Explorative and Model-based Visual Analysis of Multivariate Data

Daniel Engel

11. Juli, 2014

vom Fachbereich Informatik
der Technischen Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)
genehmigte Dissertation

Gutachter:	Prof. Dr. Hans Hagen, TU Kaiserslautern
	Prof. Dr. Bernd Hamann, UC Davis
	Prof. Dr. Thomas Wischgoll, Wright State University
Vorsitzender:	Prof. Dr. Markus Nebel, TU Kaiserslautern
Dekan:	Prof. Dr. Klaus Schneider, TU Kaiserslautern

D 386

gewidmet
Wilhem Kirstges

Preface

During the course of my studies, both at the TU Kaiserslautern and at UC Davis, I have been very fortunate to meet a great many people that have helped me in my endeavors. I want to use this opportunity and express my gratitude to them.

First, I want to thank my main advisor, Prof. Dr. Hans Hagen, who is not only an exceptional person and pioneer of the field, but also a remarkable supervisor. His engagement to educate and shape the minds of young student researchers by gradually passing on his knowledge, experience, and wisdom, is unparallel. He works hard to provide a firm frame for an open environment in which there is room to grow. Thank you for your trust in me. I am proud to have you as my "Doktorvater"!

I also want to thank my second advisor, Prof. Dr. Bernd Hamann, who has not only supervised me during my time in the states but has kept supervising me even when I was not in Davis. Over the years, he has shaped this work by great discussions, many of which over skype. Thank you for the time and effort you have put in me!

Being fortunate enough to have studied in an international research training group, I have got to meet many people in many places that have my gratitude. Naming everyone would exceed the limits of this page, so I list a representative few. In Davis, I would like to thank Ken Joy, Hank Childs, and Harald Obermayer for their hospitality and for a great time in the Lab. I also want to thank my collaborators Anthony Wexler and Keith Bein from the Air Quality Research Center for our great collaboration. In Kaiserslautern, I thank the whole group for being just the way they are! In particular, I want to express my deep gratitude to Mady Gruys, who is the beating heart of our group. She keeps this chaotic bunch running, day in and day out, has always an open ear for whatever it may be, or a shoulder to cry on, when it should come to that. Thank you for being so awesome!

Last but not least, I want to thank my friends and family. I thank Inga Scheler for the friendship, trust, and support that we have shared over the years, ever since the day I was fortunate enough to stumble into your office. I also want to thank Mathias Hummel and Olga Beketova for being so incredible. Wholeheartedly, I thank Ursula Wessoly for the enduring love and support that we share. Finally, I want to thank my parents, Petra and Klaus Engel, who I have so much to thank for and want to address in German. Ich danke Euch für die bedingungslose Liebe und Unterstützung, die Ihr mir mein Leben lang geschenkt habt. Ihr habt mir alles gegeben, was ich mir hätte wünschen können!

Thank you!

Abstract

Researchers and analysts in modern industrial and academic environments are faced with a daunting amount of multivariate 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 improving interactive abilities, in order to support data analysts in gaining new insights through visual exploration and hypothesis building.

In the research field of exploratory data visualization, this thesis contributes new approaches in dimension reduction that tackle a number of shortcomings in state-of-the-art methods, such as interpretability and ambiguity. By combining methods from several disciplines, we describe how ambiguity can be countered effectively by visualizing coordinate values within a lower-dimensional embedding, thereby focusing on the display of the structural composition of high-dimensional data and on an intuitive depiction of inherent global relationships. We also describe how properties and alignment of high-dimensional manifolds can be analyzed in different levels of detail by means of a self-embedding hierarchy of local projections, each using full degree of freedom, while keeping the global context.

To the application field of air quality research, the thesis provides novel means for the research of aerosol source contributions. Triggered by this particularly challenging application problem, we instigate a new research direction in the area of visual analytics by describing a methodology to model-based visual analysis that (i) allows the scientist to be "in the loop" of computations and (ii) enables him to verify and control the analysis process, in order to steer computations towards physical meaning. Careful reflection of our work in this application has led us to derive key design choices that underlie and transcend beyond application-specific solutions. As a result, we describe a general design methodology to computing parameters of a pre-defined analytical model that map to multivariate data. Core applications areas that can benefit from our approach are within engineering disciplines, such as civil, chemical, electrical, and mechanical engineering, as well as in geology, physics, and biology.

Contents

CHAPTER 1

Introduction

"Applications trigger new basic research problems." - Hans Hagen

Simulation and experimental data acquisition enable scientists and engineers to gather vast amounts of data. Thereby, more and more application domains are producing progressively larger and inherently more complex (multivariate) data sets. Visualization, the first and foremost step of data analysis, is presented acute challenges by multivariate data because its full depiction requires more degrees of freedom than what is feasible for physical display devices, as well as for human perception. Visualization is to find visual representations of this data that are comprehensible to the user and feasible to the user's tasks and goals. While conveying as much information within the data as possible, this amounts to no small scale of abstraction and approximation. Thereby, the challenge lies in finding solutions that are scalable, interpretable, and interactive. This thesis addresses novel techniques for explorative and model-based visual analysis that meet these challenges.

Working closely with application scientists and engineers, this research combines domain expertise from key disciplines, such as machine learning, human-computer-interaction, and visualization, to develop fundamentally new methodologies that are required to support interactive visual data exploration and analysis. The thesis contributes to the research field by methods in two topics: explorative and model-based visual analysis.

In *explorative visual analysis*, the focus is laid on utilizing dimension reduction for the visualization of high-dimensional data. Dimension reduction finds an optimal representation of points in lower-dimensional space that best reflects the inherent properties, such as variance or similarity, of points in high-dimensional space. The benefit is hereby that these techniques are scalable and highly intuitive in terms of their visual representation, for example, for identifying outliers or groups in the data. Complemented with higher-detail views, dimension reduction can facilitate a powerful interface for explorative visual analysis.

Nevertheless, state-of-the-art methods for dimension reduction also have a number of shortcomings, such as ambiguity, by which the applicability of the technique immensely suffers. This is inherent by concept, as the abstraction of complex proximity relationships yields approximation errors that are not conveyed in the point arrangement of projections. Such ambiguous projections, for example, showing a group of points being very close that

are actually far apart in high-dimensional space, can easily lead to wrong conclusions about the data set. Based purely on point arrangement, projections are generally hard to interpret for scientists as they do not convey actual data values or ways to ascertain if the global layout reflects local properties of the data. With the limited degrees of freedom available to map high-dimensional data, these problems can, in all likelihood, not be solved from a mathematical point of view.

However, we show in this work that such shortcomings can in fact be countered effectively by combining methods from different disciplines to aid in the visual representation of dimension reduction. We describe how concepts of information visualization may be used to gain additional degrees of freedom in lower-dimensional embeddings, or to help with better interpretability and interactivity in adjusting both view and model of the lower-dimensional mapping. Incorporation of level-of-detail approaches for data abstraction and new concepts for visual verification to evaluate the error and ambiguity of a mapping are further essential approaches by which we improve existing work. In particular, this thesis contributes two novel methods that render dimension reduction more interpretable, interactive, and effective in the context of explorative visual data analysis.

The first method [ERHH11, REM*12, EHHR13] incorporates data values within the lower-dimensional embedding to enhance analytical capabilities, illustrate the structure of the data, and counter ambiguity of lower-dimensional projection. It is the first method to achieve an embedding of coordinate values within dimension reduction and incorporates many intuitive interaction mechanisms to analyze and explore high-dimensional data. The second method [EKHS14] utilizes graph abstraction techniques to achieve a dimension reduction that incorporates different levels of detail. Thereby, the user is able to explore high-dimensional data in different levels interactively by traversing a hierarchy of consecutive self-embedding projections corresponding to compositional sub-regions of the high-dimensional manifold. Both methods represent fundamentally new algorithmic and analytical approaches of explorative visual data analysis.

In addition to our contributions in explorative visual data analysis, we describe a fundamentally different and new visualization approach that we call *model-based visual analysis*. Our work was triggered by the application of air quality research, where collaborating research partners have the task to characterize chemical constituents within airborne particles based on data acquisition by single particle mass spectrometry (SPMS). As of now, no visualization method was able to depict these results in a comprehensible manner that would facilitate the analytical task. Measurements provided in form of mass spectra represent complex mixtures of various sources. The goal of visual analysis is to extract these sources that constitute the individual particles in form of latent variables of the data. However, as both mass spectrum and mixture model are non-negative, common methods of dimension reduction, such as spectral decompositions, are not applicable to this task. Further, neither the visualization of the mass spectra nor the within comprised features would facilitate the analytical task of characterizing the intermixing of each individual particle.

In contrast to previous work conducted in explorative visual analysis, the analytical goal of the application is not merely to explore or comprehend their data, but to conduct an

analysis based on a specific model. Visualizing this analytical interpretation of the data cannot be facilitated by the methods for data exploration. In particular, methods like independent component analysis fail to portray the physical mixture model of particle constituents. Hence, results are not meaningful in terms of atmospheric processing and not useful to atmospheric scientists. Methods for data clustering cannot handle the subtle differences between chemical constituents and cannot unravel ambiguity in the data known as isobaric interference. As a consequence, data analysis in the domain has involved a copious amount of manual investigation by the scientist - a highly time consuming and subjective process.

Triggered by this application problem, our work instigates a fundamentally new basic research direction for visualization: model-based visual analysis. Dealing effectively with this problem necessitates the practical need for visualizing the analytical reasoning of an optimization process that is specifically designed to map to the application's analysis model. Our methodology entails a visual interface to this optimization that (i) allows the scientist to be "in the loop" of computations and (ii) enables him to verify and control the optimization process, in order to steer computations towards physically and mathematically correct interpretable results.

The model for the intermixing of chemical constituents is incorporated both in the visual representation, as well as in the semi-automatic analysis process [EGG*12]. Approximation errors, that are a necessity to the non-convex optimization process, are further analyzed in a new methodology based on the information content of the feature basis and visually presented to the analyst in an explorative visualization [EHH*13]. These methods have been integrated in a single framework for the visual analysis of SPMS data, tested, and evaluated in the application area, serving as a proof of concept. Using this framework, our collaborators have been able to (i) reproduce established findings in mere a fraction of the time, (ii) process and analyze considerably more data than in previous studies, and (iii) gain new insights enabled by the visualization.

The outstanding success of our methodology in the application of air quality research naturally prompts the question of its applicability to other domains. We evaluate this potential and describe the general design choices that underlie and transcend beyond our application-specific solution [EHHS14]. The task facilitated by this methodology is generalized to computing parameters of a pre-defined analytical model that map to given data. To achieve this task, our methodology represents an interdisciplinary and integral approach involving mathematical modeling, machine learning, visualization, and human-computer interaction. In particular, it involves the incorporation of the scientist in every step of both the design and the usage of the final system. With the generalization to model-based visual analysis, our goal is to provide guidelines for the design of tools that can help scientists solve mathematically ill-posed problems in a semi-automated manner, where the computer does the "heavy lifting" and maintains unsubjective mathematical rigor, while the scientists oversees and steers computations towards physical meaning. A broad range of application areas could be identified for our approach.

The manuscript is structured as follows. Chapter 2 introduces the bare minimum basic concepts in terms of mathematical, visual, and analytical approaches for high-dimensional

data by summarizing established principles. Chapter 3 and Chapter 4 describe our own work in the area of explorative and model-based visual analysis, respectively. Within each chapter, the sections consecutively build on one another, while the order of chapters reflect both the level of user involvement and knowledge of the field needed to follow the work described.

CHAPTER 2

Fundamental Concepts

This section summarizes fundamental concepts and related work in the scope of this thesis. It includes basic notations and conventions used throughout this manuscript. After a brief overview of the mathematical background, the key concepts of multivariate analytics and visualization are presented.

2.1 Mathematical background

It is assumed that the reader is familiar with standard concepts of point, vector, and matrix operations. This section merely gives a brief overview of the essential mathematical background needed for multivariate data analysis and recapitulates some of the main definitions.

2.1.1 Matrix algebra

A *matrix* $\boldsymbol{A}$ represents a rectangular array of elements with n rows and m columns, n,$m \in N$, of order $(n \times m)$. The element in the ith row and jth column is referred by $a_{i,j}$, whereas $a_{i,\bullet}$ refers to the entire ith row, a $(1 \times m)$ matrix, and $a_{\bullet,j}$ refers to the entire jth column, given by a $(n \times 1)$ matrix. The *matrix product* C of matrices A and B is defined by

$$c_{i,j} = a_{i,\bullet} \times b_{\bullet,j} = \sum_{1 \leq k \leq p} a_{i,k}\ b_{k,j}, \tag{2.1}$$

for $1 \leq i \leq n$, $1 \leq j \leq m$ and the orders stated below.

$$\begin{bmatrix} a_{0,0} & \dots & a_{0,p} \\ a_{n,0} & \dots & a_{n,p} \end{bmatrix} \times \begin{bmatrix} b_{0,0} & b_{0,m} \\ \vdots & \vdots \\ b_{p,0} & b_{p,m} \end{bmatrix} = \begin{bmatrix} c_{0,0} & c_{0,m} \\ c_{n,0} & c_{n,m} \end{bmatrix}$$

$$(n \times p) \qquad (p \times m) \qquad (n \times m)$$

Matrix multiplication is associative, i.e. $A(BC) = (AB)C$, but not commutative, i.e. $AB \neq BA$ in general. The *transpose matrix* A^T of a matrix A is obtained by interchanging its rows and columns. Also, a *square matrix* $(n \times n)$ is said to be symmetric if $A = A^T$,

thus $a_{i,j} = a_{j,i}$ for $1 \leq i,j \leq n$. The *identity matrix* I is a square matrix for which the diagonal entities $a_{i,i}$ are unit measures, e.g. 1, and other entries are zero, $a_{i,j} = 0$ for $i \neq j$.

A m-dimensional *point* represents a $(1 \times m)$ matrix P, whereas the transpose P^T of order $(m \times 1)$ equals its corresponding *position vector*. Throughout this work, we conveniently interpret a vector $v \in \mathbb{R}^m$ as the position vector of a point $P \in \mathbb{R}^m$. For the purpose of this manuscript, the interchanging of terminology between point and vector, thus, refers to the transposition of the underlying matrix. Further, the *length of a point* $P \in \mathbb{R}^m$, also known as the Euclidean *norm*, refers to the *Euclidean distance* from the point's positional vector v to the origin and is defined as

$$||v||_2 = \sqrt{\sum_{1 \leq i \leq m} v_i^2}. \tag{2.2}$$

Vectors $v_1, v_2 \in \mathbb{R}^m$ are said to be *orthogonal* to each other when

$$v_1^T \times v_2 = 0 \tag{2.3}$$

and *orthonormal* to each other when

$$v_1^T \times v_1 = v_2^T \times v_2 = 1. \tag{2.4}$$

A set of vectors $v_1, ..., v_n$ are said to be *linearly dependent* if constants $c_1, ..., c_n$ can be found such that

$$c_1 \ v_1 + \cdots + c_n \ v_n = 0. \tag{2.5}$$

This implies that at least one of the vectors can be expressed as a linear combination of the other vectors, thus implying redundancy. If no such constant can be found, the set is said to be *linearly independent*.

The *rank* of a square Matrix A is defined as

$$\begin{aligned} rank(A) &= \text{number of linearly independent rows of A} \\ &= \text{number of linearly independent columns of A.} \end{aligned} \tag{2.6}$$

Thus, a matrix of order $(n \times n)$ with rank n shows no redundancies in the sense that none of its columns could be expressed by a combination of its other columns, while the same holds for its rows.

The *trace* of a square matrix A of order $(n \times n)$ equals the sum of its diagonal entries, denoted as

$$trace(A) = \sum_{1 \leq i \leq n} a_{i,i}. \tag{2.7}$$

For the *determinants* of square matrices A and B holds that

- $det(A) = det(A^T)$

- $det(AB) = det(A)det(B)$

Particularly, consider a square matrix A, for which a decomposition of the form $A = PLU$ exists, so that P is a permutation matrix, L a lower triangular matrix with diagonal entries $l_{i,i} = 1$ and U is a upper triangular matrix. Then the determinant of A equals the product of the diagonal entries of U, thus

$$det(A) = \prod_{1 \leq i \leq n} u_{i,i}. \tag{2.8}$$

This decomposition is generally referred to as LU decomposition. The above stated definitions have been derived from [CC80] and [Har08].

2.1.2 Eigenvectors and eigenvalues

In the following, we describe how eigenvectors and -values of matrices represent characteristic properties of their matrices. A matrix A is said to have an *eigenvector* γ with an *eigenvalue* λ if

$$A\ \gamma = \lambda\ \gamma. \tag{2.9}$$

Square matrices of order $(m \times m)$ can have up to m linearly independent eigenvectors $\gamma_i \in \mathbb{R}^m$ with such corresponding eigenvalues. Naturally, $A\ \gamma' = \lambda'\ \gamma'$, for an infinite number of vectors γ' that are linear dependent to γ, so that $\lambda\gamma = \lambda'\gamma'$. Thus, γ is often chosen with unit length, $||\gamma||_2 = 1$. For square symmetric matrices, the eigenvalues λ_i are real and their eigenvectors are mutual orthogonal to each other, forming a coordinate system.

