DOI: 10.1111/j.1467-8659.2011.01941.x

Eurographics / IEEE Symposium on Visualization 2011 (EuroVis 2011)
H. Hauser, H. Pfister, and J. J. van Wijk
(Guest Editors)

Volume 30 (2011), Number 3

Structural Decomposition Trees

D. Engel †1, R. Rosenbaum2, B. Hamann2, and H. Hagen1

1University of Kaiserslautern, Germany
2Institute of Data Analysis and Visualization at the University of California, Davis, USA

Abstract

Researchers and analysts in modern industrial and academic environments are faced with a daunting amount of multi-dimensional data. While there has been significant development in the areas of data mining and knowledge discovery, there is still the need for improved visualizations and generic solutions. The state-of-the-art in visual analytics and exploratory data visualization is to incorporate more profound analysis methods while focusing on fast interactive abilities. The common trend in these scenarios is to either visualize an abstraction of the data set or to better utilize screen-space.

This paper presents a novel technique that combines clustering, dimension reduction and multi-dimensional data representation to form a multivariate data visualization that incorporates both detail and overview. This amalgamation counters the individual drawbacks of common projection and multi-dimensional data visualization techniques, namely ambiguity and clutter. A specific clustering criterion is used to decompose a multi-dimensional data set into a hierarchical tree structure. This decomposition is embedded in a novel Dimensional Anchor visualization through the use of a weighted linear dimension reduction technique. The resulting Structural Decomposition Tree (SDT) provides not only an insight of the data set's inherent structure, but also conveys detailed coordinate value information. Further, fast and intuitive interaction techniques are explored in order to guide the user in highlighting, brushing, and filtering of the data.

Categories and Subject Descriptors (according to ACM CCS): I.4.4 [IMAGE PROCESSING AND COMPUTER VISION]: Image Representation—Multidimensional I.5.4 [PATTERN RECOGNITION]: Clustering—Similarity measures

1. Introduction

Due to enhanced data acquisition and analysis methodologies in almost all application domains, more and more truly massive and high-dimensional data sets are being produced that require the development of fundamentally new approaches for analysis. Mathematically based approaches have recently been demonstrated as being especially valuable and promising in this problem domain. However, gaining insight into high-dimensional data by a meaningful visual representation is a still unsolved research problem.

From the various methods proposed in literature, two fundamentally different approaches can be identified: (1) value and (2) relation visualizations. The first approach focuses on the visualization of the individual dimension contributions of each multi-dimensional data point. This is usually achieved by specific visual mappings. Parallel coordinates and color maps are amongst this class of approaches. Although these techniques allow for a quick visual access to the details of each data point, they lack of an overview and are usually subject to strong visual clutter. The second group of visualizations abstracts from those details and aims to visualize the relations of the points in the high-dimensional data space. Common means to achieve this are projection into a low dimensional presentation space. While projections are an excellent approach for showing relations in the data, they immensely suffer from ambiguity that are imposed by dimension reduction. They can easily lead to wrong conclusions about the data set.

This work combines the two distinct approaches of value and relation visualizations into a single holistic technique that benefits from both methods. To achieve this a novel value visualization is embedded into a point projection. Cen-

[†] d_engel@cs.uni-kl.de

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and
350 Main Street, Malden, MA 02148, USA.

tral to our technique is the calculation and display of a *structural decomposition tree* (SDT), which

1. removes ambiguities within the low-dimensional point representation,
2. visualizes the data points' coordinates together with their composition, and
3. serves as an efficient tool for visual exploration.

The SDT is designed to convey the composition of a given data set and is calculated with regard to optimal length and extend in order to minimize redundancies and clutter. To achieve this, the data is hierarchically decomposed under the criterion of commonalities and projected in a way that maximizes the distances between branches of different composition. The SDT provides an overview to general properties of the data as well as a detailed comparative view to individual coordinate values. Ambiguities imposed by the projection are solved by the branch structure. As shown by different examples, the branch structure is also a great means to identify clusters of data points.

After the presentation of related work (section 2), we introduce SDTs, discuss related problems, and describe the main ideas of our approach (section 3). This is followed by details of the construction of SDTs (section 4), as well as their interaction mechanisms (section 5). Results are discussed (section 6) and underlined by a case study (section 7. Lastly, concluding remarks are summarized (section 8).

2. Related work

Several surveys containing different categorizations of multi-dimensional (m-D) data visualizations have been proposed, such as [dOL03], [HGM*97], [GTC01]. For example, the taxonomy of basic techniques by Ward et al. [WGK10] is derived by the emphasis on the visual primitives used to represent the data. We choose to further abstract these taxonomies and focus on the information that the visualizations are to convey, being either coordinate *values of* data points or *relations between* data points. Therefore, we categorize the field of m-D data visualization into the two basic approaches of value- and relation visualization.

Value visualizations allow detailed analysis by visualizing the coordinate values of every data point. Heatmaps, Glyphs, Scatter Plot Matrices, and Parallel Coordinates can be considered as members of this category. A common problem with these techniques is that they are often not scalable with regard to the amount and dimensionality of the data. The visualizations become less comprehensible as the number of dimensions increases and often get cluttered as the number of data elements increases. Instead of proposing new representations for m-D data, researchers have been mainly focused on overcoming the previously mentioned drawbacks. The emphasis has been on enhanced cluster visualization ([JLJC05], [ZYQ*08], or [AdO04]), brushing techniques ([EDF08] or [HLD02]), and better utilization of screen space [MM08]. However, clutter reduction through dimension ordering ([PWR04] or [YPWR03]) is often regarded as the main research focus in the realm of value visualizations. Based on data point correlations, dimension ordering techniques focus on the arrangement of dimensions in the visual representations they are applied to. The research conducted by Ankerst et al. was the first to formally state this arrangement problem [ABK98]. While these approaches are great improvements to coordinate visualization techniques, they still face scalability issues. Even if the dimensions are ordered and the data filtered perfectly, the information displayed may still be overwhelming for the user and no clear overview can be established.

