technique has been expanded in [10] and [11]. 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 data values, **relation visualizations** are designed to convey data relationships. 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 well-known 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, center). 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 [12] projecting and displaying the basis vectors along the data points (see Figure 1, center). These DAs are also an appropriate means to adjust the projection interactively [13]. 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 [14], a better incorporation of domain-appropriate analysis techniques, e.g., brushing and filtering [15], or computational speed gains [16].

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 [17]. Few publications 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 [10, 18, 19], [20], or [21]. 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, right). This is 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. The resulting tree has minimal overall branch length, reducing the redundancies considerably. Appropriate representation of the individual dimension contributions

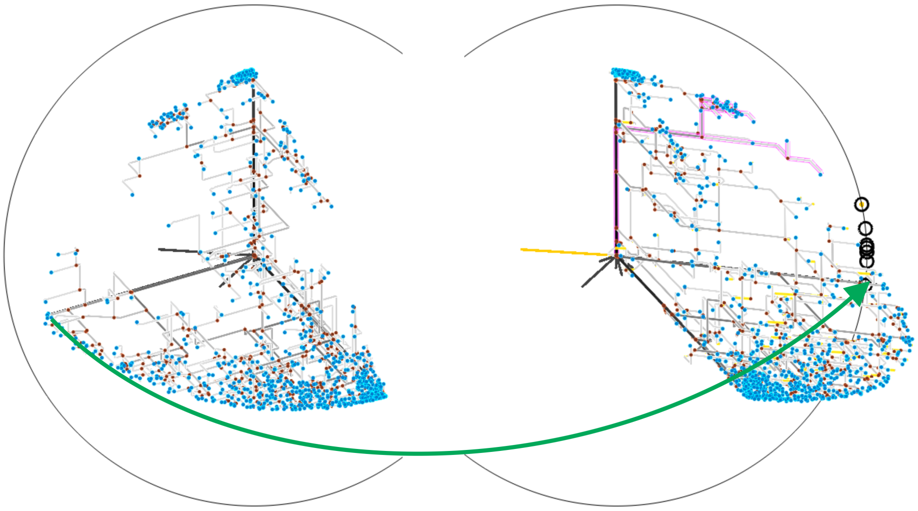

Fig. 2: The main interactions provided by SDTs: *Repositioning of DAs* (green arrow) allows for an intuitive adjustment of the projection. *Dimension highlighting* (yellow DA) conveys the individual contributions of a dimension in the data (yellow tree segments). *Path highlighting* (purple tree branch)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

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, right). The initial SDT projection maximizes the space between tree paths and allows for a better interpretation of the visualization [1].

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. Different means to emphasize and filter dimensions and line segments are provided to facilitate investigation and interpretation of the structural decomposition of the data. Published research concerned with SDTs mainly focuses on its technical foundations. Although, semantic aspects of an SDT representation were discussed in [2], the authors were mainly concerned with the application of SDTs in visual cluster analysis. General statements to practical implications, guidelines, and a concrete use case were not provided.

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 so that $X \in \mathbb{R}^{(n\times m)}$. The projection of X to $\widetilde{X} \in \mathbb{R}^{(n\times p)}$ is defined by the linear mapping of m-D data points X_i to p-D display points $\widetilde{X}_i$, for $1 \leq i \leq n$, by the linear combination of DAs $a_j \in \mathbb{R}^p$ with the corresponding coordinate $X_{i,j}$, for $1 \leq j \leq m$:

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

This technique, the mapping in star coordinates [13], 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 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 [1], 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 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 $\widehat{\Gamma}$:

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

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

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

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 constructed from X by the linear combination of all $X_{\bullet,j}$ with coefficients $(\widehat{\Gamma}\widehat{\Gamma}^T)_{i,j} = \sum_{1\leq k\leq p} \gamma_i^{(k)} \gamma_j^{(k)}$. 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 $(\widehat{\Gamma}\widehat{\Gamma}^T)_{i,j} = \sum_{1\leq k\leq p} S_{i,j}^{(k)}/\lambda_k = \widetilde{S_{i,j}}$, 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.