Eigenvectors can be interpreted as vectors whose orientation does not change by the application of the belonging matrix in form of a transformation. The orientation therefore represents a characteristic feature of the matrix. The amount of scaling, given by the eigenvalues, represents a quantification of this characteristic feature. Figure 2.1 illustrates this interpretation.

The number of eigenvalues $\lambda_i \neq 0$ equals the rank of the matrix. An analogous definition of eigenvectors and eigenvalues is given in [FH08] by

$$A = \Gamma\ \Lambda\ \Gamma^T, \tag{2.10}$$

which is called the eigendecomposition of A, whereas Γ is the matrix of column wise arranged eigenvectors and Λ is the diagonal matrix of eigenvalues. Thus, for a $(m \times m)$ matrix A,

$$\Gamma = \quad (\gamma_1, ..., \gamma_m) \quad \text{and} \tag{2.11}$$

$$\Lambda = \quad (\lambda_i)_{1 \leq i \leq m} = \quad diag(\lambda_1,...,\lambda_m). \tag{2.12}$$

There is a relationship between the determinant, trace and eigenvalues of a square matrix A, which should be noted. The trace of a $(m \times m)$ matrix A equals the sum of its

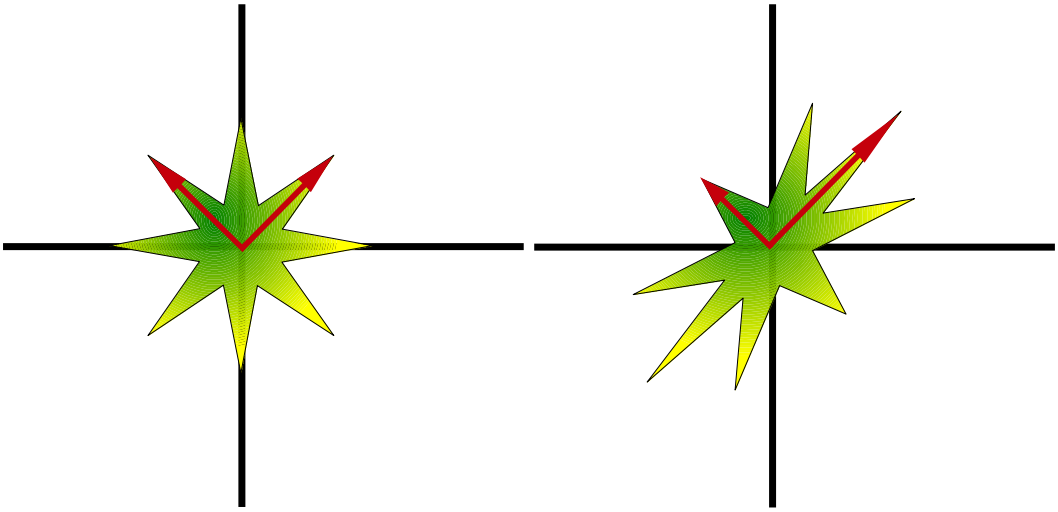


Figure 2.1: Geometric interpretation of eigenvectors as transformation invariant orientations, scaled by their corresponding eigenvalues.

eigenvalues,

$$trace(A) = \sum_{1 \leq i \leq m} \lambda_i, \tag{2.13}$$

whereas the determinant of A is given by the product of its eigenvalues,

$$det(A) = \prod_{1 \leq i \leq m} \lambda_i. \tag{2.14}$$

For a more detailed consideration, see [Har08].

2.1.3 Statistics

In the following, we introduce the essential techniques of descriptive statistics that are the basis of multivariate analysis. For more information, [Ren02] is recommended. We consider input data given in the form of a $(n \times m)$ matrix

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,m} \\ \cdots & \cdots & \cdots \\ x_{n,1} & \cdots & x_{n,m}, \end{pmatrix} \tag{2.15}$$

where each row represents one m-dimensional data element. The n data elements are considered as random samples with m variables. The $n \times m$ observations are object to

statistical analysis and description techniques. Throughout this work, we consider only continuous variables. We handle X as observation data and are only interested in the properties of the given multivariate samples, not their population. Thus, when referring to a variable's properties, we are considering the properties of the observations of this variable throughout all samples.

Given n m-dimensional data elements by a $(n \times m)$ matrix X, we consider n samples, each containing m observations (scalar values). We define the (sample) *mean of a variable* y_j, $1 \leq j \leq m$ as the arithmetic average of n observations $x_{1,j}, ..., x_{n,j}$, given by

$$\overline{y_j} = \frac{1}{n} \sum_{1 \leq i \leq n} x_{i,j}. \tag{2.16}$$

We denote the *mean of* X as the m-dimensional point

$$\overline{X} = (\overline{y_1}, ..., \overline{y_m}), \tag{2.17}$$

having the variable means in each corresponding coordinate. It represents the geometric center position of the n points in m-dimensional space. Similarly, $min(X)$ is referred to as the geometric minimum point $(min(y_1), ..., min(y_m))$.

Furthermore, a variable y_j is said to be *centered* if

$$\sum_{1 \leq i \leq n} (x_{i,j} - \overline{y_j}) = 0 \tag{2.18}$$

and analogously, X is said to be centered if all its variables are centered.

The *standard deviation* s_{y_j} of a variable y_j, $1 \leq j \leq m$, is defined as

$$s_{y_j} = \sqrt{\frac{\sum_{1 \leq i \leq n} (x_{i,j} - \overline{y_j})^2}{n}}. \tag{2.19}$$

It gives a measure of deviation, regarding the arithmetic mean distance over all the observations of a variable to the variable's mean.

A similar descriptor is given by the variance s_j^2 of a variable y_j, $1 \leq j \leq m$, defined as

$$s_{y_j}^2 = \frac{\sum_{1 \leq i \leq n} (x_{i,j} - \overline{y_j})^2}{n}. \tag{2.20}$$

It is the square of the standard deviation and therefore gives no exact distance measure.

The above listed measures of spread are concerned with a single variable. For multivariate analysis, it is often important to investigate the relationship between two variates. *Covariance* is a measure of how one variable's variance relates to another variable's variance.

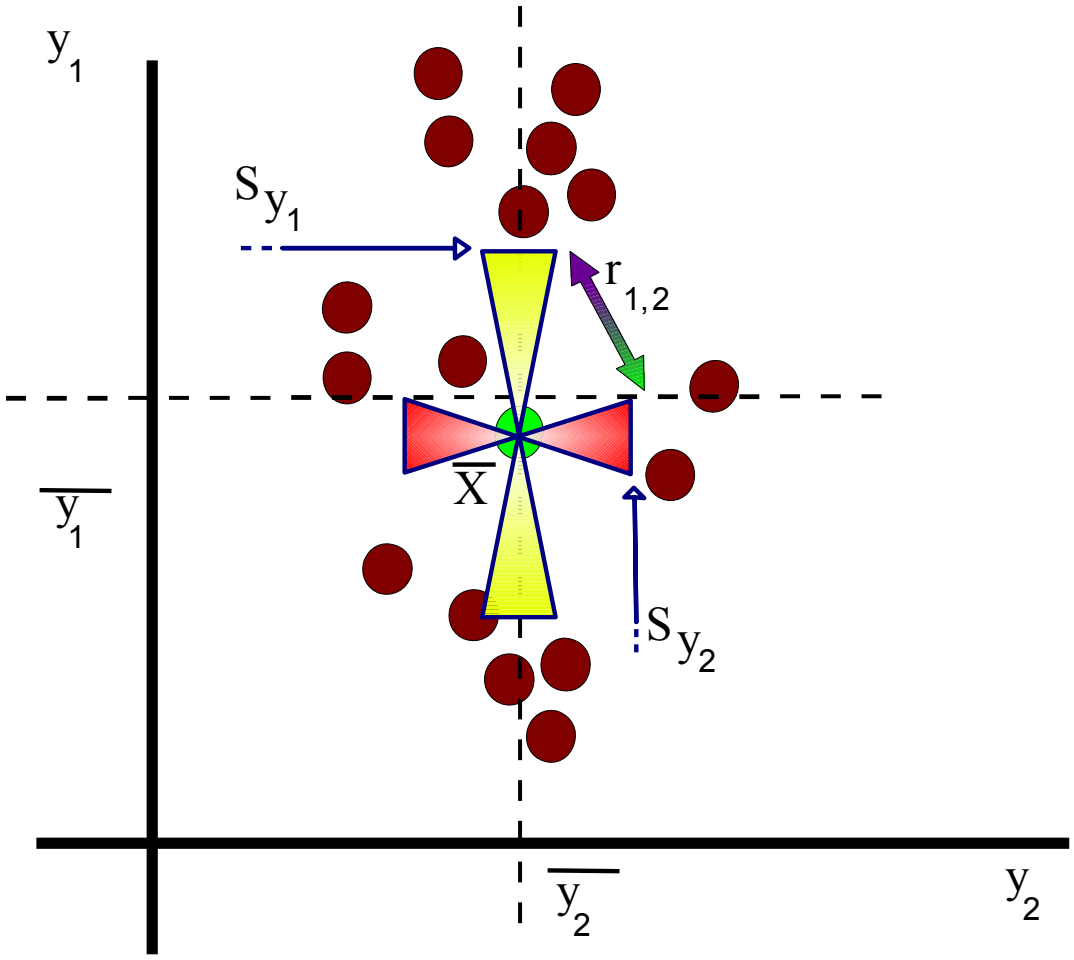


Figure 2.2: Geometric interpretation of the statistical descriptors. The data's mean $\overline{X}$, its variables' mean $\overline{y_1}$ and $\overline{y_2}$, their standard deviation s_{y_1} and s_{y_2}, as well as their correlation $r_{1,2}$ is displayed.

The covariance $s_{i,j}$ of two variables y_i and y_j, $1 \leq i,j \leq m$, is defined as

$$s_{i,j} = \frac{\sum_{1 \leq k \leq n} (x_{k,i} - \overline{y_i})(x_{k,j} - \overline{y_j})}{n}. \tag{2.21}$$

Note that the covariance of two equal variables is its standard deviation, thus $s_{i,i} = s_{y_i}^2$.

With all m variables, we form a covariance matrix S of order $(m \times m)$ with entries $s_{i,j}$ as covariances for each combination of variables. For a centered $(n \times m)$ data matrix X, we find that

$$S = \frac{1}{n} X^T X. \tag{2.22}$$

Note that covariance is translation invariant regarding different means. However, it is scale dependent regarding different spreads. If one is interested in a measure of linear relationship between two variables, that is scale invariant, the *correlation* between them is a better descriptor. The correlation $r_{i,j}$ of two variables y_i and y_j is obtained by dividing the variables' standard deviations s_i and s_j from the covariance $s_{i,j}$, thus

$$r_{i,j} = \frac{s_{i,j}}{s_i \; s_j}. \tag{2.23}$$

In this regard, correlation can be seen as a standardized covariance, where a correlation of 1 indicates full correlation, 0 none and -1 opposed correlation.

However, if all variables share the same scale, covariance gives a sufficient comparison of the linear relationship between two variables. Consider two n-dimensional vectors v_i and v_j, accounting for the n observations in the centered variables y_i and y_j, i.e. $v_i = (x_{1,i} - \overline{y_i}, ..., x_{n,i} - \overline{y_i})$, v_j analogous. Let σ be the smaller angle between them. We find that the cosine of σ equals the correlation of y_i and y_j [Ren02], since

$$r_{i,j} = cos \; \sigma = \frac{v_i^T v_i + v_j^T v_j - (v_j - v_i)^T (v_j - v_i)}{2\sqrt{(v_i^T v_i) \; (v_j^T v_j)}} = \frac{v_i^T v_j}{\sqrt{v_i^T v_i} \; \sqrt{v_j^T v_j}} \tag{2.24}$$

Hence, if v_i and v_j are perpendicular to each other, the cosine of the angle between them is zero. Therefore, y_i and y_j show no correlation in the data. To illustrate a zero-correlation, consider that

$$v_i^T \; v_j = \sum_{1 \leq k \leq n} x_{k,i} \; x_{k,j} = 0$$

follows from their orthogonality and the above sum being zero states, that for any observation $x_{k,i} \neq 0$, $x_{k,j} = 0$ and vice versa.

2.2 Visual approaches

This section addresses the visual representation of multivariate data. Ideally, such a representation allows the viewer to visually identify all coordinates for any number of dimension and for any number of elements. This task is most likely infeasible. Instead, we discuss different approaches their corresponding foci.

Most visualization techniques focus on one of two things - being feasible to visualize a high number of relatively low dimensional elements, or being able to account for a lower number of very high dimensional elements. One of the main issues in these techniques is deriving a good representation for coordinates. Since physical display devices can only facilitate a maximum of three dimensions, one can no longer represent the coordinates of higher-dimensional objects by its position within a scene. Instead, one has to derive a new representation that accounts for all possible coordinates of an object for any number of dimensions. This is a typical problem of information visualization, where one is concerned with finding meaningful representations for non-scientific data. On the other hand, most multidimensional data is scientific. This work is concerned with continuous real

multidimensional data resulting from scientific measurements. The data is interpretable by metrics and describable by statistical or analytical methods.

The aspects involving the visualization of multidimensional data could therefore be separated into the representation of multidimensional objects and the appliance of analytical methods to simplify, summarize and abstract, as well as the supply of a framework that allows an interactive, exploratory data investigation. In this regard, multidimensional visualizations may be seen as part of visual analytics or exploratory data analysis. This section introduces the most common visualization techniques for multidimensional data, discusses their advantages and drawbacks, and provides further references.

Although many techniques apply to categorical data as well, we focus on visualization techniques for tables of numerical data. These so called table visualizations have been categorized by several surveys, e.g. [HG97], [GTC01], or [AdO04], which represent the main sources of this section. Ward et al. have also given an excellent summary of existing techniques in this field [WGK10].

2.2.1 Heatmaps and Multiple Linegraphs

Heatmaps and Multiple Linegraphs may be considered as one of the most basic table visualizations. Given a table of n m-dimensional numerical data, these visualization techniques represent the data through a substitution of the numerical values by colors and line segments, respectively.

Line Graphs are commonly used to display piecewise continuous functions of one variable. Multiple Line Graphs gives the name to the technique of mapping m functions over each

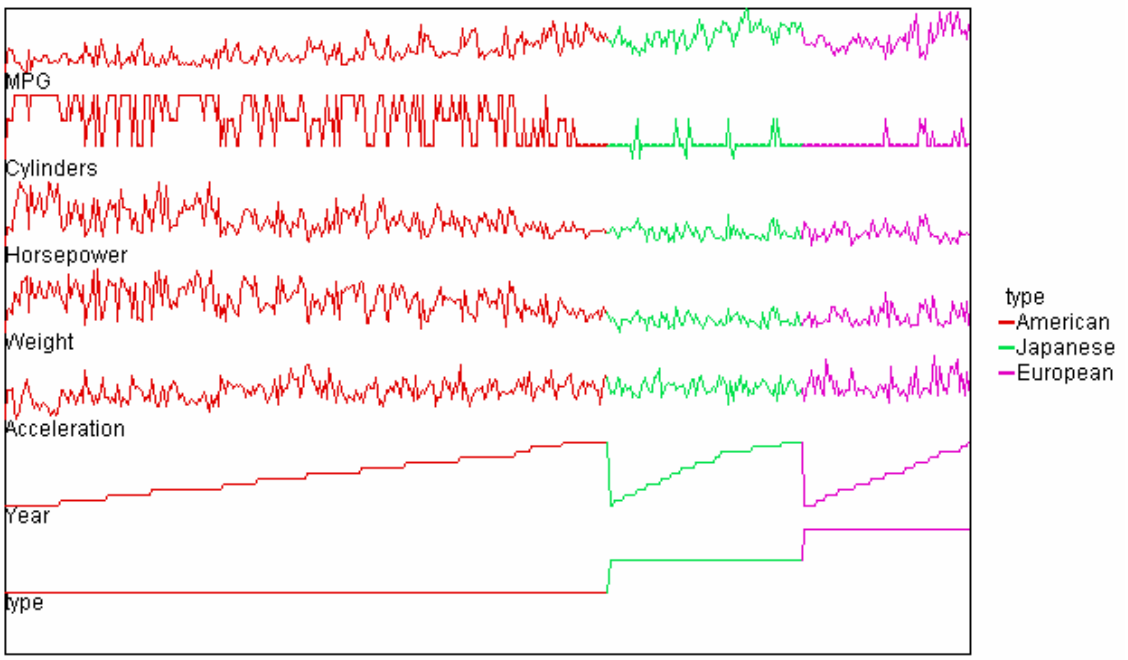


Figure 2.3: Multiple Line Graphs visualization of the ASA car specs data set. Data values are visualized as piecewise continuous functions per variable. Image courtesy of [GTC01].