The second category *relation visualization* is also referred to as dimension reduction techniques or point projections in literature. These techniques display m-D data by projecting points onto a lower dimensional space, so that distance relations between points in the projection space reflect specific relationships between the data points in m-D space. Since these relationships may be too complex to be completely conveyed in lower dimensional space, projections are in general ambiguous. As one of the first dimension reduction techniques to be proposed, Principal Components Analysis (PCA) conveys distance relations in m-D by orthogonally projecting into a plane that is aligned to capture the greatest variance of the data. Remarkably, PCA achieves this through a computationally fast linear transformation. The resulting projection is a genuine view that does not distort the data. In contrast, Multidimensional Scaling (MDS) is based on general similarity measures between data elements and computes a lower dimensional representation accordingly. This is achieved by formulating an optimization problem that is computationally more complex than linear transformations and distorts the data. Similarly to MDS but computationally less expensive for high-dimensional data, linear dimension reduction may also be weighted by similarities, as introduced by Koren et al. [KC04]. This method is incorporated into our approach due to its fast, robust, and highly flexible projection approach. Relation visualizations establish a good overview and are often incorporated in multi-view systems as exploratory devices, e.g., as in [POM07]. Depending on the application, research focuses on better representation of specific data structures, e.g., scientific point cloud data [OHJS10], a better incorporation of domain-appropriate analysis techniques, e.g., brushing and filtering [JBS08], or computational speed gains ([IMO09] or [PSN10]).

Successful representations of complex data often utilize metaphors of commonly understandable concepts, such as topological landscapes [WBP07]. Projections are especially hard to interpret since they convey no visual connection to the original dimensions. However, Dimensional Anchor Visualizations (DAVs) [HGP99] incorporate similar concepts by using dimensions as (often interactive) display objects that determine the mapping of m-D points for their projection. With the assistance of these understandable visual references, the user may influence and better interpret the projection process. RadViz [HGM*97] is an example for DAVs. It is a technique that illustrates a non-linear projection process in the form of spring forces that are connected to each data point and the Dimensional Anchors. Similarly, the Star Coordinates approach by Kandogan [Kan01] utilizes an even more intuitive projection process. This well-known projection treats DAs as unit vectors that are uniformly distributed along a circle and maps m-D points by linear combinations, as shown in Figure 1. We choose this approach as the basis for our technique due to its intuitive interaction ability and mapping process. The user may interact with the DAs by changing their end position, thus creating a new projective view on the data set. This provides an intuitive interface for viewing transformations by which the contributions of certain dimensions can be emphasized or neglected. However, the effectiveness of the presentation is strongly dependent on the quality of the initial projection. The approach discussed in [STTX08] is based on dimension ordering to find initial arrangements for the DAs in order to attain clusters within the data set. However, coordinates cannot be conveyed by these methods and the representation remains ambiguous.

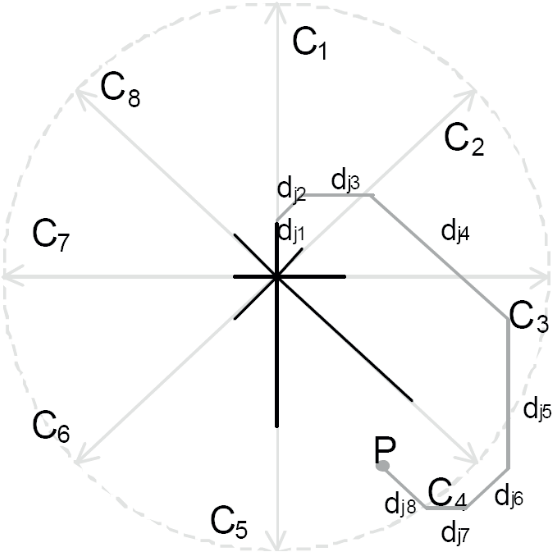

Figure 1: *A point $P = (d_{j,1}, ..., d_{j,8}) \in R^8$ is projected by the star coordinate system by the linear combination of its dimension anchors $C_1, ..., C_8$ with the point's coordinates as coefficients [Kan01]. However, many points can be projected to the same location, making this representation highly ambiguous if linear combinations are not shown.*

Few publications exist that combine the two approaches of value and relation visualizations. To the best of our knowledge, there is no related work that directly resembles our approach. However, some recent publications have tackled this combination from other perspectives. The following approaches integrate a projection method into Parallel Coordinates: Yang et al. [YPWR03] presents an importance-oriented dimension ordering approach that utilizes PCA, Johansson et al. [JJ09] propose a dimensionality reduction method that enables user-defined metrics and use this method to reduce clutter, enhance clusters and filter outliers, and Yuan et al. [YGX*09] allow the abstraction of a subset of dimensions by integrating scattered points arranged by MDS. Yang et al. have also utilized dimension hierarchies [YWRH03] or MDS [YPH*04] to display relations between dimensions and used pixel-oriented methods to display data values in form of glyphs for each dimension. In comparison, we present a novel approach that integrates a visualization of coordinates into a linear projection.

3. Main idea

The main idea of SDTs is to show how data points are projected. For this display, we use the Star Coordinates [Kan01] as basis. This projection is defined by a linear combination of unit vectors and coordinates for each dimension. The "projection path" is intuitively visualized by line segments as shown in Figure 1. Data point coordinates can be depicted and a unique path for each data point eliminates the ambiguities of the projection. However, this simple display of linear combinations increasingly clutters the display when the number of data points is large, rendering the benefits of the projection (the overview) useless. SDTs overcome this problem by having the following characteristics:

(1) In structured data, many points' coordinates are similar to some degree. Consequently, large parts of their linear combinations are similar. Such shared line segments can be aggregated, resulting in a more compact representation. One way to achieve this is by edge bundling. However, in this particular visualization, the information that is encoded in the orientation of edges (contributions of coordinates in one dimension) is easily lost when the edge is bent. Instead of using edge bundling (in geometry space), we compute a hierarchy of linear combinations (in data space) for all data points. At each level of this hierarchy, additional contributions explain the data's composition. The result is a tree in which each inner node is a compositional limiting point for the commonalities of succeeding nodes. Data points are the leafs of this tree and their coordinates are given by the sum of individual contributions along the path from leaf to origin (see Figure 2). We use *hierarchical clustering* to compute this hierarchy and achieve a tree with minimal overall branch length, thereby greatly reducing redundancies.