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

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.

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

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 issue because it may lead to 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. 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. The quality of the projection, with regard to one dimension, is reflected by the amount of its lost variance due to dimension reduction. To indicate this information, an SDT display provides *variance points* for each dimension. Each variance point consists of two circles. While the outer circle's radius represents s_i, the dimension's standard variation, the inner circle represents $\widehat{s_i}$, the part of the dimension's standard variation that is reflected by the projection. Assessing the ratio between both circles, $(\widehat{s_i}/s_i)^2$, thereby allows one to infer the quality of the projection with regard to a data dimension. Thereby, variance points provide guidance for interactive exploration.

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. Ambiguity is often a severe problem and stems from the principal illustration of "collapsed" subspaces. Points that only differ in the subspaces that are disregarded by dimensionality reduction are consequently 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. This display, however, introduces further graphical primitives into the data representation, leading to occlusion problems and visual clutter. How to solve these issues by proper interactive exploration is discussed in the following sections.

4 Means For Interaction: Purpose and Guidelines

Interactions within SDTs can be mainly classified as interactions with the *data* or the *dimensional anchors* as well as *changes of the view*. In this section, we describe the available interaction methods from a general, functional standpoint, state their individual aims, and complete with novel guidelines on how to interact with SDTs. This will provide users with quick insight and reference to the available methods. Using these guidelines, SDTs convey an intuitive visual mapping that can be remembered and from which the user can quickly learn

- how the data is assembled, spread, where clusters are, or which pattern they follow,
- how parts of the data are connected, differ, or how they relate to each other, and
- what properties they have, e.g., intra-cluster variances, shape, or alignment.

4.1 Interactions with the Data

This class of interactions allows the user to highlight or filter parts of the data. Associated techniques are usually strongly task and application depend.

Interaction: **Dimension highlighting**
Aim: Emphasizing dimension contributions of the data
Guidelines: This interaction (see Figure 2) allows the user to emphasize all line segments corresponding to the coordinates of a dimension and thus helps to investigate the

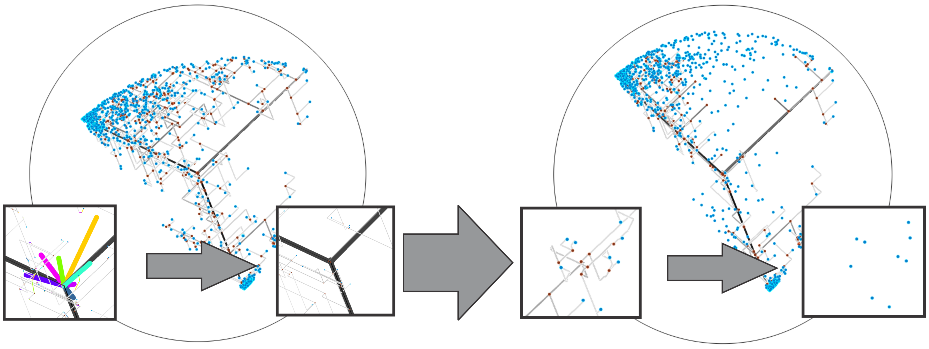

Fig. 3: Interactive complexity and clutter reduction taking advantage of the capabilities of SDTs: dimension filtering (left) reduces the number of branch segments in the tree, node collapsing (right) the number of displayed subtrees. Additional means for zoom and pan interaction allow the users to drill-down into the presentation and data to obtain momentary detail (bottom).

structural decomposition of the data. The selection of too many dimensions decreases its usefulness. Only DAs of current importance to the user should be selected.

Interaction: **Path highlighting**
Aim: Emphasizing data points and subtrees
Guidelines: Path highlighting (see Figure 2) emphasizes interesting pathes and branches within the SDT. During selection the user should focus on paths that lead through cluttered regions as they might no be easily followed and take unexpected ways.