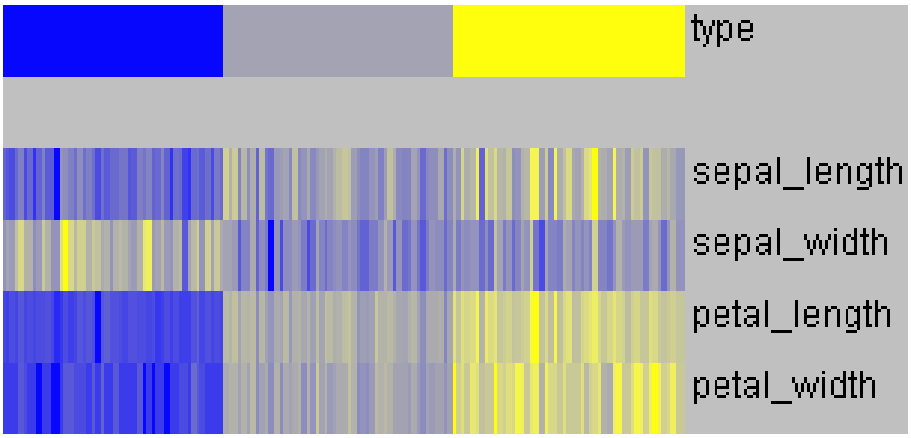


Figure 2.4: Heatmap visualization of the iris flower data set. This technique encodes data values in color space, while keeping the basic layout of a table. Image courtesy of [GTC01].

corresponding column, representing the observations in this variable as connected line segments. Each line graph is laid out horizontally with the discretized dimension over the number of observations and with the vertical continuous dimension (scaled to a specific height) for the mapping of the observations' numerical values. The high familiarity with this data representation accounts for an intuitive readability.

Instead of using the vertical dimension for the plotting of numerical values, Heatmaps use color as their dimension to represent numerical values. Numerous possible color mappings can be chosen to represent values by colors. This technique can be considered as one of the first information visualization techniques in computer graphics and is still used extensively because of its intuitiveness and efficient pattern recognition abilities. Several extensions have been developed, such as Survey Plots [Loh95] and Table Lens [RC94]. Their efficiency can be greatly improved by a meaningful ordering of the columns and rows.

Since lines and colors are a more feasible form for human perception than numerical values, these techniques enhance the readability of numerical data tables. The average viewer is more likely to recognize patterns in these pictures compared to reading the original table of numerical values. However, for many dimensions and elements, the representation takes up a lot of space and comparison becomes harder. While the overall value distribution in each variable may be displayed in an intuitive fashion, further analysis of the data may find its limits for this representation.

2.2.2 Glyphs and Icongraphic displays

Glyphs and icons are common techniques in information visualization. Usually each object is designed to express a number of features equal to the dimension of the data and represents a single data element. One of the most famous examples are the Chernoff faces [Che73], where data dimensions are mapped to facial features. These features are facial attributes like angles, width or length of eyes, nose, mouth or eyebrows.

The human ability to precisely distinguish an enormously broad range of facial characteristics make this technique potentially very suitable for encoding very high dimensional

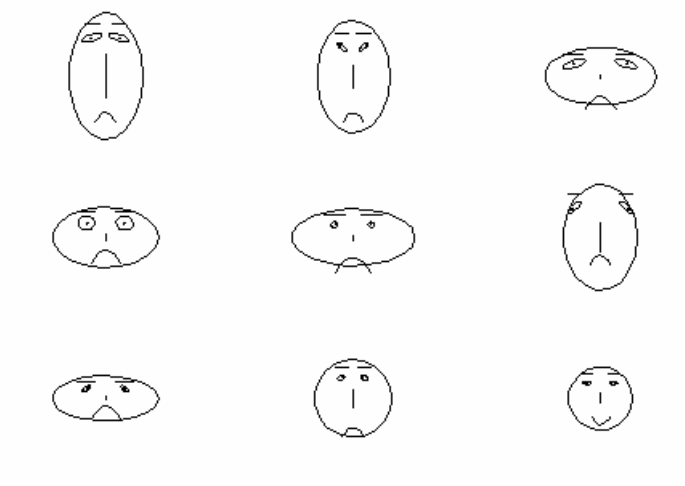


Figure 2.5: Exemplary Chernoff Faces. Data values are encoded in attributes of facial features, such as mouth width, each corresponding to a dimension. Image courtesy of [GTC01].

data points in an extremely low space-taking representation.

Another type of glyph and icon visualization encodes distinctive patterns in a texture that is to represent patterns in the data. Similar to a Chernoff face, a single icon, also known as a stick figure, shows a distinct layout for a feature set and represents one data element [PG88]. These icons are then packed together in a dense display, thus mapping the dimension values of the data onto features of the resulting texture.

Other techniques include Recursive Patterns [KAK95] and Dimensional Stacking [WLT94], where dimensions are embedded within other dimensions. By these embeddings, the re-

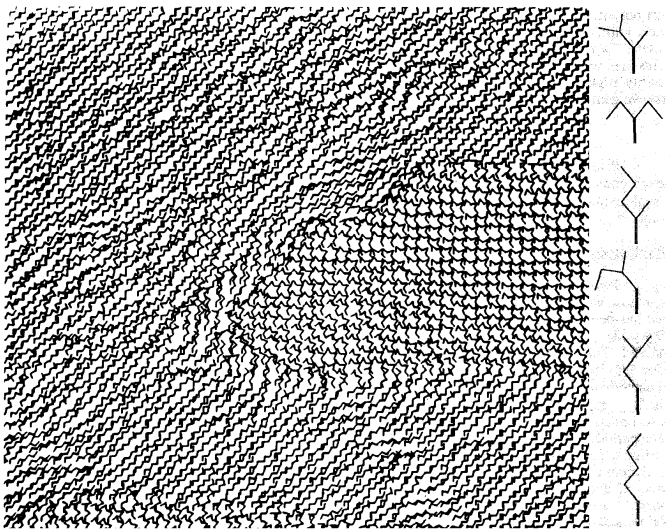


Figure 2.6: Icongraphic displays encode data values in icon features, thereby, mapping patterns in the data to patterns of the resulting texture. Image courtesy of [PG88].

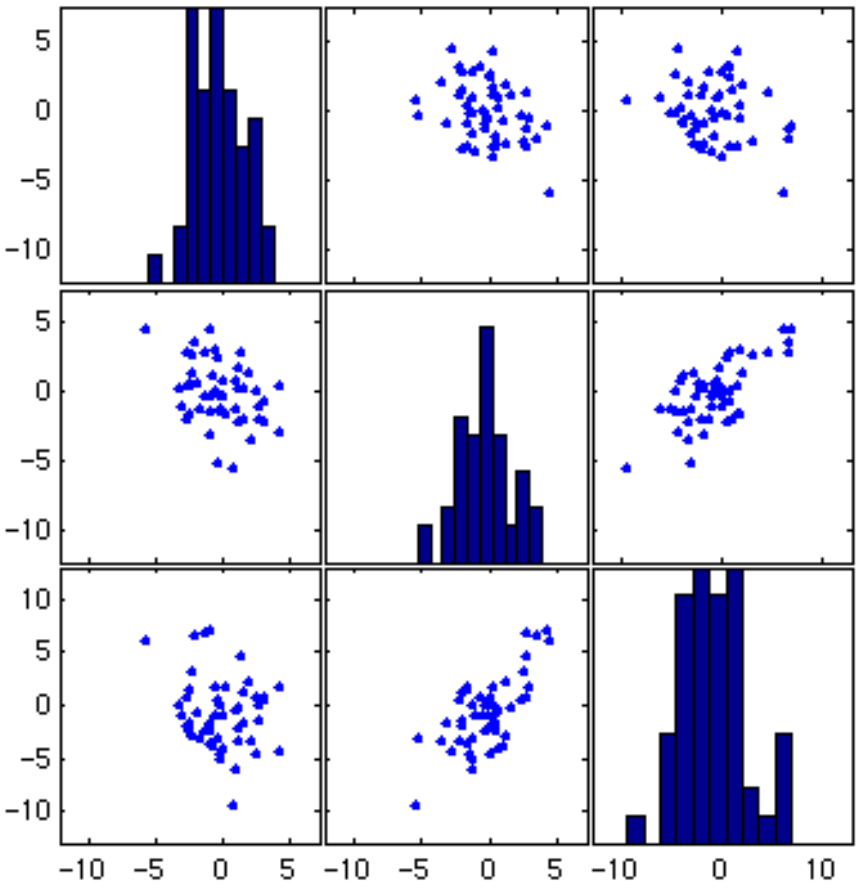


Figure 2.7: Scatter plot matrices show projections of the data along the data dimensions. The diagonal shows the distribution of each dimension. Image courtesy of [GTC01].

sulting displays encode an imposed hierarchical structure of the data. Whereas differences between data points can be recognized very easily by the means of these visualization techniques, however, the quantification of the original coordinates by the viewer usually faces a problem.

2.2.3 Scatter plot matrices

The well-known 2D scatter plot is probably the most commonly utilized visualization technique for 2D points in Euclidean space. Mappings and transformations can freely be applied to the point projection. When more (non-spatial) attributes are associated with each point, one usually encodes these by the display of objects (instead of points) with an equal number of associated attributes, e.g., color, size, or shape.

One of the earliest techniques for displaying m-dimensional data was the scatter plot matrix. This $(m \times m)$ matrix of 2D scatter plots contains $m(m-1)/2$ scatter plots, each plotting one pairwise combination of dimensions. Over time, numerous variants and extensions have evolved. Often, diagonal plots show the data's distribution in the specific

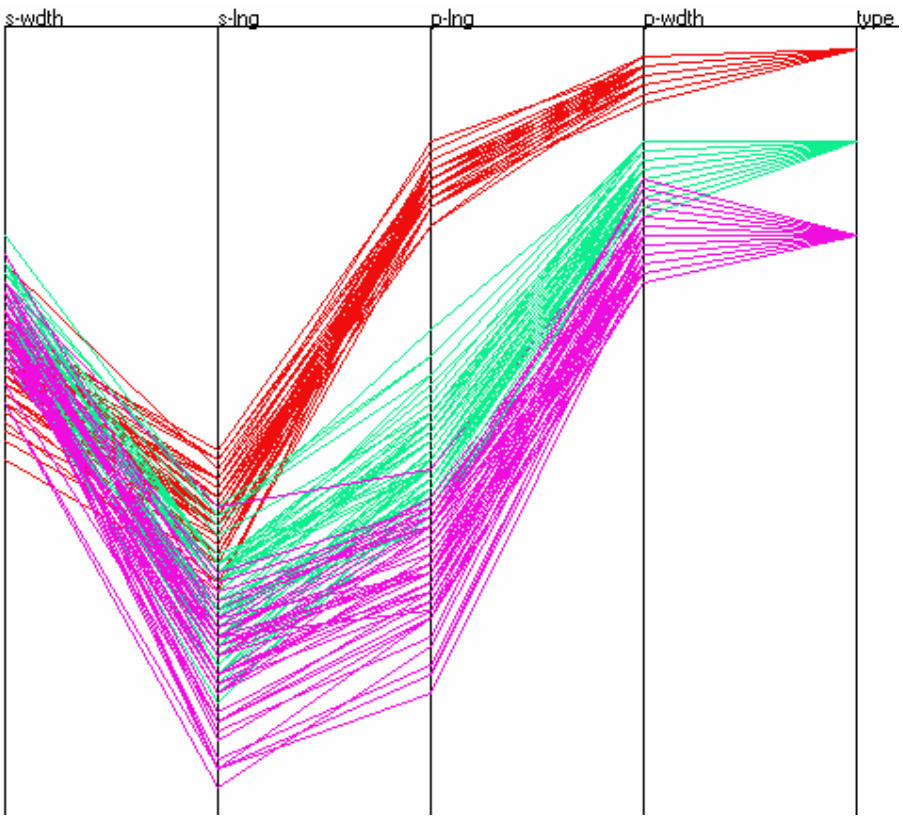


Figure 2.8: Parallel Coordinates visualization of the iris flower data set. Data points are visualized as piece-wise linear functions, thereby, revealing patterns and outliers quickly. Image courtesy of [GTC01].

dimension, for example by a histogram. Other variants are, for example, hyperbox [AC91] or prosection [STDS95].

Techniques that allow the visual linking of features from one plot to another greatly improve the readability of the plots, as well as analytical processes. It is such interactive techniques, which make a scatter plot matrix suitable for data mining. However, without such linking and highlighting enhancements, the scatter plots become hard to compare and one might quickly lose focus in analysis.

2.2.4 Parallel Coordinates

Inselberg's *Parallel Coordinates* [ID90] belongs to the most famous visualization techniques in information visualization. The core idea is to represent multidimensional points by the means of a projective coordinate system, in which uniform coordinate axis for each dimension are placed in parallel to each other. Points are then represented by polygonal line chains, which intersect the parallel axis, mapping the point's coordinate in the respective dimension.

More formally, the coordinate system constructed for the representation of m-dimensional numerical data, can be described as follows. On the two-dimensional projective xy-plane

(titled as P^m), m lines, so called parallel coordinate axes $y_1, ..., y_m$, are placed sequentially along positive x direction, in parallel, equidistant and perpendicular to the x axis. In addition, the parallel axes have the same positive orientation than the y axis and y_0, as the first axis of the sequence, is defined by $x = 0$. With d being the distance between the parallel axes, a point $p \in \mathbb{R}^m$ is represented by the polygonal line l having the vertices $(d(i-1), p_i)$, for $1 \leq i \leq m$, in xy-coordinates.

Inselberg claims that the value of his technique can only be fully appreciated, when the dualities between the projective parallel space P^2 and the Euclidean space $\mathbb{R}^2$ are considered. A good summary of the dualities between P^2 and $\mathbb{R}^2$ is given by [Hin88], [Ins85] and [ID94], as described in the following.

The *fundamental duality* is given by the fact that points in $\mathbb{R}^2$ map into lines in P^2, whereas lines in $\mathbb{R}^2$ map into points in P^2. The latter observation is given by the fact that a set of infinitely close points, forming a line in $\mathbb{R}^2$, map into a set of infinitely close lines, intersecting in a single point in the projection plane of P^2.

This relationship can be explicitly expressed by conveying the definition of the implicit line formed by the points in $\mathbb{R}^2$ directly to the point's coordinates in P^2. For a line $l : y = mx + b$ in $\mathbb{R}^2$, with $m < \infty$, the points on l form an infinite family of lines, which

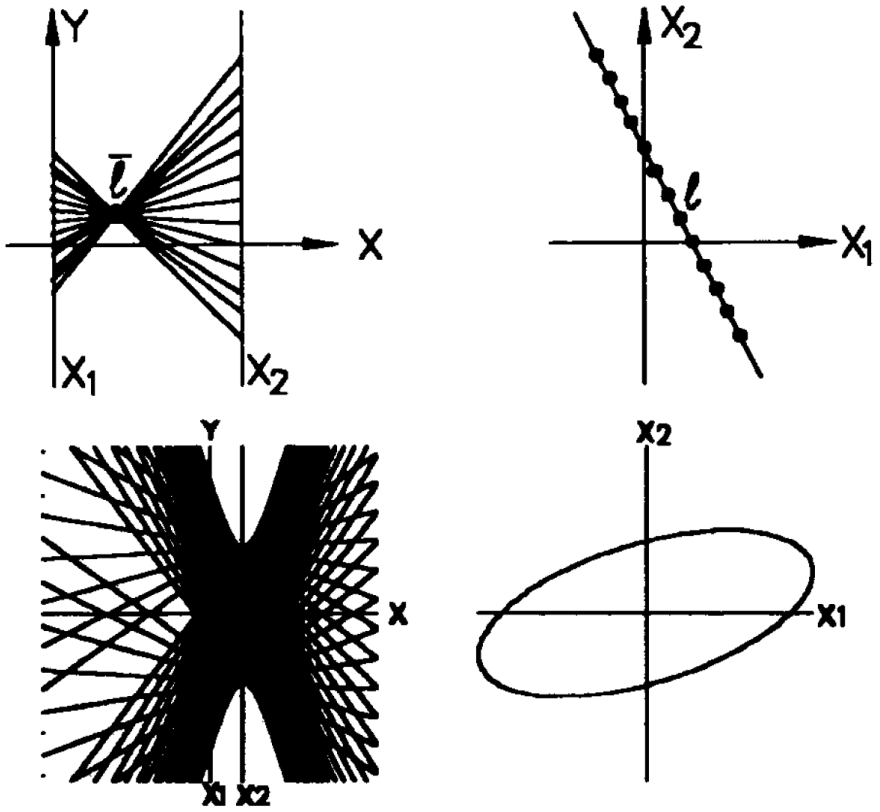


Figure 2.9: Dualities of Parallel and Euclidean space: A (point-) line is mapped to a (line-) point and a (point-) curve is mapped to a (line-) curve. Image courtesy of [ID94].