(2) Another aspect that highly influences visual clutter is the Dimensional Anchors' (DAs) initial arrangement. While this arrangement problem can be formulated as a 1-D optimization problem, this is computationally expensive. Instead, we apply a linear projection and use a sophisticated weighting scheme that greatly enhances the display of this structure. In particular, this optimizes the starting configu-

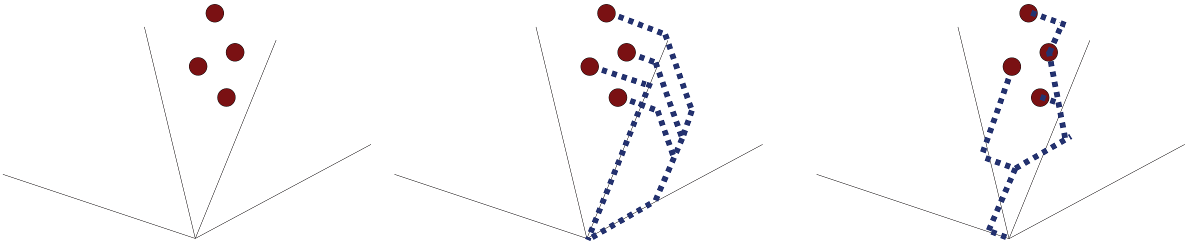

Figure 2: *By the representation through dimensional anchors alone (left), a simple but ambiguous view is achieved. Ambiguity may be solved by the display of the point's dimension contributions (center). However, this presentation highly clutters the view due to many and redundant line segments. A tree embedding, based on the structural composition of the data, achieves a trade-off in form of an unambiguous and less cluttered view (right).*

ration of DAs towards maximizing the space between tree paths.

(3) Special consideration is also placed on the *visual representation* of the SDT. For the different ways to depict the coordinate contributions, an ordering problem arises. To guarantee interactive capabilities, we employ a simple but fast ordering heuristic. In order to further enhance the recognition of data structure, the branches within the SDT encode the number of elements within this subtree in branch thickness and gray-scale.

(4) Appropriate *means for interaction* have also been developed to handle occlusion and guide in exploratory cluster analysis. Since visual analysis and exploration is indispensable, we enhance the intuitive interaction methods of DAs by novel techniques to aid in brushing, filtering, and selection. For example, one interaction technique hints at possibly interesting configurations of the projection.

The following section explains the details of these aspects. To generate an unambiguous view, we restrict SDTs to represent positive values only. Otherwise, the user would have to identify opposite-directed line segments as negative values of the respective dimension, which proved to be extremely counter-intuitive in our experiments. The high likelihood of line crossings was another essential factor for this restriction. However, in most cases this can be reasonably overcome by a translation of the data.

4. Construction of structural decomposition trees

In this section our algorithm will be introduced, as well as a detailed description of the achieved properties. Two preprocessing steps are necessary before the SDT can be visualized. First, a *hierarchical clustering* method computes the decomposition of the data for the visualization. Secondly, an *initial projection* is computed that emphasizes this structure. Finally, point coordinates and precomputed structure are visualized in a new *visual representation* that allows fast interaction. It should be noted that this step is computationally efficient in relation to the precomputations.

4.1. Hierarchical clustering

In order to create a structural decomposition, the data set is clustered hierarchically. There are many methods that perform hierarchical clustering, mostly varying in their use of inter-individual and inter-group metrics. Often, approaches are tailor-made to fit the specific requirements. In our case, the clustering step should generate an ideal tree structure, achieving a nesting with minimal redundancies, i.e., minimal overall line length. This structure should be ready for display, designed to minimize calculations at running time in order to support rendering at high frame rates.

The unfortunate restriction of drawing only positive coordinate values has proven to be a challenge for the computation of a non-redundant structure. In many clustering schemes, the average of cluster elements is compared as the representative elements for aggregation. This technique is referred to as group average linkage in literature [ELL01]. However, drawing the mean of a group (as part of the decomposition of the group's elements) would require to draw negative coordinate values, in order for the decomposition to hold. Since we have chosen to draw only positive coordinate values, we can only draw the minimum of the coordinates for each dimension as each stepwise decomposition of a group. These minimum commonalities therefore have to be the representatives for comparing clusters in our method.

Consider a matrix $X \in \mathbb{R}_+^{n \times m}$ of n m-dimensional data points where $x_{k,q}$ refers to the q'th coordinate of the k'th data point, as well as a complete and disjoint partition in clusters $C_1, \dots, C_{n_c}$ containing indices of data points, i.e., $C_i \subset \{1, \dots, n\}, 1 \leq i \leq n_c$. We define the inter-group proximity measure δ to quantify the measure of compositional commonalities between two clusters i and j as

$$\begin{aligned} \delta_{i,j} &= |\mathtt{min}(C_i \cup C_j)|, \text{ for } \mathtt{min}(C_i \cup C_j) \in \mathbb{R}^m \\ &= \sum_{1 \leq q \leq m} \mathtt{min}(C_i \cup C_j)_q, \end{aligned} \tag{1}$$

where $\delta_{i,j} \geq 0$. In analogy to the L^1 norm, we interpret δ as the length of a path in non-Euclidean space. We further define $\mathtt{min}$ of a collection of indices C as

$$\mathtt{min}(C) = (\min_{k\in C}(x_{k,1}),\dots,\min_{k\in C}(x_{k,m})),\ \text{for}\ C\subset\{1,\dots,n\} \quad (2)$$

and interpret it as the geometric minimum element, $minE$. It represents the point of the convex hull of m-dimensional points in a collection C that is closest to the origin, i.e., has the lowest norm as defined above.