Interaction: **Node collapse**
Aim: Data filtering
Guidelines: This interaction causes subtrees and data points to disappear from the SDT representation. A single subtree is then represented by a characteristic point only (see Figure 3, right). 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. Most appropriate regions to apply node collapse are cluttered areas or uninteresting subtrees. The user, however, should always bear in mind that data filtering was applied.

4.2 Interactions with the Dimensional Anchors

The layout of an SDT visualization consists of the different DAs. As their alignment strongly influences the projection of the data, allowing for their interactive modification is a powerful means for a variety of purposes. As there is no restriction on their placement, interactions can change the (1) *angle* or (2) *length* of an DA, or (3) *both*. Each kind of modification can be used to achieve a distinct aim.

Interaction: **Move of a DA to a corresponding variance point**

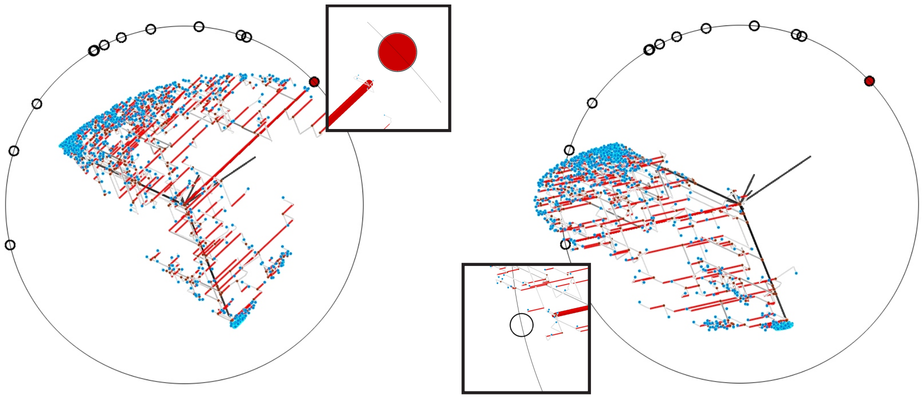

Fig. 4: Variance points help to find other promising projections of the data. Large variance points (left) indicate projections most suited to convey the variance in the data. Opposite variance points (right), even when not accounting for much variance, often lead to strongly different projections helping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

Aim: Exploration of hidden subspaces
Guidelines: Subspaces hidden by dimension reduction can contain further information important for the analyst. They are made available by a successive exploration of individual dimensions via their respective variance points. This leads to different but still orthogonal projections 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.

Interaction: **Move of the SDT stem to another position**
Aim: Solving occlusion issues
Guidelines: 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.

Interaction: **Orthogonal placement of two DAs**
Aim: Discovery or verification of correlation between two dimensions
Guidelines: The orthogonal placement of two DAs emphasizes potential correlation between two dimensions and thus enables the viewer for its visual discovery or verification. Correlations can be identified by following the development of the point contributions from the origin along the direction of the respective dimension vectors. As an example, increasing contributions for both dimensions indicate a linear correlation