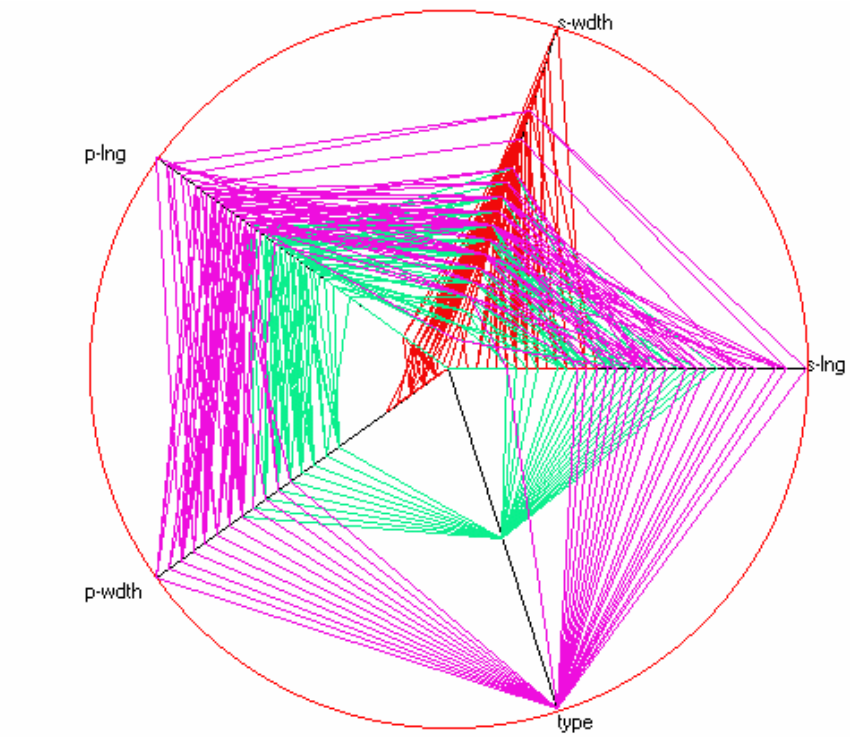


Figure 2.10: Star Plot visualization of the iris flower data set. Data is mapped on closed piece-wise line segments along circularly arranged coordinate axes. Image courtesy of [GTC01].

intersect in the point $(d/(1-m), b/(1-m))$, for $m \neq 1$, in P^2. Furthermore, vertical lines $l : y = c$ are mapped to $(0,c)$. Whereas, for lines with $m = 1$, the notion of an *ideal point* is derived, for which the corresponding parallel lines intersect at the slope $y/x = b$. Other dualities include that a conic (point) curve in $\mathbb{R}^2$ is mapped to a conic line curve in P^2, defined by the lines representing tangents to the curve. For example, an ellipse in $\mathbb{R}^2$ is mapped to an hyperbola in P^2, as shown in figure 2.9 [ID90].

Transformations also show interesting dualities. Rotations in $\mathbb{R}^2$ map to translations in P^2 and vice versa, making a transformation, that is visually hard to detect in $\mathbb{R}^2$, easier to detect in P^2. For a detailed discussion of these geometric dualities as well as the consideration of multidimensional space, the reader is referred to [Ins09a]. These interesting geometric properties of P^m, together with the fact that the data presentation with its uniform treatment of dimensions, is very intuitive and easy to understand, have made this technique very successful. However, some major drawbacks should be mentioned. The needed display space to represent multidimensional data increases linearly to the number of dimensions. More profoundly, for a sizable number of data elements, the line chains become increasingly visually cluttered and hard to compare. Furthermore, its effectiveness for detecting features in the data is highly dependent on the dimension ordering, i.e. the sequential ordering of the parallel axis.

However, over the past decades, many advances in the visual representation have been made, especially in enhancing cluster detection and selection highlighting. These techniques

well exceed the common alpha blending and edge-bundling, as well as the limits of this work. Examples include [HLD02], [HW09], [YGX*09], [AdO04], [MW02] or [MM08].
Closely related, though not quite as famous, is a technique named *Circular Parallel Coordinates* [HG97], also known as Star Plots. Although they can be considered as glyphs, they are introduced in the following to highlight the similarity to dimensional anchor visualizations, discussed in the next section.

As the name suggests, circularly arranged coordinate axes are utilized to represent multidimensional data points. The analogy of a Star Plot derives from the presentation of a point in $\mathbb{R}^m_+$ by a closed polygonal line chain, which may look like a star for some points. Analogous to Parallel Coordinates, the polygonal line intersects each dimension axes at their corresponding coordinate value. The technique is restricted to positive real data, as most representations that use circularly arranged dimension axes. Similar to Parallel Coordinates, Star Plots become cluttered for many data elements and its effectiveness is dependent on the arrangement of the dimensions.

2.2.5 Dimensional Anchor visualizations

Inspired by Parallel Coordinates, many other techniques have been developed that treat dimensions uniformly and represent them in form of a line or a direction in space. The following techniques can be categorized as Dimensional Anchor visualizations [HGP99], whereas a dimensional anchor corresponds to a single dimension. The anchor is usually represented by a line, whereas its position, length and orientation play an essential part for the data projection. Often, such anchors are circularly arranged position vectors, which makes the visualization one of the most compact regarding the needed display space.

Dimensional Anchor Visualizations typically project a multidimensional data element as a single point onto (usually two-dimensional) space, whereas the point's position within the projection hints to its coordinates. This high abstraction level has the benefit that the representation can account for many data elements without cluttering the display. However, since the point positions in such a projection are highly ambiguous, the visualization can lead to misinterpretations of the data. The amount of enlightenment that derives through this data display is highly dependent on the effectiveness of the chosen projection.

Star Coordinates [Kan00] are associated with this category of visualizations. They feature a basic, yet intuitive projection technique. Each dimension is represented by circularly arranged position vector, a dimensional anchor. The projection of an m-dimensional data element is found by the linear combination of these vectors multiplied by the numerical values in the corresponding dimensions. Star Coordinates offer simple but very powerful interaction and the possibility for enhancement. The technique is formally introduced in a later chapter.

RadViz [HG97] (Radial Coordinate Visualization), also belongs into the category of dimensional anchor visualizations. RadViz uses spring constants to find the position of the projected data points. Illustratively, each data point has m springs attached to it, each located at the end of the corresponding dimensional anchor. The springs' force equals its coordinate in the dimension. The data point is projected to the position that has a sum of zero total spring forces.

PolyViz [HG97], as well as its extension XRadViz [HGP99], are similar techniques. Here

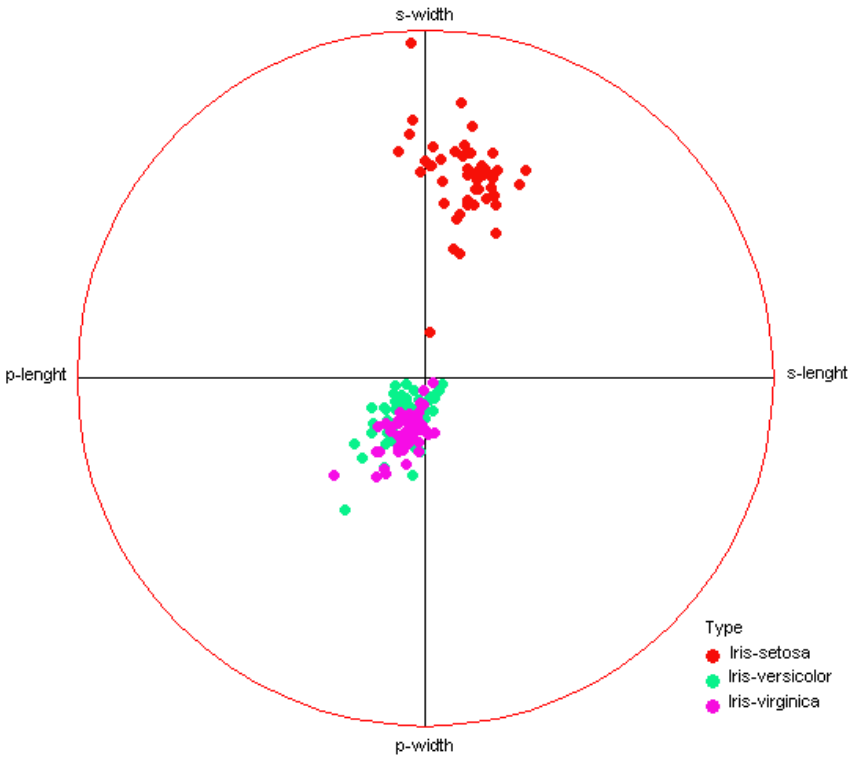


Figure 2.11: RadViz visualization of the iris flower data set. Data points are projected using spring forces from anchors arranged on a circle. Image courtesy of [HG97].

the anchors are arranged as line segments of a polygon, showing the data's distribution in the dimension along each line segment. However, the data elements are projected as in RadViz.

2.2.6 Other techniques

Projection Pursuit [GTC01] is another interesting technique. As the name would already suggest, the purpose of Projection Pursuit lies in finding good data projections that account for sought-after data properties. The user may choose, possibly with the help of analytical methods, to define a curve in multidimensional space that is of particular interest. The technique then generates a sequence of projections by moving a projection plane along this curve.

A quite different projection technique is a method called Andrew's Curves [And72]. Similarly to a Fourier transformation of a data point, the method plots a curved line for each data point $(x_1, ..., x_m)$, using the following function:

$$f(t) = \frac{x_1}{\sqrt{2}} + \frac{x_2}{sin(t)} + \frac{x_3}{cos(t)} + \frac{x_4}{sin(2t)} + \frac{x_5}{cos(2t)} + ...$$

Plotted for in the interval $[-\Pi, \Pi]$, f can represent many dimensions and highlight clusters within the data in a visually appealing way. However, for large data sets, computation

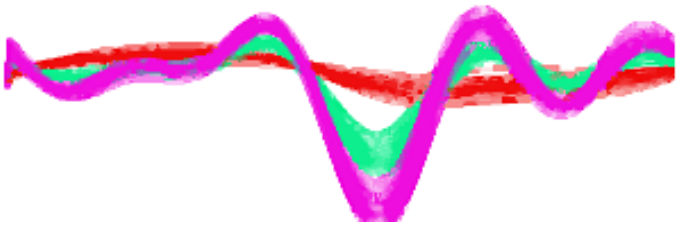


Figure 2.12: Andrew's Curve visualization of the iris flower data set. Data points are visualized by characteristic curves that reflects features. Image courtesy of [HG97].

time can be a disadvantage of this technique.

To allude to a visualization that is not based on a projection in Euclidean space, Hyperbolic Multi-Dimensional Scaling [WR02] utilizes the hyperbolic plane to incorporate focus and context into the projection. Using the circular Poincaré model, under which the H2 has an exponential growth of length and area toward the disc's boundary, they formulate a projection method that applies Sammon's MDS to H2.

2.2.7 Comparison and observations

For all presented representations, it can be concluded that no one method can be recommended above all others. All visualizations are scale dependent, meaning that they need standardized data to achieve their full effectiveness.

For a small number of elements in a high number of dimensions, glyphs are a rather intuitive and effective representation. However, their efficiency is dependent on the ordering of the dimensions and to the mapping of dimensions to the glyph's features.

For a small number of dimensions, scatter plot matrices can represent a high number of elements in an intuitive way and give a good overview on the data, although their analytical abilities are somewhat limited. Also, they tend to use up a lot of display space, as do heatmaps and line graphs.

Dimensional anchor visualizations and parallel coordinates also give a good overview.

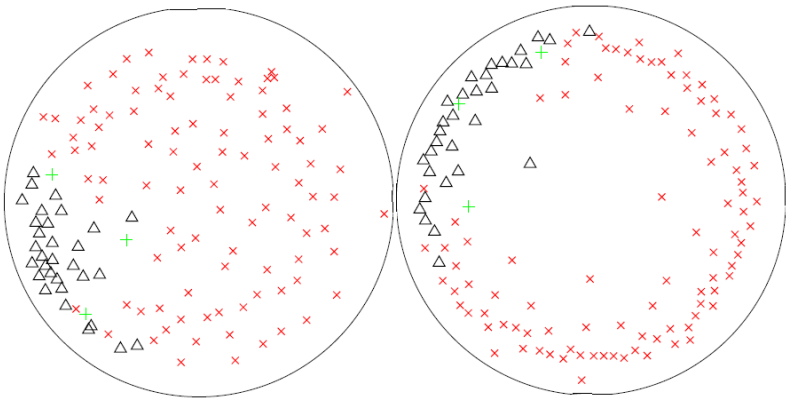


Figure 2.13: Hyperbolic plane visualization of the iris flower data set. Data points are projected on the hyperbolic plane, providing focus and context. Image courtesy of [WR02].

While the latter displays all data without abstraction, the former abstracts the data values by a point projection. This leads to reduced visual clutter but also to inherent ambiguity with respect to coordinate values. However, both techniques show the most potential for improvements and many interaction techniques are possible within these displays.

Interaction is the key fundament of exploratory data visualization. The cognitive admission and comparison of data patterns is greatly improved by interaction techniques like linking, brushing, browsing, zooming and highlighting. In addition to interactive means for exploration, most state-of-the-art techniques facilitate or directly incorporate analytical methods in order to work effectively, for example, a good (starting) projection or dimension ordering. Basic analytical concepts are introduced in the following.

2.3 Analytical approaches

A number of techniques have been introduced that aid in the visual representation of multidimensional data, as it has been given in raw form. However, with increasing complexity of the data and the given limitations of available display space, it is often inevitable to make compromises regarding the information that can be reasonably displayed.

Often, the question arises on what the focus of visualization for a specific data set should be. In order to derive such a focus, the given information content has to undergo a selection that reflects some measure of importance for the user. To do so, the data has to be analyzed and possible sources of interest have to be discovered. This section provides a small insight in techniques that derive such knowledge from data and it is fair to say that knowledge about the underlying data, if utilized accordingly, may aid in numerous aspects of its visual representation.

Stating the precise boundaries between data mining and knowledge discovery is a difficult task. The interdisciplinary context of knowledge discovery makes most techniques hard to assign to a single field. Data mining involves the analysis of observational data sets with the intent of discovering relationships within the data and the summary of these in an understandable and useful way for the user [HSM01]. The output formed by these techniques is often called patterns or features of the data set and generally specify a mathematical model that is of interest to the investigator. These models may be classified in the two categories, descriptive and predictive [Han05]. While descriptive models characterize observed properties of the data, predictive models describe the inference that is drawn from the data.

The main concepts of data mining cover data processing, -characterization, -discrimination, -clustering, -regression, -classification, as well as the visual representation circumscribed in the fields of exploratory data analysis and model visualization [dOL03], [Han05]. Data preprocessing techniques include data cleaning, -reduction and -filtering. Filtering and data reduction may also encompass outlier analysis, which purpose is to find objects that do not comply with the general behavior of the data set. It can be seen as noise reduction in this context. For more information about data mining techniques, the interested reader will find, for example, [Han05] or [HSM01], to give a more comprehensive introduction. The described content in this section is, to a big extend, summarized from the work of Everitt et al. [ELL01], which can be recommended as an excellent reference and can serve as a source for further information about this topic.

2.3.1 Metrics

An issue of central importance for all analytic techniques is the comparison of objects. This section illuminates on methods for quantifying how "close" objects are together, i.e. how (dis-)similar they are. Common transformations between similarities s and dissimilarities d include:

- $d = 1 - s$
- $d = c - s$, for some constant c
- $d = \sqrt{2(1-s)}$

While some metrics measure similarity and others dissimilarity, we abstract these notations where possible by using the single term proximity δ. Further, We only consider measures for continuous real data, although some of these techniques may apply to categorical data as well. In the following, we discuss measures for the proximity between objects and between groups.

Measures of proximity

Designing a generally applicable metric is an immense challenge. Metrics are almost always data dependent and should be tailored for every application to fulfill the right needs, i.e. provide the right quantification of proximity. However, there are a range of generally applicable proximity measures for multivariate elements. Most of these are distance metrics, which fulfill the triangular inequality

$$\delta_{i,j} + \delta_{j,k} \geq \delta_{i,k}, \tag{2.25}$$

with δ being the distance measure for objects i, j and k.

Common inter-individual proximity measures $\delta_{i,j}$, deriving from a $(n \times m)$ data matrix X include [ELL01]:

$$\textit{Manhattan distance} \qquad \delta_{i,j} = \sum_{1 \leq p \leq m} |x_{i,p} - x_{j,p}| \tag{2.26}$$

$$\textit{Euclidean distance} \qquad \delta_{i,j} = \sqrt{\sum_{1 \leq p \leq m} (x_{i,p} - x_{j,p})^2} \tag{2.27}$$

$$\textit{Minkowski distance} \qquad \delta_{i,j} = \Big(\sum_{1 \leq p \leq m} |x_{i,p} - x_{j,p}|^r \Big)^{\frac{1}{r}} \ , \ r \geq 1 \tag{2.28}$$

$$\textit{Angular separation} \qquad \phi_{i,j} = \frac{\sum_{1 \leq p \leq m} x_{i,p} x_{j,p}}{\sqrt{\sum_{1 \leq p \leq m} x_{i,p}^2} \sqrt{\sum_{1 \leq p \leq m} x_{j,p}^2}} \tag{2.29}$$

$$\textit{Pearson correlation} \qquad \phi_{i,j} = \frac{\sum_{1 \leq p \leq m} (x_{i,p} - \overline{x_{i,\bullet}})(x_{j,p} - \overline{x_{j,\bullet}})}{\sqrt{\sum_{1 \leq p \leq m} (x_{i,p} - \overline{x_{i,\bullet}})^2} \sqrt{\sum_{1 \leq p \leq m} (x_{j,p} - \overline{x_{j,\bullet}})^2}} \ , \quad \text{with } \overline{x_{i,\bullet}} = \frac{1}{p} \sum_{1 \leq p \leq m} x_{i,p} \tag{2.30}$$

Most commonly used is the *Euclidean distance*, which quantifies proximity of multivariate data elements by interpreting each variable as a dimension in Euclidean space and measuring the physical distance between the objects within this space. As one might guess, this is not always ideal, since not all variables are equally scaled or should be treated equally.