A clustering method computing the desired decomposition can be summarized as the following binary hierarchical agglomerative process:

1. Generate a starting set G of n single-element-clusters C_i, each containing a different data point.
2. Iterate the following steps until G contains only a single (root) cluster.
   a. Search for the most appropriate pair within G, i.e., clusters C_i and C_j, for which $\delta_{i,j} = \max_{C_a,C_b\in G}(\delta_{a,b})$.
   b. Aggregate this pair to a cluster C_{new}, append this cluster to the set G and remove the original clusters C_i and C_j from G.
3. The single remaining cluster in G represents the root element of the structural decomposition.

Note that many bottom-up approaches have spatial distorting properties, like it is often the case when shortest pair distance or single linkage is used [ELL01]. Early approaches with common distance measures have lead to highly redundant structures and cluttered displays due to the rapid deterioration of the minimum representative through the aggregation within the hierarchy. Keeping big minimum commonalities has proven to be a key property for our structure. Therefore, we have developed a cluster criterion that forms maximal fitting cluster representatives, so that the aggregation steps along the hierarchy (from long to short minima) ensures the right spanning of the tree. The result is illustrated in Figure 3 and shows highly favorable properties.

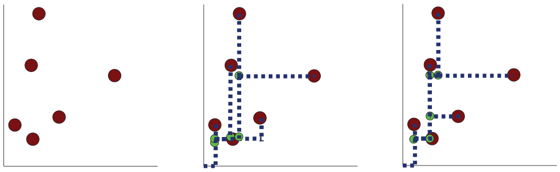

Figure 3: *2D data points without (left) hierarchically clustered by a shortest distance criterion (center). Spatial distortion with this metric leads to the tendency to low decomposition points and thus, to high redundancies. Clustering based on the criterion of the highest minimum commonality (right) achieves an embedding that minimizes redundancies, and shows no such spatial distortion effects.*

This clustering scheme is specially tailored for our visualization. In this context, it provides an optimal solution to reduce the overall redundant lines in terms of length, as we will show in the following. At each step, the two clusters are aggregated that have the maximum length of their joined $minE$. We find that this maximization of $|minE|$ is essentially equivalent to the minimization of the length of discrepancy to the (joined) father-node for each of the clusters. In other words, for C_i and C_j being aggregated, we denote $\theta_{i,j}$ as the discrepancy $|\mathtt{min}(C_i)| - \delta_{i,j}$ and find that at each step, both $\theta_{i,j}$ and $\theta_{j,i}$ are minimized if $\delta_{i,j}$ is maximized. Therefore, we achieve the minimization

$$\theta_{i,j} + \theta_{j,i} \leq \theta_{k,l} + \theta_{l,k}, for 1 \leq k,l \leq n_c, \quad (3)$$

for the aggregation of two clusters C_i and C_j, at any step of the clustering process. This observation is of key interest for our technique, since these discrepancies $\theta_{i,j}$ represent the length of the lines drawn for each (child-)node (C_i) of the SDT. Thus, the overall length of the SDT's line segments is stepwise minimized, which optimizes the representation in terms of compactness and minimal redundancies.

4.2. Initial projection

As discussed in section 2, there are many ways to project data. While one usually wants a projection to preserve relative $m-$D distances between data elements, our observation is that this does not necessarily lead to an ideal embedding for any data representation. For SDTs, the space between tree paths has a crucial influence on line intersections and visual clutter. Therefore, the Dimensional Anchors (DAs) need to be arranged in a way that maximizes this space.

While any linear projection method can be used to adjust the DAs, we use a weighted linear projection scheme according to [KC04] because it is fast, robust, and very flexible. By following this approach, pairwise weights are used to influence the covariance matrix so that its eigenvectors (and the two-dimensional projection given by the two 'highest' eigenvectors) reflect the pairwise dissimilarities given by or imposed to the data. By the means of these weights, objects are projected further away if they are highly dissimilar and vice versa. Consequently, we use a weighting that emphasizes the computed hierarchical data structure.

The initial projection is defined by the two eigenvectors, γ_1 and γ_2 of largest eigenvalues, computed from the weighted covariance matrix $X^T \mathscr{L}\ X$, where $X \in \mathbb{R}_+^{n\times m}$ is centered beforehand. The pairwise weights $\mathscr{L}_{i,j}$ are chosen to relate to the distances between nodes within our hierarchical decomposition. Let $dist^t_{C_i,C_j}$, for two clusters (nodes) C_i and C_j within our hierarchy, be the structural distance measure. This measure is formally defined as the number of edges along the path from C_i to C_j. The Laplacian matrix $\mathscr{L}$ is defined as in [KC04] and pairwise dissimilarities are ${dist^t_{C_i,C_j}}^2$. As shown in [KC04], we find that our projection maximizes

$$\sum_{i<j}(dist^t_{C_i,C_j}\ dist^p_{i,j})^2, \quad (4)$$

where $dist^p_{i,j}$ represents the Euclidean distance between the two projected points within the p-dimensional projection. Clearly, this equation is maximized by projecting those data points far from each other that hold a greater structural distance. Through this scheme, the projection is optimized to display the structural decomposition, which leads to an optimal separation of different tree paths (according to (4)). Note that this process is achieved through a linear transformation, preserving the genuine data properties as a true projective view. Since this structural distance disregards the actual proximity of points within the m-dimensional space, this weighting is also robust to outliers.