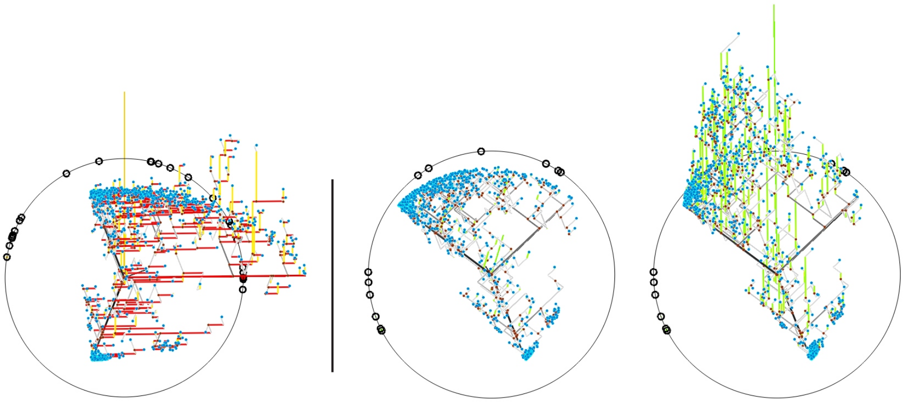

Fig. 5: The orthogonal placement of two DAs (red and yellow color) in the presentation can simplify the evaluation of correlations between the associated dimensions (left). In order to overcome potential point cluttering within the initial projection (center), a DA (green color) can be interactively stretched (right).

for the associated dimensions (see Figure 5, left). Visual emphasis of the involved dimension contributions by dimension highlighting helps revealing such characteristic patterns. Due to the fact that SDTs are projected into a two-dimensional presentation space, correlations between more than two dimensions must be explored successively.

Interaction: **Enlarging or shrinking a DA**
Aim: Exploration of data distribution, discovery of data clusters, conveyance of value contributions
Guidelines: As the length of a DA proportionally influences the position of the projected data, the associated points can be stretched or compressed easily (see Figure 5, center/right). This allows for an investigation of the data distribution of the associated dimension. Thereby, it is useful to enlarge and shrink the DA multiple times and in different directions to discover the representation where the distribution is conveyed best.
Dimensions causing a visual separation of data points usually contribute to clustering. Enlarging the length of a DA enhances separation and thus can help identifying such clusters. 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 emphasizing value contributions in the tree. Strong contributions can easily be identified by their strong response to length changes.

Interaction: **Move of a DA to the origin of the projection**
Aim: Dimension filtering
Guidelines: To reduce clutter, it is meaningful to filter out less interesting dimensions by placing their anchors at the origin of the projection (see Figure 3, left). Appropri-

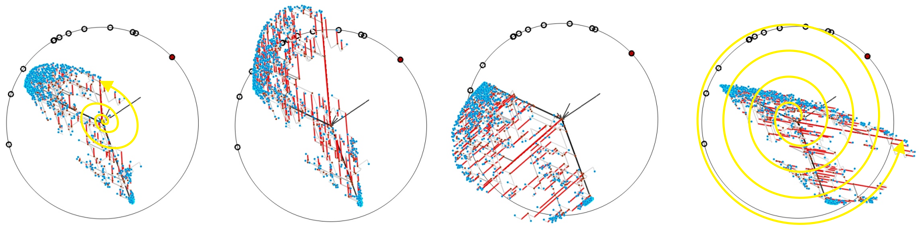

Fig. 6: Moving a DA in circles activates the motion parallax effect of the human visual system letting the tree and the data points appear more "plastic". By providing many different coordinated projections, characteristics of the data can be identified or verified. Best results are obtained by using a DA corresponding to a dimension with high variance.

ate 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, but removes many SDT branches from the representation. 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.

Interaction: **Continuous circular movement of a DA**
Aims: Discovery of m-D data clusters, exploration of data distribution
Guidelines: The movement of a DA enables the motion parallax effect of the human visual system to create a pseudo three-dimensional impression of the two-dimensional SDT representation (see Figure 6). This lets the points and the tree appear more "plastic" and results in more insight about the structure and potential clusters in the data. During the interaction, point clusters can be identified by their constant grouping. Circular movement leading to similar projections at each turn helps the human visual system to memorize the gained 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 strongly contributes to higher tree branches, e.g., one that has a high variance in data values. As such a dimension strongly affects the top of the SDT leaving its stem nearly unchanged, it can increase motion parallax. Appropriate dimensions can easily be found by dimension highlighting emphasizing all line segments corresponding to the coordinates of a dimension and thus conveying their distribution.