Another popular measure of distance is the *Manhattan metric*, also called "city-block" or "taxi cab" metric. Motivated by the Manhattan city blocks, its distance measure equals the length of the path between two locations within such a street configuration. It serves as an excellent example for metrics in general, since it gives a measure of proximity, subject to a specific interpretation of space, which is determined by the application. Both the Manhattan (l_1 *norm*) and the Euclidean distance (l_2 *norm*) are special cases of the general *Minkowski distance*, also known as l_r *norm*.

Distance measures have an inherent interpretation of their correspondent space, through which they derive their measure of distance. However, there are other ways to interpret the proximity of two objects, which are not related to a spatial norm.

The above noted correlation coefficients suggest that the correlation between two objects may be used to quantify their relationship. These coefficients $\phi_{i,j}$ indicate a correlation varying from [−1,1] and may be transformed into a dissimilarity measure by $\delta_{i,j} = (1 - \phi_{i,j})/2$.

The *angular separation* coefficient is often employed in specific applications, in which variables are of the same scale and when the investigator is not interested in absolute sizes of samples, but in its (relative) composition. The name derives from the fact that this measure equals the cosine of the angle between the two m-dimensional sample vectors. Since the angle $\angle(x_{i,\bullet}^T, x_{j,\bullet}^T)$ is independent of the length of the vectors, as well as in which orientation they should differ, it is a measure of how close the relative composition of one sample is to the composition of another sample. For example, perpendicular vectors indicate zero correlation. Often, all samples $x_{i,\bullet}$ are normalized, when the investigator is only interested in the composition of the samples. The angular separation coefficient is then equal to the *dot product* of two vectors, given by $\sum_{1 \leq p \leq m} x_{i,p} x_{j,p}$.

Pearson's correlation coefficient differs from the angular separation in the employment of the arithmetic mean ($\overline{x_{i,\bullet}}$) for the m observations in the a sample. This differs severely from the common correlation between variables ($r_{p,q}$), where the variable's mean (over all samples) is employed to each corresponding observation. In this regard, correlation between objects tries to standardize the samples, not the variables. Therefore, it considers the angle between two m-dimensional vectors, starting at the mean of the observations.

It should be noted that correlation coefficients are to some extend contested in literature, since they are problematic when variables have been measured in different scales. For more information, see [ELL01].

Measures of inter-group proximity

With a given proximity metric, defined for pairs of multivariate elements, this section considers methods for evaluating the proximity between groups of elements.

There are two major ways for deriving such inter-group proximities. They are based on either inter-individual proximities or group summaries. The set of inter-individual

proximities for two groups, $G_1, G_2 \subset \mathbb{R}^m$, entitles the set of all proximities $\delta_{i,j}$ with $i \in G_1$ and $j \in G_2$. These values are often stored in a symmetric ($|G_1| \times |G_2|$) proximity matrix Δ. Given Δ, common interpretations of inter-group distances δ_{G_1,G_2} include

$$\textit{nearest neighborhood distance} = \min_{1 \leq i \leq |G_1|, 1 \leq j \leq |G_2|} (\delta_{i,j}) \tag{2.31}$$

$$\textit{furthest neighborhood distance} = \max_{1 \leq i \leq |G_1|, 1 \leq j \leq |G_2|} (\delta_{i,j}) \tag{2.32}$$

$$\textit{average neighborhood distance} = \frac{1}{|G_1| + |G_2|} \sum_{\substack{1 \leq i \leq |G_1|, \\ 1 \leq j \leq |G_2|}} \delta_{i,j} \tag{2.33}$$

Inter-group proximities may also be interpreted based on group summaries. These commonly include the mean element, the element having arithmetic group mean values for all variables, as well as the minimum or maximum elements. Such representative elements are then substituted in the calculation of proximities between groups, thus making it to a consideration of individual proximities between representative elements. Using the Euclidean distance metric for instance, the inter-group proximity between groups G_1 and G_2, derived from the group representative elements σ_1 and σ_2, is obtained by

$$\delta_{G_1,G_2} = \delta_{\sigma_1,\sigma_2} = \sqrt{\sum_{1 \leq i \leq m} (\sigma_{1_i} - \sigma_{2_i})^2}. \tag{2.34}$$

Also common is the consideration of within-group covariance for interpreting proximity between groups. The Mahalanobis distance D^2 for instance, is obtained by

$$D^2 = (\overline{x_1} - \overline{x_2})^T \; S^{-1}_{G_1 \cup G_2} \; (\overline{x_1} - \overline{x_2}), \tag{2.35}$$

with $\overline{x_1}$ and $\overline{x_2}$ being the group mean elements and $S_{G1 \cup G2}$ the pooled covariance matrix of both groups.

By considering the within-group covariances, the Mahalanobis distance takes the shape of groups into account for quantifying proximity, i.e. similarity. However, for high differences in within-group covariances, this metric is an inappropriate inter-group measure. For this case alternative measures were suggested, such as

$$\delta_{G_1,G_2} = \frac{2 \; b_t^T \; (\overline{x_1} - \overline{x_2})}{\sqrt{b_t^T \; S_1 \; b_t} + \sqrt{b_t^T \; S_2 \; b_t}}, \tag{2.36}$$

with $b_t = (t \; S_1 + (1-t) S_2)^{-1} \; (\overline{x_1} - \overline{x_2})$.

Such inter-group measures are of fundamental interest in clustering approaches, which are introduced in a later section. For more information, the interested reader is referred to [ELL01].

Weighting and standardization

Throughout all presented measurements of proximities between multivariate elements, a weighting, $\omega_i \in \mathbb{R}$ for $1 \leq i \leq m$, may be assigned to the variables or samples, reflecting a measure of specific importance. Weights may also be assigned to enhance or neglect specific data structures.

As a special case of weighting, the standardization of variables to unit variance or range is common in many applications. Samples may be collected with differently scaled variables, showing different value ranges or variances. Standardization of the variables is often useful when scale-dependent proximity measures are used to compare multivariate elements.

Standardization weights, that impose a uniform treatment of variables on the data, are often applied prior to analysis. The reciprocal of the samples' standard deviation or range in each variable y_i are commonly used as weights ω_i for this standardization.

In contrast to treating variables uniformly, weights may also be applied as an enhancement of specific variables or elements. Depending on the application, the investigator might want to reduce the impact of outliers on certain analysis by the (down-)weighting of such elements. When considering groups of elements, weighting might also be applied to elements based on representative properties of their group.

By the application of weights, certain measurements of proximity can reflect favored properties for the application. In this regard, there are vast possibilities for the improvement of data analysis, when the appropriate weights are chosen. However, in most cases, the appropriate weights are found in practice through experimentation and discussion with application experts. Nevertheless, there are some general approaches to weighting. For instance, weights derived from estimates are often used for enhancing analysis results. Such estimates include the probability of a variable's property to account for sought-after data features. Weights are then applied with the aim to emphasize on variables with the highest potential regarding their properties for analysis, e.g. classification. For example, in clustering applications, weights derived from estimates of within-group variability show high potential for recovering groups.

It should also be noted that a given data set for analysis might already contain a pre-weighting of variables, namely the variables that have been measured and selected for analysis. They reflect therefore the interest of the investigator according to a binary weighting. Further weighting by subjective importance judgments, e.g. through user interaction, should be applied with caution, since such results may only reflect an existing classification of the investigator. Therefore, subjective weights might not lead to new results based on genuine data features. Weights should always be chosen to enhance specific aspects of analysis and it should be carefully evaluated if this enhancement does not restrict the overall analysis to a point, where the output is merely a consequence of the weighting.

2.3.2 Clustering

In the context of multidimensional data analysis and visualization, a group of similar objects is called a cluster. In other words, the art of bundling objects in a way that reflects their inherent relationships is called clustering. An excellent introduction to the field of cluster analysis has been presented by Everitt et al. [ELL01]. This section marks the main aspects of their extensive survey and provides a short overview of the field. The reader is

advised to recall the introduction to proximity metrics, presented in section 2.3.1, since they are a key interest to all techniques that require comparisons.

One distinguishes between two basic kinds of clustering methods. These are the hierarchical and optimization clustering techniques. *Hierarchical clustering* imposes a hierarchical structure on the data, defining the relationship of the objects by stepwise refining the data as a single unit into single-elements groups. On the other hand, *optimization clustering* is not concerned with the hierarchical coherence of the data. Its goal is to partition the data elements into a specified number of clusters, for which a numerical criterion is optimized.

As hierarchical clustering is of particular interest to this work, it is considered in more detail. However, a brief revision of basic optimization criteria is presented. Algorithms that efficiently compute approximations of such optimization problems are not part of this work. The interested reader is referred to [PS82] or [Mit98], which give an introduction to the computation of optimization algorithms.

As with proximity measures, no one clustering method can be recommended above the others in all cases. Factors like the graphical display of the output, the needed robustness toward noise in the data, the desired mathematical properties of the clusters or the produced structure, as well as computational expenses need to be considered. Clustering methods should always be designed to recover the types of clusters that are suspected in the data and to fit the application dependent needs [ELL01].

Hierarchical clustering

Hierarchical clustering techniques are concerned with the assembly of an hierarchical structure for all elements of the data. Therefore, the output may be interpreted as a directed tree, having groups of elements (being clusters themselves) as inner nodes and the data elements as leafs. The cluster of all data would be represented by the root of this tree.

One differentiates between two basic techniques of hierarchical clustering, which are defined by the direction of the tree's construction. *Agglomerative* methods build the hierarchy bottom-up. They begin with single-element groups and stepwise join the most similar groups together, forming a new group, represented by an inner node, within the hierarchy. Once joined groups are never touched again, which ensures an unambiguous structure. However, this benefit comes with the price that mistakes cannot be undone.

Analogously, *divisive* methods construct a hierarchy from top to bottom. This way they produce a series of partitions in each hierarchy step, beginning with a single cluster that contains all the data, to a n clusters with a single element.

Depending on the different presumptions of inter-individual and inter-group proximities, hierarchical clustering methods attempt to find the stepwise optimal divisions of the data. The interpretation of inter-group proximities is of particular importance, since different measures of such produce highly different resulting hierarchy structures.

Everitt et al. have revised many *agglomerative methods*, defined by different inter-group proximity measures. In analogy to inter-group proximities discussed in section 2.3.1, common agglomerative methods are presented below [ELL01] and illustrated in figure 2.14.

- *complete linkage*: furthest neighbor proximity
- *single linkage*: nearest neighbor proximity

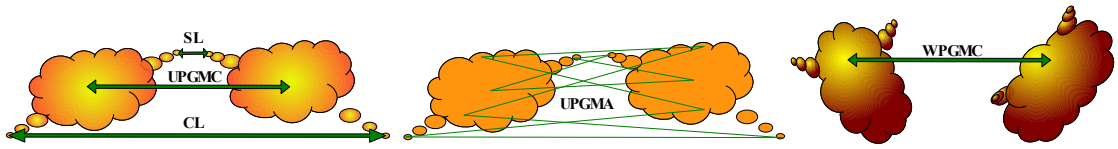


Figure 2.14: An illustration of different inter-group proximity relations in agglomerative clustering methods. The relevant distances when comparing two point clouds are shown for the criteria complete linkage (CL, bottom-left), single linkage (SL, top-left), group average linkage (UPGMA, center), centroid linkage (UPGMC, mid-left), and median linkage (MPGMC, right).

- *group average linkage (UPGMA)*: averaged unweighted pairwise inter-individual group member proximity
- *centroid linkage (UPGMC)*: unweighted inter-individual group mean (centroid) proximity
- *median linkage (MPGMC)*: inter-individual weighted group mean (median) proximity

Median linkage imposes an equal weighting for the computation of a newly joined groups' mean, so that groups with many members do not dominate smaller groups regarding the joined mean.

Lance and Williams' recurrence formula constitutes a method that produces such a nested structure by optimizing a criterion in a stepwise manner. Their proposed flexible scheme for the update of inter-group proximities, due to the fusion of groups i and j, is given by

$$\delta_{k,(ij)} = \alpha_i \delta_{k,i} + \alpha_j \delta_{k,j} + \beta \delta_{i,j} + \gamma |\delta_{k,i} - \delta_{k,j}|, \tag{2.37}$$

where $\delta_{i,j}$ is the proximity between groups i and j. Lance and Williams used this scheme with the parameterization $\alpha_i = \alpha_j$, $\gamma = 0$, $\beta < 1$ and $\alpha_i + \alpha_j + \beta = 1$. Other parameterizations can be found, for which this scheme matches the above presented agglomerative methods. For example, single linkage corresponds to the parameterization $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ and $\gamma = 1/2$.

Ward proposed a different method, in which the fusion of two clusters is based on an overall error sum E. Similar to an optimization step, the clusters are joined that minimize the increase in the sum of within-cluster errors. For a disjoint partition of the n m-dimensional data elements into groups, $\{1, ..., n\} = G_1 \cup ... \cup G_{n_g}$, this total error is defined as

$$\begin{aligned} E &= \sum_{1 \le g \le n_g} E_g, \quad \text{with} \\ E_g &= \sum_{i \in G_g} \sum_{1 \le j \le m} (x_{i,j} - \overline{y_{j(g)}})^2, \\ \overline{y_{j(g)}} &= \frac{1}{|G_g|} \sum_{i \in G_g} x_{i,j} \ . \end{aligned} \tag{2.38}$$

Since $\overline{y_{j(g)}}$ equals the mean of the observations in variable j for group g, the group error criterion E_g essentially equals the within-cluster sum of variance for each variable.

Empirical studies reveal that agglomerative methods may suffer from the effect of chaining, reversals and inversions. Furthermore, they often fail to recover non-spherical clusters that are present in the data. Ward's method works well for many cases, whereas single linkage should not be used with noise between clusters within the data. It is argued that because of these problems, divisive methods are likelier to recover the main structure of the data [ELL01].

Divisive methods operate globally and stepwise minimize *disorder* (heterogeneity) within each group, similar to direct optimization algorithms. The disorder D is reduced, by the split of a cluster (ij) in two clusters i and j, by $D_{(ij)} - D_i - D_j$. Consequentially, $D_i + D_j < D_{(ij)}$. Analogously to agglomerative methods, the split is chosen that achieves the highest of such reductions.

Divisive techniques include the monothetic and polythetic division. Given their naming, monothetic methods base their split on a single variable. They optimize homogeneity or association with other variables and are often used for binary data, where they tend to minimize the number of splits.

Polythetic divisive methods base their decision for a split on all variables. At each step, they form two splitter groups with the pair of most dissimilar elements as seeds and adjoin the rest of the elements to the group having the more similar seed. For example, with the Euclidean distance as the inter-individual proximity measure, this methods cuts a (perfect) ellipse exactly in half, with a split perpendicular to the major axis.

As a remark, interesting mathematical properties of hierarchical clustering methods are introduced in the following. A basic property, that should be considered, is the *ultrametric property*, which states

$$\delta_{i,j} \leq max(\delta_{i,k}, \delta_{j,k}), \text{ for all clusters } i, j \text{ and } k, \tag{2.39}$$

where $\delta_{i,j}$ is the proximity between clusters i and j. For most agglomerative methods, this holds for the pair (i,j), at the step when they are grouped together.

A related property to consider is the property of *spatial distortion*. The with single linkage associated chaining effect, for which dissimilar objects are drawn into the same cluster through noise, is an example for this effect. The property of *spatial conservation* states

$$\begin{aligned} &min(\delta_{u,i}, \delta_{u,j}) \leq \delta_{u,(ij)} \leq max(\delta_{u,i}, \delta_{u,j}), \\ &\qquad \text{for all clusters } i, j \text{ and objects } u, \end{aligned} \tag{2.40}$$

where $\delta_{u,(ij)}$ is the proximity of objects u to the joined cluster of i and j. In other words, it states that the proximity of u to a joined cluster (ij) is neither smaller than the minimal proximity, nor greater than the maximal proximity to one of the original clusters. Through the fusion of two clusters, these bounds are conserved.

Group average linkage would be an example, for which this property holds. Contrary to

conserving this proximity relation, spatial distorting methods interpret proximities to be closer or further after the merging of clusters. For example, complete - and single linkage are two spatial distorting methods.