In order to derive the DA arrangement, the original m unit vectors are projected. The ith DA's end position, $a_i \in R^2$, is given by

$$a_i = (\begin{pmatrix}0\\1\end{pmatrix}\gamma_{1_i} + \begin{pmatrix}1\\0\end{pmatrix}\gamma_{2_i})^T, \qquad (5)$$

and may be normed or scaled (e.g., by eigenvalues) depending on the application. Note that not all dimensions are necessarily present (via DA) within this projection and that DAs may be collocated and of different scale. Since our approach is general, there may be applications where this mapping has to be adjusted due to specific data properties. However, we claim that collocation of anchors according to their dimension's correlation and a scale according to their contribution within the data is a faithful abstraction method. This way, correlating variables can be easily assessed within our visualization, which is a useful feature for an intuitive visual assessment of the data's properties.

4.3. Visual representation

Once the data is decomposed into a hierarchical structure and the Dimensional Anchors (DAs) are arranged to bring out this structure accordingly, the SDT can be visualized. The coordinates of a node (cluster) C_i are given by the precomputed $\mathfrak{min}(C_i)$. The node's position within the projection, $pos^p(\mathfrak{min}(C_i))$, is given by the linear combination with the DAs and $\mathfrak{min}(C_i)$ as corresponding coefficients. This position, however, does not convey the coordinates in an unambiguous way. As discussed in section 3, the linear combinations reflecting the point coordinates have to be visualized in order to assure an unambiguous visualization. Therefore, we visualize the path leading to each point's location - the actual linear combination.

For a node C_i, we have at least $m'!$ possible combinations to draw the path leading to the node, where $m' = |\{j \mid \mathfrak{min}(C_i)_j \neq 0, 1 \leq j \leq m\}|$. Therefore, the arising problem is to find a meaningful order of these m' line segments without spending large computational expenses on this arrangement. Since the DAs' orientation and scale are meant to be interactively changed, the resulting layout of these line segments changes accordingly. We want to keep rendering speed at interactive levels, while achieving a visually uncluttered representation. Consequently, an optimization process according to exact quality measures (e.g., line crossings) is not an option. However, we have discovered a good heuristic for this ordering being dependent on length and orientation of the line segments. For any line segment $\vec{s_j} = \vec{a}_j c_j$, the dot product to the normed direction from father C_k to child node C_i determines the drawing order of these connected segments. Thus, we draw decreasing with $\vec{v_c} \bullet \vec{s_i}$ for $\vec{v_c} = pos^p(\mathfrak{min}(C_i))^T \quad pos^p(\mathfrak{min}(C_k))^T$.

The actual rendering of the tree is a straight-forward recursion. Starting at the root C_r of the tree, $minE(C_r)$ is drawn and for each successive child node, the discrepancy to the father node is drawn. Thus, for every node C_i, being the child of a father node C_k, the discrepancy $minE(C_i) \quad minE(C_k)$ is drawn. As discussed with (3), this discrepancy is minimized in a stepwise fashion by our clustering algorithm which consequently reduces visual clutter. Another way of dealing with visual clutter and line crossings is to use color and shape for a better visual recognition of different paths. For the line segments of a node, an appealing color and width configuration is found to relate to the number of nodes within the current subtree, ranging from dark to light and broad to thin with decreasing element count.

It should be noted that the rendering of a SDT is potentially faster than in other value visualizations. By exploiting commonalities in our precomputation steps, the decomposition can reduce the overall objects that have to be drawn significantly. For example, while scatter plot matrices draw nm^2 points and parallel coordinates draw $n(m \quad 1)$ line segments for every data set, a SDT may have any number of line segments in $[0,(2n \quad 1)m]$, depending on the commonalities in the data. This may be of benefit for large data sets.

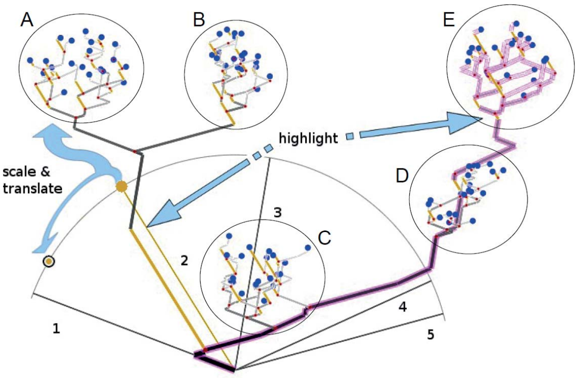

Figure 4: *A 5-D data set with 5 point-clouds is shown. SDTs best display differences and commonalities within the structural assembly of the data. Further analysis can be conducted by adjusting the projection, highlighting, or filtering.*

5. Interaction with structural decomposition trees

Effective visual analysis of high-dimensional data requires interaction. In this section we introduce the *means for interaction* provided by SDTs and illuminate on their eligibility for interactive visual cluster identification. Figure 4 illustrates these techniques.

Although the SDT's starting projection has desirable mathematical properties, interactive *adjustment of the projection* is usually needed to gain further insight. As Dimensional Anchors (DAs) are the visual representatives of the basis vectors of the m-dimensional coordinate system, rearranging their end points provides an intuitive interface for modifying the data projection. Scaling of the DAs is particularly useful to investigate the contribution of a dimension to the data set, where large contributions are stronger emphasized than small ones. Also, these interactions do not require additional computational expenses as the scene must only be redrawn.

However, a problem that is caused by the unconstrained placement of the DAs is that orthogonality of the projection is easily lost. As this might lead to misinterpretations of the data, novel visual clues have been implemented and made accessible through interaction. *Variance points* are placed along the unit circle in order to indicate angles that lead to an orthogonal projection for a selected dimension. They are computed as a combination of the first principal component with all other components and thus represent the different positions of this DA on the unit circle determined by PCA. The corresponding eigenvalues are also the foundation for another beneficial property of variance points. As their values represent the respective (weighted) variance in the data, we encode their values in the point's size by a ratio to the largest eigenvalue. Thus, further insight to the amount of information hidden by the current projection can be conveyed.