4.3 Interactions to Change the View

Interaction: **Zoom&Pan of the current viewing region**
Aim: Providing overview or detail
Guidelines: Changing view point and direction is a common means in interactive data exploration. Following the information visualization mantra [22] visual data analysis should successively repeat the stages: (1) providing an overview to the data, (2) filtering data that are of minor interest, and (3) drilling-down to uncover interesting details.

Overview and detail within this process can be obtained by panning and zooming into the representation (see Figure 3).

5 Results and Use Case

In the following, the guidelines and implications described in this paper are applied in the form of a case study on a real-world data set from the application of air quality research. Data has been provided by the Air Quality Research Center of UC Davis and obtained by single particle mass spectrometry [23]. We discuss the application of SDTs to wood stove source sampling data from Pittsburgh, Pennsylvania. The focus for this investigation lies in quantifying the relationship between isotopes ambient during biomass combustion. Biomass combustion emits copious amounts of gases and particles into the atmosphere and plays a key role in almost all present day environmental concerns including the health effects of air pollution, acid rain, visibility reduction, and stratospheric ozone depletion. The raw 256-dimensional data has undergone application-specific data transformations as well as dimension reduction to the dimensions most important for the investigation purposes of our collaborators: m/z 24 (C_2^+), 27 ($C_2H_3^+$), 36 (C_3^+), 39 ($^{39}K^+$ / $C_3H_3^+$), and 41 ($^{41}K^+$ / $C_3H_5^+$). The data are highly unstructured. Due to this characteristic, the SDT consists of a small stem and many small branches. The achieved representation of individual coordinate values, however, still allows for an accurate data investigation as shown by the following findings.

Figure 7 a), shows the initial projection for 1000 particles randomly selected from the sampling campaign. This first view clearly reveals two main clusters corresponding to m/z 39 and to a mixture of m/z 24, 27, 36, and 41, respectively. The *Dimensional Anchor arrangement* suggests a positive correlation between m/z 24 and 36 by their DAs' co-location, as well as low variance in m/z 41 by the DAs low length. Verification of both indicators based on *dimension highlighting* and *variance points* reveals that the variance in m/z 39 is only partially reflected in the projection. This can be seen in Figure 7 b), where a secondary placement of the DA is suggested at the bottom of the circle. Highlighting and *circular movement* of the DA, however, reveals that the overall variance is considerably lower than in other dimensions.

Further investigation of the large cluster on the ride side of the view reveals that two stems branching of in dimensions 27 ($C_2H_3^+$) and 39 ($^{39}K^+$ / $C_3H_3^+$) are the primary contributors to this cluster. Detailed investigation by *zooming and panning* reveals that these points indeed show mixtures of all dimensions, mainly residing in mid-value ranges and of similar intra-cluster variance. This is shown in Figure 7 c), where a parallel coordinates plot of these dimensions is included for reference. By using *dimension and branch filtering*, further insight is gained in the relationship between dimensions 24, 27 and 39. In Figure 7 d), DA 36 is moved at the center of the projection due to its correlation with these dimensions. In the figure, the selection of dimensions 24 and 39 reveals that low value contributions of $^{39}K^+$ / $C_3H_3^+$ are present in the majority of the cluster located along the DA of 24 (C_2^+), while in the clusters of 27 ($C_2H_3^+$) and 39 ($^{39}K^+$ / $C_3H_3^+$), these mixtures are filtered out. Successive investigation of the rela-