Everitt et al. [ELL01] have summarized a number of admissible properties that have been discussed in literature, which include

- *Clump admissibility*: regarding the within-cluster proximities in relation to the between-cluster-proximities
- *Convex admissibility*: regarding the intersection of convex hulls of clusters in Euclidean space
- *Point proportional admissibility*: regarding the alteration of cluster boundaries, subsequent to the replication of objects
- *Monotone admissibility*: regarding the change in the clustering, subsequent to monotonic transformations

It should be noted that hierarchical methods are not limited to producing a nested structure, but can also be used to find a number of partitions of the data. The choice, at which step to stop the further fusion or segmentation of clusters, however, represents a general problem.

Often the number of clusters in the data is unknown and cannot be specified in advance. However, an indicator for the "best cut" is given by a large change in fusion level or by a low change in disorder level. In order to find this cut, common methods evaluate the clump admissibility of the structure, which criteria is based on the ratio of between-cluster to within-cluster variance.

Optimization criteria

As optimization clustering is not part of this work, we do not dwell into the vast field of approximation algorithms. Nevertheless, we introduce the basic criteria for optimization clustering. Such methods find a partitioning of the individual data elements into a specified number of groups by optimizing a numerical criterion.

An optimal criterion should describe a partitioning that exhibits groups with coherent and distinct properties or structure. Furthermore, groups should be well distinguishable from other groups by possibly application-dependent properties like spatial positioning, shape, density, distribution, size, spatial orientation or variable trends. Everitt et al. [ELL01] have discussed the most common optimization criteria, which are presented in the following.

Given a dissimilarity matrix Δ of n data elements, basic optimization criteria give attention to the *homogeneity and separation* of the clusters that are to be produced. As a quantification for the lack of homogeneity within a single group, measures include the sum of dissimilarities between each pair of objects in the group. For a group defined by object indices $G \subset \{1, ..., n\}$, this measure is given by

$$h_1(G) = \sum_{u,v \in G, u \neq v} \delta^r_{u,v}, \tag{2.41}$$

where $r \in 1{,}2$. The employment of squared dissimilarities is a common method to emphasize on strict structure.

Instead of considering all pairwise dissimilarities, a measure known as *star index* evaluates the sum of dissimilarities to the group's seed (a center element), which is given by

$$h_2(G) = \min_{c \in G} \left(\sum_{u \in G} \delta_{c,u}^r \right). \tag{2.42}$$

A simple measure of a group's separation is given by the sum of pairwise dissimilarities between the group's members to the members of other groups. For a disjoint partition $\{1, ..., n\} = G_1 \cup ... \cup G_{n_g}$, this represents

$$s_1(G_j) = \sum_{u \in G_j} \sum_{1 \le i \le n_g, i \ne j} \sum_{v \in G_i} \delta_{u,v}^r. \tag{2.43}$$

In some cases, the minimum separation of a group may give a stricter and appropriate measure,

$$s_2(G_j) = \min_{\substack{1 \le i \le n_g, i \ne j, \\ u \in G_j, v \in G_i}} \left(\delta_{u,v}^r \right). \tag{2.44}$$

With one of the above measures of group homogeneity or separation, an optimization criterion might be defined by the sum over all groups, e.g.

$$\sum_{1 \le i \le n_g} h_1(G_i). \tag{2.45}$$

Clearly, the minimization of the above criteria is equivalent to the maximization of the sum for s_1. Ratios of such criteria might also be interesting, for example

$$\frac{\sum\limits_{1 \le i \le n_g} g_2(G_i)}{\sum\limits_{1 \le i \le n_g} h_1(G_i)}, \tag{2.46}$$

represents a balance of the maximization of the minimal separation and the minimization of the group's heterogeneity.

More interesting may be criteria directly derived from a $(n \times m)$ data matrix X. However, the principal considerations are similar. With regard to homogeneity and separation, a dispersion matrix T can be decomposed into a *within and between group dispersion matrix*.

The dispersion matrix T, defined as

$$T = \sum_{1 \le i \le n} (x_{i,\bullet}^T - \overline{X}^T)(x_{i,\bullet}^T - \overline{X}^T)^T, \tag{2.47}$$

is decomposed into a within-group dispersion matrix W,

$$W = \sum_{1 \leq i \leq n_g} \sum_{u \in G_i} (x_{u,\bullet}^T - \overline{X_{G_i}}^T)(x_{u,\bullet}^T - \overline{X_{G_i}}^T)^T, \tag{2.48}$$

and a between-group dispersion matrix B,

$$B = \sum_{1 \leq i \leq n_g} (\overline{X_{G_i}}^T - \overline{X}^T)(\overline{X_{G_i}}^T - \overline{X}^T)^T, \tag{2.49}$$

so that $T = W + B$.

Recall that for a square matrix A, $trace(A)$ equals the sum of A's eigenvalues, whereas $det(A)$ equals the product of its eigenvalues. The following criteria may derive from the above decomposition. To minimize the within-group sums of squares over all variables, one minimizes $trace(W)$, which is equivalent to maximizing $trace(B)$.

Since $trace(W)$ is *scale dependent* and imposes a *spherical structure* on the data, the maximization of the ratio $det(T)/det(W)$ may be of better use if standardization is not wished. Since T, and therefore $det(T)$, remains constant for all group permutations, the minimization of $det(W)$ is equivalent and is a commonly used criteria. It is scale independent and can find non-spherical clusters.

However, $det(W)$ assumes that clusters have approximately the same size and shape. For clusters with *different shape*, the criteria

$$\prod_{1 \leq i \leq n_g} det\,(W_{G_i})^{|G_i|} \tag{2.50}$$

will likely uncover these clusters better than $det(W)$. However, it is restricted to a minimum number of group members, $|G_i| > m$, to avoid singular dispersion matrices.

For uncovering groups of *different size*, the criteria could be adjusted to

$$\prod_{1 \leq i \leq n_g} det\left(\frac{W_{G_i}}{|G_i|^2}\right)^{|G_i|}. \tag{2.51}$$

The interested reader is referred to [ELL01] for a more detailed description.

CHAPTER 3

Explorative Visual Analysis

Due to enhanced data acquisition and analysis methodologies in almost all application domains, dimensionality and amount of data have increased steadily. Since the comprehension of interrelations in human visual recognition is usually limited to three dimensions, gaining insight into high-dimensional data by a meaningful visual representation is a highly relevant and still very much an open research topic. Although different approaches have already been described to depict the coordinate values of multidimensional data points, the static nature of basic visual representations renders them ineffective to facilitate cognition of the data when its amount or complexity exceeds a certain threshold.

Methods for explorative visual analysis add an interactive component to the visual representation in order to explore the data, identify patterns, trends, and other relevant characteristics more effectively. One fundamental approach that has proven to perform well for the exploration of high-dimensional data is the projection from high- to a low-dimensional space. Through the use of such dimension reduction techniques, the often highly complex relationships between multidimensional data elements can be abstracted and encoded into distance relationships between points of a lower-dimensional embedding. Thus, dimension reduction provides visually scalable embeddings of the intrinsic properties of the data. Such embeddings also represent an ideal visual interface for selection and subsequent detailed analysis in a multi-view environment. Designed appropriately, approaches using dimension reduction are ideally suited to gain an overview, assess data relationships, and identify outliers, clusters, or structure. However, problems include false interpretation under ambiguity and insufficient degrees of freedom, as well as the correct separation of salient data features.

This chapter describes our results and approaches using dimension reduction to facilitate explorative visual analysis. Our main contributions are:

- Visualizing values in dimension reduction
- Level-of-detail in dimension reduction

We describe how methods of dimension reduction can be effectively combined with methods from other disciplines to render visual data exploration more powerful and informative, for example, by visualizing higher-level structure, level-of-detail, error, and ambiguity.

The remainder of this chapter is structured as follows. *Section 3.1* [EHH12] provides an overview of dimension reduction techniques suited to aid in explorative data analysis. It is an introduction to the main concepts and related work. A comparison of state-of-the-art methods describes the differences and relationships between dominant research foci.

Section 3.2 [ERHH11, REM*12, EHHR13] describes a novel technique that combines clustering, dimension reduction, and multidimensional data representation to form a multivariate data visualization that incorporates both detail and overview. This unique combination counters the individual drawbacks common to dimension reduction and multidimensional data visualization techniques, namely ambiguity and clutter. Thereby, a specific clustering criterion is used to decompose multidimensional data into a hierarchical tree structure. This decomposition is embedded in a novel dimensional anchor visualization through the use of weighted linear dimension reduction. The resulting Structural Decomposition Tree (SDT) provides not only insight on the data's inherent structure, but also conveys detailed coordinate value information. Fast and intuitive interaction techniques are facilitated in order to guide the user in highlighting, brushing, and filtering of the data.

Section 3.3 [EKHS14] describes a novel method that combines global and local projections into a level-of-detail approach for dimension reduction. Based on the construction of a relative neighborhood graph, we apply hierarchical clustering using the Mahalanobis distance to abstract the data's underlying multidimensional manifold. Thereby, our clustering criterion is designed to capture the shape and topology of the manifold in different compositional levels. Our visualization facilitates interactive exploration of these levels by means of self-embedding projections, each using full degrees of freedom in locally optimal depictions of the corresponding compositional part of the data. Thereby, our method facilitates an interactive exploration of both global and local data properties.

3.1 Survey of related work in dimension reduction

This section provides an introduction to the concepts of visualizing high-dimensional data using dimension reduction. Thereby, we survey methods of dimension reduction that focus on visualizing multivariate data. Our aim is not to be exhaustive but to describe the mathematical concepts and ideas underlying the algorithms, to provide an overview of basic approaches, and to review select state-of-the-art methods. Implementation details, although important, are not discussed. The reader should be aware that there are numerous dimension reduction methods that focus on the various aspects of data analysis. For example, methods for feature reconstruction or classification are closely related to those considered here but are not discussed because their focus is not on visualization. The reader will find that, due to its long history, there are numerous surveys on dimension reduction. For example, authors focus on a specific subset of techniques[SWH*06] or investigations[ST03], provide a broad overview[Bur10], or historical background[LV10].

The remainder of the section is structured as follows. Section 3.1.1 represents the core of the survey - a detailed introduction to the concepts of dimension reduction. After a formal problem statement is given, we divide the basic approaches in two classes: projection (Section 3.1.2) and manifold learning (Section 3.1.3). We also provide a taxonomy for these methods that can act as a classifier for which data the methods are most suited. Section 3.1.4 reviews two recently developed but fundamentally different approaches to

non-linear multivariate data visualization and offers a qualitative comparison between them. The object of this investigation is to infer common trends between different concepts of dimension reduction. Finally, concluding remarks are provided in Section 3.1.5. The work is published in [EHH12].

3.1.1 Dimension reduction

Methods for dimension reduction compute a mapping from high- to low-dimensional space. The formal problem setting can be described as follows. Let $X \in \mathbb{R}^{(n\times m)}$, a set of n points in m-dimensional data space, and two metric distance (or dissimilarity) functions, $\delta_m : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ and $\delta_t : \mathbb{R}^t \times \mathbb{R}^t \to \mathbb{R}$, over data space $\mathbb{R}^m$ and target space $\mathbb{R}^t$ respectively, with $m,t \in \mathbb{N}^*, t \ll m$, be given. A mapping function ϕ that maps the m-dimensional data points ($x_i \in X$) to t-dimensional target points ($y_i \in Y$), i.e.,

$$\begin{aligned} \phi : \quad \mathbb{R}^m &\to \mathbb{R}^t \\ x_i &\mapsto y_i, \text{ for } \ 1 \leq i \leq n, \end{aligned} \tag{3.1}$$

is defined s.t. ϕ "faithfully" approximates pairwise distance relationships of X by those of $Y \in \mathbb{R}^{(n\times t)}$. Thus, the goal is to map the proximity of points in data space to points of equal proximity in target space, i.e., $\delta_m(x_i, x_j) \approx \delta_t(y_i, y_j)$, for $1 \leq i,j \leq n$. In particular, an adequate mapping is designed to ensure that remote data points are mapped to remote target points.

Since the target space usually has lower degrees of freedom than those required to model distance relationships in multi-dimensional space, the mapping ϕ adheres to an inherent error that is to be minimized by its definition. Often, ϕ is defined to minimize the least squares error $\epsilon(\phi(X))$:

$$\epsilon(\phi(X)) = \sum_{1\leq i,j\leq n} W_{i,j} \ (\delta_m(x_i, x_j) - \delta_t(y_i, y_j))^2, \ \textit{for } W \in \mathbb{R}^{(n\times n)}, \tag{3.2}$$

where W is a weight matrix that can be used to define the importance of certain data relationships or dimensions. For example, this may be used to disregard outliers by defining $W_{i,j} = 1/\delta_m(x_i, x_j)$ (for $\delta_m(x_i, x_j) \neq 0$) [KC04].

Formally, the above definitions require both data and target distance functions to be metric. That is, both functions must adhere to the properties of positive definiteness, symmetry, and the triangular inequality. For human perception, the most intuitive distance metric is the Euclidean distance, $L_2(p,p') = \sum_{1\leq i\leq q} \sqrt{(p_i - {p'}_i)^2}$ for $p,p' \in \mathbb{R}^q$. Due to this fact, the Euclidean distance is often chosen as the metric for the target space, $\delta_t = L_2$. However, the distance (or dissimilarity) measure of the application domain, δ_m, is in most cases not Euclidean and may in some cases not even be metric. For example, psychometric dissimilarities can be non-metric. Methods that deal with non-metric distances, for example, by mapping the rank order of data points, are introduced in Section 3.1.3.

In the following, we review and discuss several algorithms that realize a suitable mapping ϕ as defined above. We divide them into two basic approaches of the following underlying

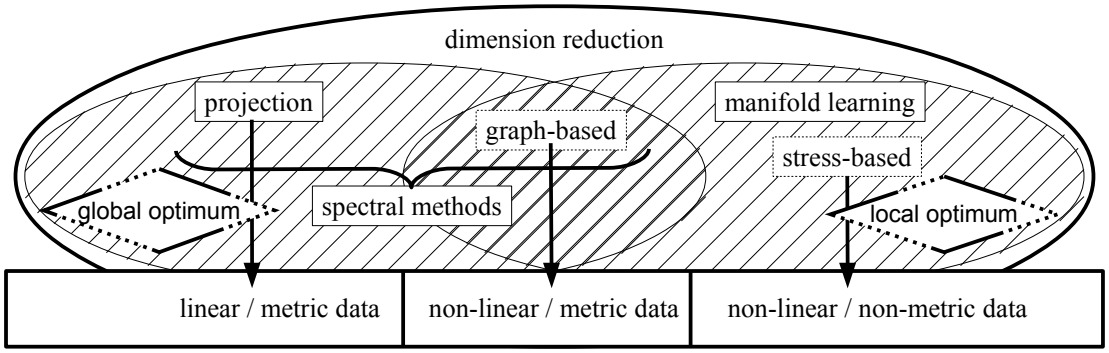


Figure 3.1: A taxonomy of dimension reduction methods. We distinguish between projective methods, graph-based methods, and methods for manifold learning, by being applied for the corresponding type of data.

principal geometric ideas. If the data lies within a linear subspace of lower dimensionality, then it can be re-expressed by a linear basis transformation without loss of information. These bases can be ordered according to their contribution to the mapping error $\epsilon\phi$ and the t bases are used that minimize this error. However, if the data is non-linear and lies on an unknown manifold of lower dimensionality, then distance relationships along this manifold can be learned in an unsupervised manner and used for data mapping.

A careful taxonomy of the methods considered here is formulated in the following and illustrated in Figure 3.1. Methods that are solely based on linear inner product transformations are defined as projection techniques, while those that are able to ascertain distance relationships in a non-linear data structure are defined as manifold learning techniques. These techniques can be further grouped in two basic approaches. Focusing on metric data spaces, the first approach is graph-based. These methods model the data as a graph and utilize optimizations of graph theory to learn manifold distances in data space. The second approach is stress-based and focuses on the embedding directly, i.e., learning the mapping that minimizes the mapping error (stress) in target space. Based on iterative optimizations, they can learn the embedding of non-metric distances.

3.1.2 Projection-based methods

Projective techniques display multi-dimensional data by projecting points onto a lower-dimensional space such that distance relationships between points in the projection space reflect specific relationships between the data points in multi-dimensional space. Since these relationships may be too complex to be completely conveyed in lower-dimensional space, projections (and all mappings considered here) are in general ambiguous. We define a projection by the use of a projection in the geometric sense - projecting the data based on a (linear) inner product transformation. The geometric idea behind this approach is to express the data by a set of "condensed" variables that approximately model the (unknown) underlying factors and reduce redundancies. The two main approaches are to project based on variance or inner product relations and both are, in an Euclidean setting, interchangeable.