SDTs also support a variety of methods to highlight *data elements and properties*. *Dimension highlighting* emphasizes all line segments corresponding to the coordinates of a dimension by a distinct color. With this interaction, distribution of data in a dimension can be assessed more easily. Also, in high-dimensional data sets, the orientation of line segments might be hard to map to the corresponding dimension, which can be overcome by this means. Through the selection of a line segment, the respective DA is highlighted. A means to highlight structural parts of the data is *path highlighting*. It enables the user to select a node of the tree, after which the path from root to this node is highlighted, as well as all subsequent paths of the subtree. Path highlighting can select individual or a set of data points and is especially useful to investigate the structural decomposition of the data or when used as a selection tool in combination with other detailed data views. Should there exist dense regions, it also shows the unique path of items overcoming clutter. The introduction of SDTs also offers an intuitive option to simplify the presentation by *filtering data points or dimensions* currently of no interest. A simple filtering metaphor from graph visualization comes natural to the tree layout - *node collapsing*. The crotches of the SDT are ideal points to collapse and hide subtrees.

Presently, no stand-alone technique is capable to ideally support all analysis tasks for high-dimensional data. There is a clear trend towards multi-view-systems that link several techniques to combine their individual benefits. Providing a convenient overview of the data, as well as an intuitive interface for selection and filtering is a critical property of such systems. The unique support of intuitive interactions (zoom, pan, data selection, dimension highlighting, viewing manipulation) makes SDTs a suitable candidate to act as an overview and interface for such systems.

6. Results

To evaluate the visual representation of data structures by SDTs, we investigated both artificial and benchmark data sets. We observed that our approach is profoundly robust for connected data, e.g., ellipsoids or curves, that can be represented linearly and has an inherent structure. Quite notably, non-linear, non-convex, and even higher genus shapes of different sizes are presented well by the structurally weighted projection approach, as Figure 5a-c shows. We observed that closely connected compositional parts of curves or higher genus shapes share the same subtree within our computed hierarchy structure. Since the projection is optimized to emphasize these structural distances and since this structural distance is a topological feature of the data, we can in fact note that our approach is topology-preserving in terms of an optimization regarding the alignment of the projection plane.

Further, two popular and real-world multi-dimensional data sets (Iris and Cars) have been chosen to act as benchmark data sets and to provide a comparison to other techniques using the same data sets, e.g., [PWR04], [SYHX08], [YGX*09], or [WGK10]. In Figure 5d, the SDT of the Iris data reveals three compositionally distinct groups. The strong stem indicates a clear structure and shows that the sepal width accounts for the most commonalities among the collected species. Two of the three species are similar but can be distinguished in their different variance, as well as different magnitudes, in petal length and width. In Figure 5e, five groups can be distinguished resigning at three different levels of the tree. These (vertical) levels represent car properties ranging from "efficiency" (specs of high acceleration, low weight, MPG, and displacement) to "high power" (the opposite) . The initial projection correctly hints at correlations between these properties. European and Japanese cars are dominant within the lower "efficiency" tree level (bottom right), while American cars reside in all levels of the tree (left, vertical side).

These results show that structural relations between clusters within data are well represented and differences in coordinates between clusters can be perceived easily. SDTs are

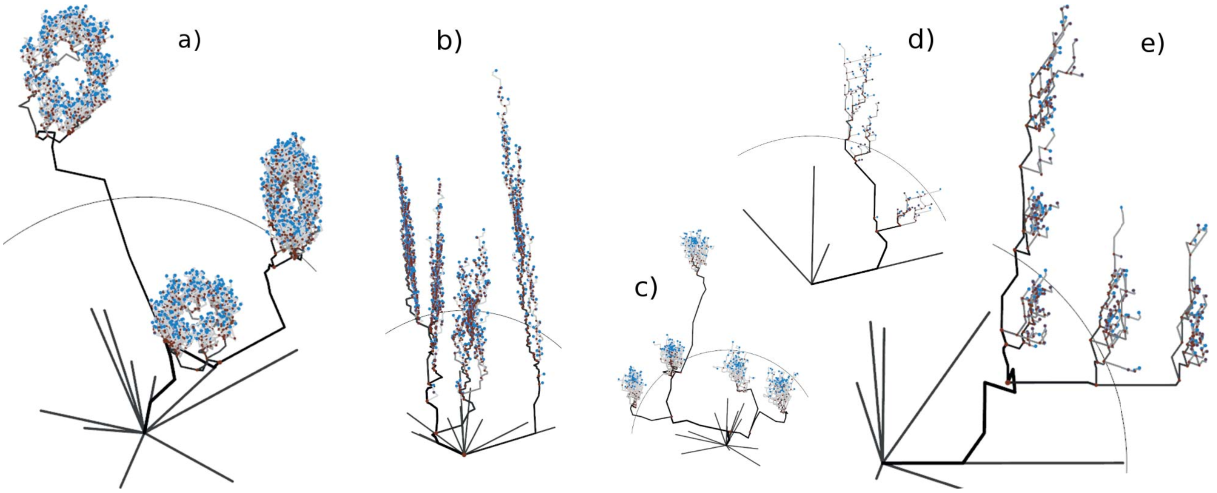

Figure 5: *Artificial data sets: **(a)** 3 tori in $\mathbb{R}^{10}$, **(b)** 4 ellipsoids in $\mathbb{R}^{10}$, **(c)** 5 point clouds in $\mathbb{R}^{15}$; Benchmark data sets: **(d)** Iris and **(e)** Cars data set.*

most suited to depict a general impression of a data set and to convey an intuitive visual mapping of m-D data that can be remembered. The user can learn easily

- 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 with regard to intra-cluster variances, shape, or alignment.