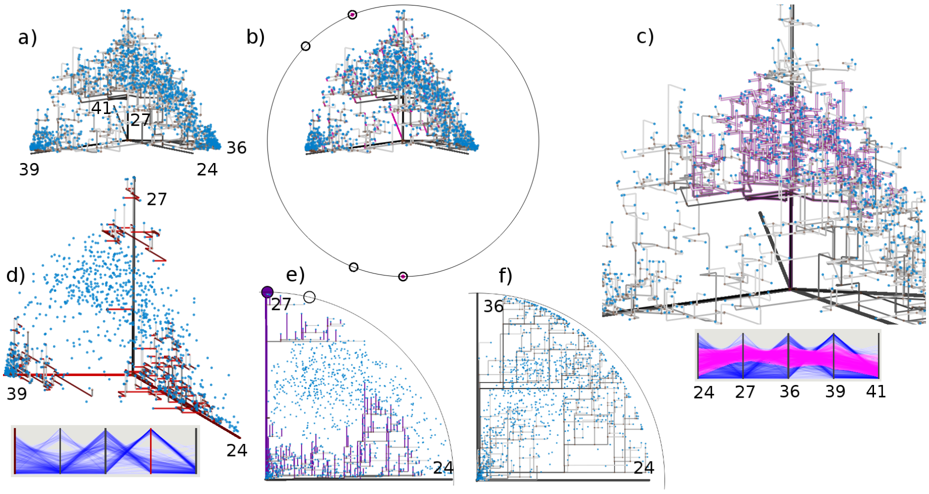

Fig. 7: SDT of 1000 mass spectra obtained from a source sampling campaign in Pittsburgh, Pennsylvania: **(a)** Initial projection of the dimensions relevant to analysis indicates correlation between 24 and 36, as well as low variance in 41. **(b)** Verification of 41's variance points conveys the information loss for this dimension due to dimension reduction. **(c)** Navigation and selection options allows to quickly identify the cluster residing in mid-value ranges of the dimensions. **(d)** Options for filtering allow to further adjust the view to current needs. **(e)** The display of the data's structure avoids misconceptions while **(f)** adjusting the projection for verification of assumptions.

tionship between 24 (C_2^+) and 27 ($C_2H_3^+$) is conducted by the *orthogonal placement of DAs*, as shown in Figure 7 e).

Next to its highly interactive capabilities, the SDT's strength lies in displaying the underlying structure of the data, thus, enhancing the projection by conveying proximities in high-dimensional space, as well as in projective space. While the point projection, as shown in Figure 7 e), does not display any relationship between the two dimensions, the SDT shows two main branches. After investigation, it is revealed that the upper and lower branch structures originate from samples showing values in 39 ($C_3H_3^+$) and 36 (C_3^+), respectively. The co-occurrence of $C_2H_3^+$ and $C_3H_3^+$, as well as C_2^+ and C_3^+ is in perfect agreement of the wood stove source sampling study [23], where correlations between C_x^+ and $C_xH_y^+$ isotopes have been verified based on manual data analysis. Figure 7 f) shows this correlation for C_2^+ and C_3^+.

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 addresses several of these questions. We were particularly interested in practical implications and insight that can be gained from an interpretation of the initial projection of the data. We showed 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 a functional view and novel guidelines for effective interaction with SDTs. To illustrate their meaningful appliance, we performed a case study on highly complex real-world data. The results demonstrate that SDTs can be successfully used in a variety of real-world application domains to cope with the challenging problem of high-dimensional data analysis, visualization, and interactive exploration.

Acknowledgments The authors gratefully acknowledge the support of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) for funding this research (grants #RO3755/1-1, #RO3755/2-1, and #1131).