Principal Components Analysis (PCA)

As one of the first dimension reduction techniques discussed in the literature, Principal Components Analysis (PCA)[Pea01] conveys distance relationships of the data by orthogonally projecting it on a linear subspace of target dimensionality. In this specific subspace, the orthogonally projected data has maximal variance. Thereby, PCA defines a "faithful" approximation as one that captures the data's variance in an optimal way. It has been shown[KC04] that by the maximization of variance, PCA also minimizes the least squares error (3.2) for Euclidean distances in data and target space, $\delta_m = \delta_t = L_2$, under the constraint of orthogonally projecting the data:

$$\epsilon PCA = \sum_{1 \leq i,j \leq n} (L_2(x_i, x_j) - L_2(y_i, y_j))^2. \tag{3.3}$$

Remarkably, PCA achieves this through a computationally efficient linear transformation. The resulting projection is a genuine view that does not distort the data. The only major drawback of PCA is that, due to its linear nature, it does not capture non-linear data well.

For the following considerations, we assume without loss of generality that $X \in \mathbb{R}^{(n \times m)}$ is centered, i.e., the mean of all given data points has been subtracted from all data points. The PCA projection is defined as

$$\begin{aligned} \text{PCA}: \quad \mathbb{R}^m &\to \mathbb{R}^t \\ x_i &\mapsto x_i \, \widehat{\Gamma}, \ for \ 1 \leq i \leq n, \end{aligned} \tag{3.4}$$

with $\widehat{\Gamma} = (\gamma^{(1)}, ..., \gamma^{(t)}) \in \mathbb{R}^{(m \times t)}$ being the matrix storing columnwise the eigenvectors of the corresponding t largest eigenvalues of the data's covariance matrix $S = n^{-1} X^T X$. The largest eigenvalue of S, λ_1, holds the variance of the data orthogonally projected in the direction of γ_1. $\widehat{\Gamma}$, storing the t mutually orthogonal vectors in which directions the data has the largest variance, define a partial orthonormal basis in data space $\mathbb{R}^m$. The orthogonal projection onto the corresponding rank-t subspace in $\mathbb{R}^m$ is defined by $\widehat{X} = X \, \widehat{\Gamma} \widehat{\Gamma}^T$. Thereby, $\widehat{X} \in \mathbb{R}^{(n \times m)}$ is the best rank-t-approximation of X (under L_2). Using the basis $\widehat{\Gamma}$, data points x_i are projected onto this subspace such that $\widehat{x}_i = \sum_{1 \leq k \leq t} \gamma^{(k)} \text{PCA}(x_i)_k$, for $1 \leq i \leq n$.

Besides its broad applicability to visualization, PCA may be used for many more tasks. For example, a prominent gap in the eigenvalue spectra gives an upper bound for the intrinsic dimensionality of the data. Therefore, it is often used for filtering Gaussian noise or for reducing data size and computation time. PCA is a well-established technique with an extensive history. As such, many variants exist and more information can be found, for example, in [Jol02] or [Man86].

Metric Multidimensional Scaling (MDS)

Metric Multidimensional Scaling (MDS)[Tor58], also known as classical MDS, is a well-established approach that uses projection to map high-dimensional points to a linear subspace of lower dimensionality. The technique is often motivated by its goal to preserve pairwise distances in this mapping. As such, metric MDS defines a faithful approximation

as one that captures pairwise distance relationships in an optimal way; more precisely, inner product relations.

Metric MDS finds an optimal (least squares) linear fit to the given pairwise distances, assuming the distance used is metric. If Euclidean distances are given, $\delta = L_2$, metric MDS is equivalent to PCA up to scaling and rotation. However, metric MDS finds the best linear fit to any metric dissimilarities. This makes the technique more flexible to use compared to PCA. Its performance is also independent of data dimensionality, however, the method scales poorly with the number of data points.

By the method's design, the mapping error preserves inner product relations:

$$\epsilon mMDS = \sum_{1 \leq i,j \leq n} (x_i x_j^T - y_i y_j^T)^2. \tag{3.5}$$

Let a matrix of pairwise metric distances (or dissimilarities), $(\Delta)_{i,j} = \delta_{i,j}$, be given. From these metric distances, the data's Gram matrix of inner products is given by $G = HAH^T$, where $A = -1/2\delta_{i,j}^2$ and H is a centering matrix. The complete eigendecomposition of G requires $O(n^3)$ time which is, in most cases, too expensive for practical problems. However, variants of the method achieve an approximation in $O(nlogn)$ time based on a divide and conquer approach of the eigendecomposition[YLMW06]. In addition, increasingly faster solvers are being developed[KMP10].

Metric MDS is defined as

$$\begin{aligned} \text{mMDS}: \quad \mathbb{R}^m &\to \mathbb{R}^t \\ x_i &\mapsto \widehat{\Gamma}_i \, \widehat{\Lambda}, \; for \; 1 \leq i \leq n \end{aligned} \tag{3.6}$$

with $\widehat{\Gamma} = (\gamma^{(1)}, ..., \gamma^{(t)}) \in \mathbb{R}^{(n \times t)}$ being the matrix storing columnwise the eigenvectors of the corresponding t largest eigenvalues of the Gram matrix of inner products, $G = XX^T$, $G_{i,j} = x_i x_j^T$. $\widehat{\Lambda}$ is the diagonal matrix storing the roots of the t largest eigenvalues of G, $\widehat{\Lambda} = diag(\sqrt{\lambda_1}, ..., \sqrt{\lambda_t})$.

Although metric MDS works in the inner product space, the geometric intuition behind the method is very similar to that of PCA. As such, points are projected into the linear subspace of largest variance. However, this subspace is defined by metric MDS based on the eigenvalue decomposition of an $n \times n$ matrix of inner products. The duality between PCA and MDS becomes clear when considering that G has the same rank and eigenvalues (up to a constant factor) as the covariance matrix $Cov(X) = n^{-1}X^T X$ and that for the Gram matrix holds that $G = n^{-1}Cov(X^T)$. Therefore, the Gram matrix is a covariance matrix in $\mathbb{R}^n$ that reflects the same principal relationships of the data as the covariance matrix in $\mathbb{R}^m$, although, expressed in a basis system that reflects linear combinations of data points (instead of dimensions). For more information on metric MDS, the reader is referred to [HS07] or [CC94].

Although being both powerful and flexible, metric MDS leaves two questions unanswered: (1) What if the data are samples from a non-linear manifold and its proximity relationships are unknown? (2) What if these proximity relationships cannot be described by a metric?

In the following, we discuss the essential concepts that solve these two major issues.

Kernel PCA

Kernel PCA[SSM98] is considered a variant of PCA and metric MDS (due to their duality) that is capable of depicting non-linear data. Kernel PCA is based on two assumptions that make the application of (linear) PCA to non-linear data possible. The first assumption is that in the space of the data's underlying features, the data is linear. The second assumption is that there is a function that approximates the inner product of data points in this feature space. This function is called a kernel and the utilization of a non-linear kernel in a linear setting to capture non-linear data structure is known as the "kernel trick". Formally, this setting is described as follows. Let a kernel k be given that approximates inner product relations of non-linear data in their feature space, such that

$$\begin{aligned} k: \quad \mathbb{R}^m \times \mathbb{R}^m &\to \mathbb{R} \\ (x_i, x_j) &\mapsto \Phi(x_i)\Phi(x_j)^T, \ for\ 1 \leq i \leq n, \end{aligned} \tag{3.7}$$

where Φ is the mapping to feature space. Kernel PCA is defined as

$$\begin{aligned} \mathrm{K-PCA}: \quad \mathbb{R}^m &\to \mathbb{R}^t \\ x_i &\mapsto \widehat{\Gamma}_i \ \widehat{\Lambda}, \ for\ 1 \leq i \leq n \end{aligned} \tag{3.8}$$

with $\widehat{\Gamma} = (\gamma^{(1)}, ..., \gamma^{(t)}) \in \mathbb{R}^{(n \times t)}$ being the matrix storing columnwise the eigenvectors of the corresponding t largest eigenvalues of the Gram matrix of inner products *in feature space*, $G^{(k)}_{i,j} = k(x_i, x_j)$. $\widehat{\Lambda}$ is the diagonal matrix storing the roots of the t largest eigenvalues of $G^{(k)}$, $\widehat{\Lambda} = diag(\sqrt{\lambda_1}, ..., \sqrt{\lambda_t})$.

Thereby, Kernel PCA computes the eigenvectors of the covariance matrix of the data in feature space. Although this space, as well as the data coordinates therein, is unknown, the kernel maps to the data's Gram matrix of inner products in feature space. Based on the assumption of the correctness of a kernel k, the eigendecomposition of $G^{(k)}$ captures the non-linear relationships in the data by maximizing variance in feature space. As such, Kernel PCA can be viewed as a generalization of metric MDS by substituting the utilization of Euclidean dot products to generalized dot products.

It is not surprising that the bottleneck of Kernel PCA is finding the "right" kernel. Since distance relationships along the possibly non-linear sub-structures of the data are, in general, a-priori unknown, the definition of a suitable kernel requires explicit knowledge about the data. If this knowledge is not given, methods are better suited that determine distance relationships along non-linear data structures in an unsupervised data-driven manner. This is the concept of manifold learning.

3.1.3 Manifold learning

Projection-based methods work well for data that fits approximately to a linear subspace. When this is not the case, the hope for dimension reduction is that the data follows at least a non-linear pattern, i.e., it lies on a lower-dimensional manifold. The methods considered in this section are able to learn (and depict) proximity relationships of data points on

(non-linear) manifolds in an unsupervised manner. While mappings from projection-based methods can be described by linear transformations that capture known proximity relationships, this is not the case for manifold learning techniques. In particular, these techniques abstract from Euclidean distance relationships and capture distances along a manifold. Figure 3.2 illustrates the difference between projection and manifold learning based mappings.

There are two distinct approaches to learn unknown proximity relationships. These approaches are based on the data being of metric or non-metric dissimilarity. To model metric distances on a manifold, graph-based techniques are often used that retrieve local distance relationships in a data-driven way and project the data based on these metric distances. However, there are various applications that require the display of non-metric dissimilarity relationships. This problem cannot be solved by graph-based methods but only through a direct minimization of the mapping error in the embedding. This leads to the optimization of a non-convex stress function. Consequently, stress-based methods are prone to local minima and slow convergence.

Graph-based methods can be divided into two classes: global and local modeling. Global approaches first learn proximity relationships on a locally low-dimensional sub-manifold and, second, depict these relationships using, for example, projection-based methods like metric MDS. Local graph-based modeling follows a divide and conquer approach. The idea is to divide the data into small groups and to solve this embedding locally. Local systems are then "pieced together" based on overlapping or fixation points. Although the projection step finds the global optimum for the embedding, the initial retrieval of distance relationships is based on optimization problems such as shortest path problems, least squares fits, or semidefinite programming. In this regard, graph-based methods are also prone to local minima or higher computational cost.

Non-metric MDS

The ability of metric MDS to map data relationships from a dissimilarity matrix is based on the key assumption that dissimilarities are approximate squared metric distances. As for all spectral methods, this allows for the computation of a globally optimal projection. However, this also limits its application and prohibits non-metric scenarios, for example, stemming from psychometric research where metric postulates do not hold. Instead of this eigendecomposition approach, the idea of non-metric Multidimensional Scaling is to directly minimize the mapping error (3.2) with respect to a given non-metric dissimilarity matrix and possibly some weighting thereof. Unfortunately, due to the non-metric nature, the resulting stress function is non-convex and optimization thereof is prone to local minima.

For a perfect projection, it holds that $\epsilon(MDS(X, Y, W)) = 0$, where X is the input, Y the output, and W an optional (arbitrary) weighting. One way to approximate the solution is through a steepest descent approach, for example, with the Euler method[AP98]. Thereby, a step-wise iteration towards zero, where the $(k+1)^{th}$ iteration has the form $Y^{(k+1)} = Y^{(k)} + \alpha^{(k)} \nabla \epsilon MDS(X, Y^{(k)}, W)$, converges to a local minimum. The step size $\alpha^{(k)}$ can be constant or can be computed by means of line search. A disadvantage of this method is its slow convergence near a minimum. An approach to avoid this is to use higher-level, gradient-descent-type methods, for example, Newton's methods[KTT98].

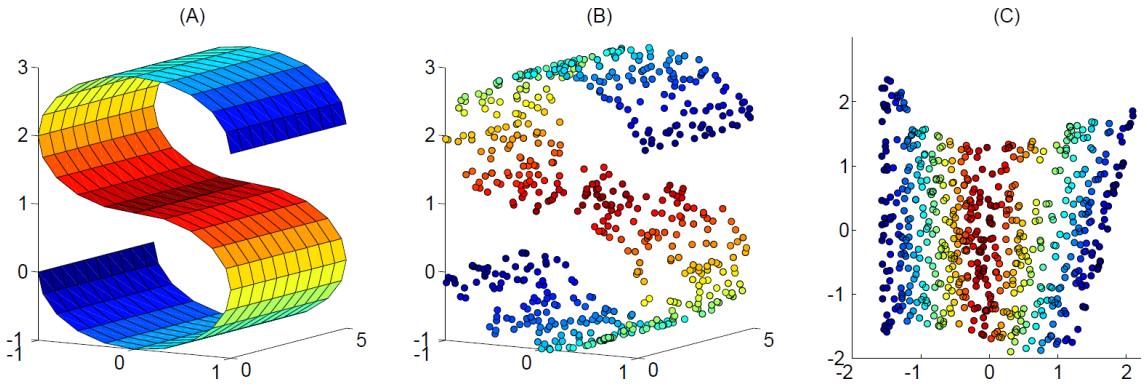


Figure 3.2: The difference between projection and manifold learning. In the swiss roll data set, data is sampled along a non-linear manifold (A). A projection of the data is shown in (B), while an embedding by manifold learning is shown in (C). The embedding by manifold learning does not provide a view of the data points in data space but a distortion thereof along the manifold in order to approximate proximity relationships. Image courtesy of [SWH*06].

These methods converge more quickly at a higher computational cost.

The exact embedding of non-metric dissimilarities in a metric target space is impossible. However, in non-metric MDS, the rank-order of dissimilarities is assumed to contain the most significant information and the main goal of the approach is to depict the rank-order in its output configuration. A well-known approach to non-metric MDS is the Shepard-Kruskal algorithm[Kru64]. At its core is a twofold optimization process that optimizes the goodness of fit with regard to the non-metric input. First, an optimal monotonic transformation of the non-metric dissimilarities to metric distances is found that preserves the rank-order of non-metric inputs. After the optimization of the rank-order distances, the output configuration is further improved iteratively, balancing both stress and monotonicity.

MDS is in all respects a hard non-convex optimization problem. Using a good initialization is therefore important. Numerous variants of MDS exist and many other methods are closely related, like Sammon's mapping[Sam69]. Especially multi-level approaches have substantially increased performance[IMO09]. For an overview, reference [BG05] is helpful.

Isomap

Instead of learning the embedding directly in target space, Isomap[TSL00a] attempts to explicitly model non-linear proximity relationships in terms of geodesic distances. As such, it can be viewed as a variant of metric MDS to model non-linear data using its (metric) geodesic distances. In order to retrieve these distances, a global graph-based optimization approach is utilized.

Geodesic distances are learned by linearly approximating the non-linear manifold. Thereby, a network of undirected neighborhood graphs is constructed in which each data point is a node and has edges to its neighbors that are weighted by the points' dissimilarity. The weights represent the local approximation of geodesic distances on the manifold. From these graphs, a square geodesic distance matrix is computed which is used for the metric MDS projection. The essential steps can be summarized as follows:

1. For each data point x_i compute an undirected k-neighborhood graph based on the k points of smallest dissimilarity to x_i and assign this dissimilarity as the edge's weight.[1]
2. The $(n \times n)$ matrix of geodesic distances $\tilde{\Delta}$ is found by computing the shortest paths through the network of neighborhood graphs.[2]
3. Project the data using $\tilde{\Delta}$ and metric MDS, as described in Section 3.1.2.

One problem of Isomap is that after double-centering of the geodesic distances, the Gram matrix of inner products is not guaranteed to be positive semidefinite. One variant that solves this issue is Maximum Variance Unfolding (MVU)[WS06]. The underlying idea behind MVU is to unfold the manifold under the constraint that local distances between neighboring points are preserved. This is optimized with respect to maximum variance.

Note that the lower-dimensional embedding of geodesic distances by Isomap involves the eigendecomposition of a dense $(n \times n)$ matrix. Like with metric MDS, this leads to significant computational effort. Further variants exist that tackle this problem, for example, by integrating a local approach[ST03].