SDTs not only show how different (*relations*) data points are but also where these differences lie (*values*) by giving an assessable connection to multi-dimensional space. This is conveyed in a compact and intuitive representation which leads to a better interpretation of the data and is the main contribution of our work. Since SDTs focus on providing an overview of the data, traditional value visualizations are better suited for detailed analysis tasks. Embedded in an interactive framework, SDTs are appropriate as a device for exploration, selection, and filtering. The benefit of the underlying projection becomes even more obvious for higher-dimensional data. One can observe that clusters are well represented, even in very high-dimensional data sets, where one can argue that purely value visualizations fail. The amalgamation of value- and relation visualization makes SDTs more powerful than linear projections and more scalable than value visualizations.

Inherent in our method, drawbacks of SDTs are shared with those of linear projections and concern the suitability for unstructured, noisy, or manifold data. For such data, SDTs resemble more "cluttered bushes" than structured trees. However, our approach is generic and offers many possibilities for adjustments, e.g., in data transformation, projection, and clustering, to better fit specific applications.

7. Case study: air quality data

We have evaluated our method to real-world data provided by the UC Davis air quality research center and obtained by single particle mass spectrometry [BW09]. The raw 256-dimensional data has undergone application-specific data transformations (normalization) as well as dimension reduction to the 13 dimensions most important for the investigation purposes of our collaborators. 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 we got during analysis.

Figure 6 a), shows the obtained initial projection for 1000 particles randomly selected from a sampling campaign at three different sites. This first view clearly reveals two main clusters corresponding to the different sampling sites. Due to a similar particle compositions, however, both campaigns ran for Fresno can only hardly be distinguished (green and blue dots) even with the support of the SDT. Three dimensions are highly significant for all campaigns: *C*-Carbon, *NOx*-Nitrogen oxides, and *Po*-Potassium. By using *dimension highlighting* it can be revealed that there are significantly higher *C* concentrations in Fresno than in Pittsburgh (see Figure 6 b)). The opposite applies to *Po*. *NOx* is more variant and can be found in similar contributions in both sites (see Figure 6 c)).

While exploring the data by *projection adjustments*, it is possible to show that dimension *C*24 representing a carbon isotope can not be found in Pittsburgh and also has only small concentrations in Fresno (see Figure 6 d)). Contributions of *C*36, another carbon isotope, can be found in similar concentrations at both sampling sites with either high or low

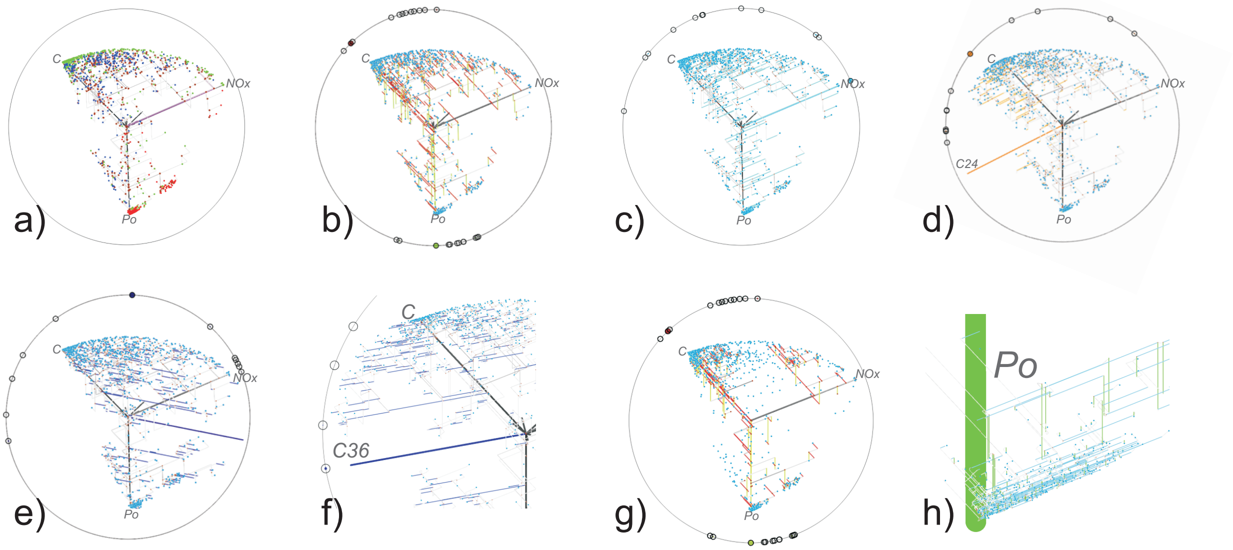

Figure 6: *Structural decomposition tree of 1000 data points obtained from three different air particle sampling campaigns: **(a)** Initial projection. The coloring of nodes is used for illustration purposes only (red: Pittsburgh, 2002; blue: Fresno, 2007; green: Fresno, 2009). Dimension highlighting applied to dimensions C, Po **(b)**, and NOx **(c)**. Adjusting the projection by moving the anchors corresponding to dimension C24 **(d)** and C36 **(e)**. An inverse correlation to C could be revealed by moving the anchor of dimension C36 to one of its variance points **(f)**. Options for filtering ((**g**); analog to (**b**)) as well as zoom and pan **(h)** allow to further adjust the view to current needs.*

values (see Figure 6 e)). Moving the corresponding dimension anchor to one of its *variance points* also indicates an inverse correlation of *C*36 to *C* (see Figure 6 f)).

Figure 6 g) illustrates the effect of dimension and node filtering leading to a reduced number of displayed tree items and thus less occlusion. Compared to Figure 6 b) only details relevant to the domain scientists are shown. Options to drill into interesting parts of the projection provide more details. As shown in Figure 6 h) one can see that all points belonging to the Pittsburgh sub-cluster show almost identical *Po* concentrations, but vary strongly in their *NOx* contents.