References

1. Engel, D., Rosenbaum, R., Hamann, B., Hagen, H.: Structural decomposition trees. Computer Graphics Forum **30**(3) (June 2011) 921–930
2. Rosenbaum, R., Engel, D., Mouradian, J., Hagen, H., Hamann, B.: Interpretation, interaction, and scalability for structural decomposition trees. In: Proceedings of International Conference on Information Visualization Theory and Applications, Rome, Italy (February 2012)
3. Johansson, J., Ljung, P., Jern, M., Cooper, M.: Revealing structure within clustered parallel coordinates displays. In: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization, Washington, DC, USA, IEEE Computer Society (2005) 125–132
4. Zhou, H., Yuan, X., Qu, H., Cui, W., Chen, B.: Visual Clustering in Parallel Coordinates. Computer Graphics Forum **27**(3) (May 2008) 1047–1054
5. Artero, A.O., de Oliveira, M.C.F., Levkowitz, H.: Uncovering clusters in crowded parallel coordinates visualizations. In: Proceedings of the IEEE Symposium on Information Visualization, Washington, DC, USA, IEEE Computer Society (2004) 81–88
6. Elmqvist, N., Dragicevic, P., Fekete, J.D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. IEEE Transactions on Visualization and Computer Graphics **14** (2008) 1141–1148
7. Hauser, H., Ledermann, F., Doleisch, H.: Angular brushing of extended parallel coordinates. In: INFOVIS '02: Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02), Washington, DC, USA, IEEE Computer Society (2002) 127–130
8. McDonnell, K.T., Mueller, K.: Illustrative parallel coordinates. Computer Graphics Forum **27**(3) (2008) 1031–1038
9. Ankerst, M., Berchtold, S., Keim, D.A.: Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In: INFOVIS '98: Proceedings of the 1998 IEEE Symposium on Information Visualization, Washington, DC, USA, IEEE Computer Society (1998) 52–60
10. Yang, J., Peng, W., Ward, M.O., Rundensteiner, E.A.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In: Proceedings of the Ninth annual IEEE conference on Information visualization, Washington, DC, USA, IEEE Computer Society (2003) 105–112
11. Peng, W., Ward, M.O., Rundensteiner, E.A.: Clutter reduction in Multi-Dimensional data visualization using dimension reordering. In: Proceedings of the IEEE Symposium on Information Visualization, Washington, DC, USA, IEEE Computer Society (2004) 89âĂŞ96 ACM ID: 1038787.

12. Hoffman, P., Grinstein, G., Pinkney, D.: 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, New York, NY, USA, ACM (1999) 9–16
13. Kandogan, E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: Proceedings of the ACM international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2001) 107–116
14. Oesterling, P., Heine, C., Jänicke, H., Scheuermann, G.: Visual analysis of high dimensional point clouds using topological landscapes. In: Pacific Visualization Symposium (PacificVis), 2010 IEEE. (Mar. 2010) 113 –120
15. Jänicke, H., Böttinger, M., Scheuermann, G.: Brushing of attribute clouds for the visualization of multivariate data. IEEE Transactions on Visualization and Computer Graphics **14** (2008) 1459–1466
16. Ingram, S., Munzner, T., Olano, M.: Glimmer: Multilevel mds on the gpu. IEEE Transactions on Visualization and Computer Graphics **15** (2009) 249–261
17. Paulovich, F.V., Oliveira, M.C.F., Minghim, R.: The projection explorer: A flexible tool for projection-based multidimensional visualization. In: Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing, Washington, DC, USA, IEEE Computer Society (2007) 27–36
18. Yang, J., Ward, M.O., Rundensteiner, E.A., Huang, S.: Visual hierarchical dimension reduction for exploration of high dimensional datasets. In: Proceedings of the symposium on Data visualisation 2003. VISSYM '03, Aire-la-Ville, Switzerland, Switzerland, Eurographics Association (2003) 19–28
19. Yang, J., Patro, A., Huang, S., Mehta, N., Ward, M.O., Rundensteiner, E.A.: Value and relation display for interactive exploration of high dimensional datasets. In: Proceedings of the IEEE Symposium on Information Visualization, Washington, DC, USA, IEEE Computer Society (2004) 73–80
20. Johansson, S., Johansson, J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. IEEE Transactions on Visualization and Computer Graphics **15** (November 2009) 993–1000
21. Yuan, X., Guo, P., Xiao, H., Zhou, H., Qu, H.: Scattering points in parallel coordinates. IEEE Transactions on Visualization and Computer Graphics **15** (November 2009) 1001–1008
22. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. Proceedings of the IEEE Symposium on Visual Languages (1996) 336–343
23. Bein, K., Zhao, Y., Wexler, A.: Conditional sampling for source-oriented toxicological studies using a single particle mass spectrometer. Environmental Science and Technology **43**(24) (2009) 9445–9452