8. Conclusions and possible research directions

We have introduced a new method for the visualization of high-dimensional data based on the idea of representing and visualizing the data's structure by a tree. This approach leads to visualizations that allow one to comprehend relations between clusters in high-dimensional data and helps to reinforce a mental mapping of these relations.

The computation and display of this structural decomposition tree (SDT) is optimized with regard to depicting minimal redundancies in order to prevent cluttering. The result is a meaningful embedding of coordinate values in a point projection. This approach tackles the issue of ambiguities introduced by projection effectively and further supports the capability of producing an overview of the data. For effective data investigation, we have developed unique interaction techniques that enhance exploratory capabilities such as projection adjusting, feature highlighting, and data filtering.

It is planned to perform additional research to tackle problems of occlusion that arise with unstructured data sets by embedding more sophisticated brushing, feature enhancing, filtering, and abstraction techniques that are successfully applied to line drawing visualizations. Future research plans also involve a better evaluation of our technique by a user study and a detailed description of the geometric properties of SDTs.

Acknowledgements

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) as a joint effort of the International Graduate School (IRTG) in Kaiserslautern, Germany (DFG-#1131), and the Institute of Data Analysis and Visualization (IDAV) at the University of California, Davis, USA (DFG-#RO3755/1-1). Special thanks belongs to Taimur Khan and Christian Wagner.

References

[ABK98] 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, 1998), IEEE Computer Society, p. 52. 2

[AdO04] ARTERO A. O., DE OLIVEIRA M. C. F.: Levkowitz h.: Uncovering clusters in crowded parallel coordinates visualizations. *IEEE Symp. on Information Visualization* (2004). 2

[BW09] BEIN K.J. Y. Z., WEXLER A.: Conditional sampling for source-oriented toxicological studies using a single particle mass spectrometer. *Environmental Sciience and Technology 43*, 24 (2009), 9445–9452. 8

[dOL03] DE OLIVEIRA M. C. F., LEVKOWITZ H.: From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics 9* (2003), 378–394. 2

[EDF08] 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. 2

[ELL01] EVERITT B. S., LANDAU S., LEESE M.: *Cluster Analysis*, 4th ed. Arnold, London, 2001. 4, 5

[GTC01] GRINSTEIN G., TRUTSCHL M., CVEK U.: High-dimensional visualizations. In *Proceedings of Visual Data Mining workshop, KDD'2001* (2001). 2

[HGM*97] HOFFMAN P., GRINSTEIN G., MARX K., GROSSE I., STANLEY E.: Dna visual and analytic data mining. In *Proceedings of the 8th conference on Visualization '97* (Los Alamitos, CA, USA, 1997), VIS '97, IEEE Computer Society Press, pp. 437–ff. 2, 3

[HGP99] 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 in conjunction with the eighth ACM internation conference on Information and knowledge management* (New York, NY, USA, 1999), ACM, pp. 9–16. 2

[HLD02] 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, 2002), IEEE Computer Society, p. 127. 2

[IMO09] INGRAM S., MUNZNER T., OLANO M.: Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics 15* (2009), 249–261. 2

[JBS08] 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. 2

[JJ09] 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. 3

[JLJC05] 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, 2005), IEEE Computer Society, pp. 17–. 2

[Kan01] KANDOGAN E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2001), KDD '01, ACM, pp. 107–116. 3

[KC04] KOREN Y., CARMEL L.: Robust linear dimensionality reduction. *Visualization and Computer Graphics, IEEE Transactions on 10*, 4 (jul. 2004), 459 –470. 2, 5

[MM08] MCDONNELL K. T., MUELLER K.: Illustrative parallel coordinates. *IEEE-VGTC Symposium on Visualization 2008* (2008). 2

[OHJS10] 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), pp. 113 –120. 2

[POM07] 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, 2007), IEEE Computer Society, pp. 27–36. 2

[PSN10] PAULOVICH F., SILVA C., NONATO L.: Two-phase mapping for projecting massive data sets. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (Nov.-Dec. 2010), 1281 –1290. 2

[PWR04] PENG W., WARD M. O., RUNDENSTEINER E. A.: Clutter reduction in multi-dimensional data visualization using dimension reordering. In *In INFOVIS âĂŹ04: Proceedings of the IEEE Symposium on Information Visualization (INFOVISâĂŹ04* (2004), IEEE Computer Society, pp. 89–96. 2, 7

[STTX08] SUN Y., TANG J., TANG D., XIAO W.: Advanced star coordinates. In *WAIM '08: Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management* (Washington, DC, USA, 2008), IEEE Computer Society, pp. 165–170. 3

[SYHX08] SUN Y., YUAN J., HU Y., XIAO W.: An improved multivariate data visualization technique. In *Information and Automation, 2008. ICIA 2008. International Conference on* (June 2008), pp. 1525 –1530. 7

[WBP07] WEBER G., BREMER P.-T., PASCUCCI V.: Topological landscapes: A terrain metaphor for scientific data. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (Nov.-Dec. 2007), 1416 –1423. 2

[WGK10] WARD M. O., GRINSTEIN G., KEIM D. A.: *Interactive Data Visualization: Foundations, Techniques, and Application.* A. K. Peters, Ltd, 2010. 2, 7

[YGX*09] 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. 3, 7

[YPH*04] 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, 2004), IEEE Computer Society, pp. 73–80. 3

[YPWR03] YANG J., PENG W., WARD M. O., RUNDENSTEINER E. A.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. IEEE Symposium on Information Visualization* (2003). 2, 3

[YWRH03] 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* (Aire-la-Ville, Switzerland, Switzerland, 2003), VISSYM '03, Eurographics Association, pp. 19–28. 3

[ZYQ*08] ZHOU H., YUAN X., QU H., CUI W., CHEN B.: Visual Clustering in Parallel Coordinates. *Computer Graphics Forum 27*, 3 (May 2008), 1047–1054. 2