Locally Linear Embedding (LLE)

In contrast to modeling a manifold by global geodesic distance relationships, LLE[RS00a] models the manifold by extracting its local intrinsic geometry. Thereby, LLE follows a local graph-based approach. The basic idea of LLE is based on the linear approximation of all data points (in complex non-linear structures) by a convex linear combination of its neighborhood. Formally, this assumption can be described by the following equation which has to hold for all data points $x_i \in X$ and their surrounding neighbors N_i,

$$x_i = \sum_{x_j \in N_i} W_{i,j} x_j \tag{3.9}$$

with $0 \leq W_{i,j} \leq 1$, $\sum_{x_j \in N_i} W_{i,j} = 1$, and $W_{i,i} = 0$, for $1 \leq i,j \leq n$. The local intrinsic geometry has the appealing property that it stays unchanged under transformations like translation, rotation or scaling. Hence, the local linear relationships of points in data space directly define the intrinsic geometry for the output points to target space. The weights $W_{i,j}$ are approximated by solving a least squares problem based on a k-neighborhood graph. In contrast to Isomap, LLE models nearest neighbors by directed graphs which leads to a more suitable approximation. With these local relationships, LLE constructs a set of global equations for the projection to target space. The method is summarized as follows:

1. For each data point x_i, compute the k neighbors N_i that are nearest to x_i with respect to the distance function δ_m.
2. Compute the weights $W_{i,j}$ that minimize the equation $\sum_{i=1}^{n} |x_i - \sum_{j=1}^{n} W_{i,j} x_j|^2$ and satisfy the constraints, $W_{i,j} = 0$ if x_j is not a neighbor of x_i, $W_{i,i} = 0$ and $\sum_{j=1}^{n} W_{i,j} = 1$ for all $1 \leq i \leq n$.

1 Often a threshold is used to model disconnected sub-manifolds.
2 This can be computed, for example, using Dijkstra's algorithm[Dij59].

3. Compute the output points y_i that minimize the equation $\sum_{i=1}^{n} |y_i - \sum_{j=1}^{n} W_{i,j} y_j|^2$.

As with Isomap, the data projection step is done by solving an $n \times n$ eigenproblem that is based on the global weight matrix W. Due to the locality of LLE, this weight matrix is sparse which leads to a significant advantage in terms of computation speed. The projection is defined by the bottom $t+1$ [1] eigenvectors of the matrix $(I-W)^T(I-W)$ that can be computed without a full matrix diagonalization[DDRvdV00].

3.1.4 Current state of research

Having introduced the main concepts of dimension reduction that can be utilized for visualization, this section reviews more recent work. We compare the two dominant and distinct approaches to non-linear dimensionality reduction, namely graph- and stress-based methods. We review one representative paper of each approach, each one being both state-of-the-art and comparable in terms of similar goals and assumptions. Because both methods stem from a different background, it is likely that they have been developed independently from each other. Our goal is to infer common trends, relations, and solutions of these independent research streams that both solve the problem of finding optimal lower-dimensional embeddings for non-linear multivariate data.

Piecewise Laplacian-based Projection (PLP)

Similar to LLE, PLP[PEP*11] makes the assumption that every data point x_i can be approximated by a convex combination of its neighbors $x_j \in N_i$ based on weights $W_{i,j}$. While LLE finds those weights through optimization, PLP uses pre-defined weights according to:

$$W_{i,j} = \frac{1}{\delta_m(x_i, x_j)} \Bigg/ \sum_{x_k \in N_i} \frac{1}{\delta_m(x_i, x_k)} \tag{3.10}$$

with δ_m being the metric distance function of the data space. Due to those pre-defined weights, the projection has no unique solution. Therefore, a set of global control points is added on a divide-and-conquer basis to solve this problem. PLP divides the data in smaller subsets, each contributing a number of control points that are globally projected to preserve global relationships among subsets. This procedure allows for corrections based on user input, which makes this method interactive. PLP is defined by the following steps:

1. Separate X into $s = \sqrt{n}$ different samples S_j for $1 \leq j \leq s$. [2]
2. For each sample S_j define the neighborhoods $N_i \subseteq S_j$ for each $x_i \in S_j$ and a set of control points $C_j \subseteq S_j$.
3. Globally project all control points $C = C_1 \cup \ldots \cup C_s$ from $\mathbb{R}^m$ to $\mathbb{R}^t$.

1 The bottom eigenvector is a unit vector and is discarded to enforce the constraint that the embeddings have zero mean. Here, bottom refers to the ordering imposed by largest to lowest corresponding eigenvalues.

2 Note that $\sqrt{n}$ is an upper bound for the number of groups in a data set of size n [PB95]. More sophisticated estimation schemes may also be used.

4. For each sample S_j, construct and solve a separate local linear system but based only on the local variables C_j and the neighborhoods $N_i \subseteq S_j$.
5. Present the resulting projected data points Y to the user who can redefine the neighborhoods. Based on the new neighborhoods, repeat the method from step three.

Paulovich et al.[PEP*11] set the number of neighbors k to ten and the number of control points in each sample S_i to $\sqrt{|S_i|}$ [1], which ensures that the number of control points of a sample corresponds to its sample size. The set of global control points C can be embedded by any appropriate mapping, for example, Paulovich et al. use the stress-based *Force Scheme*[TMN03].

After the local linear systems have been solved for each sample, the user can interact with the projected data set through its representation as a k-nearest neighbor graph and adjust neighborhoods or samples by simply moving data points within the embedding. Due to the used multi-level approach, only the linear systems of samples have to be recomputed in which the neighborhoods have been changed. Consequently, PLP can learn the embedding of large high-dimensional data sets in a semi-supervised manner.

If data does not come in a tagged format, partitioning it into samples is done by clustering methods. On the one hand, the multi-level approach leads to significantly smaller total computational cost since the linear systems, which are solved at step four, are now smaller. On the other hand, important global features may be missed due to this approach. Since the control points (randomly chosen) set the frame for global relation of local patches, there is no guaranty that global features can be preserved in all cases. However, the option of user interaction can compensate for this scenario.

Multigrid Multidimensional Scaling (MG-MDS)

As a variant of multidimensional scaling, MG-MDS[BBKY06] is based on the direct optimization of the *weighted* mapping error as a stress function $\epsilon\phi$ given by (3.2), although, the method requires distances to be metric. In contrast to PLP, weights in MG-MDS can be arbitrary and do not represent a convex combination. The basic idea is to re-state the problem of finding $\phi = \arg\min_\phi \epsilon(\phi(X); \Delta, W)$ with respect to the gradient-descent-type method as a problem of finding ϕ with $\nabla\epsilon(\phi(X); \Delta, W) = 0$ and to embed this problem into a multigrid approach, through which substantial performance improvements can be achieved. A simplified view of MG-MDS is that the re-stated problem is first solved for a *core*, a small subset of all data points. But instead of a one-step projection of the remaining data points, each of the remaining data points is projected separately, in a step-by-step projection. Hence, to project one of the remaining data points, not only the projected core but all the so far projected points are used. Obviously, this increases the computational cost, but approximation errors which occur during a big, one-step projection, can be counteracted.

MG-MDS constructs a hierarchy of grids from the data set X such that for $X = \{x_{i_1}, \ldots, x_{i_n}\}$, the hierarchy is defined by choosing x_{i_n} randomly chosen from X and

1 Note that the total number of control points amounts to $n^{3/4}$

picking x_{i_k} from $k = n-1$ to $k = 1$ so that the following equation holds:

$$x_{i_k} = \arg\max_{x \in X} \min_{l=k+1,\ldots,n} \delta(x, x_{i_l}) \text{ for } 1 \leq k \leq n-1.$$

In other words, x_{i_k} is a data point with maximal distance to all data points with higher hierarchy level. Each grid level k holds the set of all data points of the hierarchy level equal or higher than k, i.e., $X_k = \{x_{i_k}, \ldots, x_{i_n}\}$. To transfer between grid levels, multi-grid approaches offer *restriction* P_k^{k+1} and *interpolation matrices* P_k^{k-1}, such that $N_{k-1} = P_k^{k+1} N_k$ and $N_{k-1} = P_k^{k-1} N_k$. Additionally, a corresponding stress function ϵk, based on X_k, Δ_k, and W_k, determines the error.

Choosing a maximal grid level R, MG-MDS is summarized by the following steps:

1. If $r = R$, solve $\min_{X_R} s_R(X_R, T_R)$ by using Euler's or Newton's methods which are based on the gradient of s_R.
2. Otherwise, go from grid r to $r+1$, using P_r^{r+1} and ∇s_r, changing also W and Δ.
3. Apply recursively the MG-MDS method to X_{r+1} and use P_{r+1}^r to get from grid $r+1$ back to grid r.
4. During each movement from one grid to the next, a relaxation, using an SMACOF-type method[BG05], is needed to smooth the errors which occur during the movement.

Note that the existence of P_r^{r+1} and P_r^{r-1} for all $R \leq r \leq n$ is a weaker form of the convex neighborhood assumption of LLE or PLP. P_r^{r+1} and P_r^{r-1} can be found if the data points in grid level r belong to the convex combination of the points in $r+1$, and $r-1$ respectively.

Comparison

Both approaches of stress optimization and spectral decomposition solve the problem of visualizing non-linear multivariate data. However, they achieve this in completely different ways. A comparison between them is difficult because stress optimization solves the much harder problem of embedding non-metric distance relationships, while spectral methods are restricted to metric ones. Nevertheless, such a comparison has the potential of inferring valuable insights on what generic ideas and solutions help with the problem at hand. For this, MG-MDS was chosen as a representative over numerous other state-of-the-art methods that follow the stress optimization approach because its unique advantages are also restricted to the input being metric dissimilarities. Here, the relations between both methods are qualitatively discussed and their suitability for different scenarios is assessed. This comparison is based on the crucial factors that may delimit their application: online behavior, parameterization, and computational cost.

Assumptions

PLP (like LLE) makes the assumption that each data point can be represented by a convex combination of its nearest neighbors. Thereby, the data is approximated by a set of linear patches. MG-MDS, on the other hand, is based on the minimization of the stress function

through gradient methods and uses only a weak form of this assumption.[1] Hence, for data sets where the convex combination property does not hold, no suitable neighborhood can be found, or when the computation of the neighborhoods is too costly, MG-MDS is better suited to solve the problem.

Algorithmic approach
PLP and MG-MDS are similar in the sense that they do not use the whole data set at once. Instead, they use a small subset for the costly core projection[2] and then project the rest of the data with a faster method which uses the core projection. This is a definite trend and saves a significant amount of time. However, this approach requires the data set to be of sufficient size in order for a good initial core projection to be possible. Hence, for smaller data sets, methods like LLE are preferable.

Parameterization
When little is known of a data set, an extensive list of parameters often represents a burden for the analyst. However, in a visual analytics environment, the ability to tweak the mapping based on knowledge and interaction is a definite advantage. Additionally, expert knowledge is utilized that simplifies the problem of embedding. PLP requires knowledge of the "right" clustering technique, the number of clusters in the data set, the number of control points, as well as knowledge for defining the "right" neighborhood. This requires the user to have a good initial assessment on the data's structure and its global features. Therefore, when no expert is available, MG-MDS is the safer choice because it requires less user parameters (maximal grid level and core gradient method). On the other hand, PLP's ability to iteratively refine the mapping based on user interaction makes the method more suitable for visual exploration and allows one to infer this knowledge over time.

Online behavior
Considering online scenarios where an existing solution is to be adjusted with regard to new data, PLP is better suited for such purpose than MG-MDS. [3] With PLP, new data points do not chance the global projection but only the local linear system within the sample which can be computed with comparably low computational cost. In this regard, MG-MDS has to be redone for grid levels in which the new data points occur. Although, most likely, the maximal grid level $r = R$ stays unchanged, the overall computational cost is higher. Both methods, however, are based on the dimension of the data points. For online scenarios where, instead of new points, new dimensions are added to the already existing data, methods solely based on local intrinsic geometry (like LLE) are advantageous. In any case, local methods are preferable for online scenarios.

Computational cost

1 In MG-MDS, the convex combination is only a sufficient condition but does not have to hold for all data points and also does not include neighborhood relations.
2 Either the projection of the control points or the calculation at the maximal grid level $r = R$.
3 It is assumed that the new data points are not taken as control points.

Another limiting factor for the suitability of dimension reduction methods with regard to many applications are their computational costs. The cost for computing a neighborhood graph depends on the form the data is given in. In case of a distance matrix, the cost to construct a k-neighborhood graph amounts to $O(k \cdot n)$ for each of the n data points. If the data is given as an m-dimensional point set, the computational cost to define the neighborhood for each data point is $O(m \cdot n^2)$. Although, in some cases, space partitioning data structures like *K-D trees*[FBF77] can reduce this cost to $O(n \cdot \log n)$, their suitability for higher-dimensional spaces is an open research question. We therefore denote the cost to compute the distance matrix by $O(\text{Distance})$, while the cost to compute a k-neighborhood graph is denoted by $O(\text{Neighbors})$. With these considerations in mind, the computational costs of these two methods are:

PLP: $O(\text{Distances}) + O(n^{3/2}) + O(n^{9/4}) + O(s{\cdot}\text{Sample}_{n/s})$ with $s = \sqrt{n}$ being the number of samples and $O(\text{Sample}_{n/s})$ the computational cost for each sample with size n/s. For this, a uniform size over all samples is assumed. $O(\text{Sample}_k)$ is defined as $O(\text{Sample}_k) = O(k^{3/2}) + O(k^{3/2})+$ the computational cost to solve a linear system of size $k \times (k + \sqrt{k})$, with $\sqrt{k}$ being the number of control points in the sample. The two other terms are the cost to find the samples using a clustering method and the global projection of all $n^{3/4}$ control points using any $O(n^3)$ projection method.

MG-MDS: $O((n - R)n^2) + O(2Rn^2) + O(\text{Distances})$, with R being the maximal grid level. The first term is for the core projection of grid level $r = R$ using Euler's Method. The second term is for movement between these r many grid levels. By using more complex methods than Euler's method, the computational cost increases while the value of the stress function decreases. Based on the same considerations as those made by PLP, it seems that $R = n - \sqrt{n}$ is a fair initial guess for the maximal grid level.

Note that these terms are all upper bounds. The actual computational cost can be far smaller. For examples in PLP, much effort is saved since the computation of the samples and control points uses the clustering results for the computation of the neighborhoods. Also, data may already come in a gridded or tagged form that these algorithms can use and take advantage of.

3.1.5 Conclusions

The comparison of state-of-the-art methods that follow the graph- or stress-based approach shows that no single method can be preferred over another. On the contrary, the effectiveness of state-of-the-art methods mainly depends on the data and application. However, the comparison also shows that there are similar research directions. At present, especially multi-level approaches show great potential and form one of the dominant research directions in both graph- and stress-based manifold learning.

Motivation for ongoing work includes manifolds of complex non-linear geometry, more flexible and interactive embeddings, better encoding of information, and scalability to larger data sizes.We believe that only through the incorporation of multiple concepts from different research fields can methods for dimension reduction keep pace with future problems. Due to the increasing complexity of high-dimensional data sets, a two-dimensional target space is

not sufficient for the embedding. There is a major gap to close to the concepts of information visualization that may be used to gain additional degrees of freedom in an embedding. Furthermore, these concepts can help with better interpretability and interactivity in adjusting both view and model of the lower-dimensional mapping. Conceptual level-of-detail approaches for data abstraction and analysis can be entwined with algorithmic multi-level approaches. Also, the approximation of proximity relationships, as well as their embedding, is not only prone but bound to produce errors. One possible research direction to solve this dilemma is to develop concepts for visual verification that evaluate and visualize the error and ambiguity of a mapping. These concepts can be combined with semi-supervised learning approaches that incorporate user input into the mapping and visualization, thereby allowing a more effective visual exploration.

The following sections provide new conceptual approaches tackling some of the research directions mentioned above. Thereby, we contribute to the current state-of-the-art by ways to render dimension reduction more

- *interactive* in providing novel ways to parameterize embeddings interactively,
- *flexible* by incorporating user input for hypothesis testing and more focused analysis,
- *intuitive* by using visual metaphors, and
- *informative* by incorporating more information than proximity relationships, e.g., higher-level structure, level-of-detail, alternative solutions, ambiguity, and error.

3.2 Combining relationship and value visualization

From the various methods proposed in literature for multivariate data visualization, 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 multidimensional 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 proximity relationships of the points in high-dimensional data space. Common means to achieve this are projection into a low dimensional presentation space. As described in the previous section, projections are an excellent approach for showing relations in the data. However, they immensely suffer from ambiguity that is imposed by dimension reduction. Thus, they can easily lead to wrong conclusions about the data set.

This section describes a novel visualization technique that 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. Central 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 data 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.

The section is structured as follows. After the presentation of related work (Section 3.2.1), we introduce SDTs, discuss related problems, and describe the main ideas of our approach (Section 3.2.2). This is followed by details of the construction of SDTs (Section 3.2.3), interpretation of the initial projection (Section 3.2.4), as well as their interaction mechanisms (Section 3.2.5). Results are discussed (Section 3.2.6) and underlined by a case study (Section 3.2.7. Lastly, concluding remarks are summarized (Section 3.2.8). The work is published in [ERHH11, REM*12, EHHR13].

3.2.1 Related work

Several surveys containing different categorizations of multidimensional 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 multidimensional 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 and are introduced in Section 2.2. 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 multidimensional 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 referred to as dimension reduction techniques in literature. 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]). As introduced in Section 3.1, these techniques display multidimensional data by mapping points to a lower dimensional space, so that proximity relationships between points in the projection space reflect the relationships between the data points in multidimensional space. Since these relationships may be too complex to be completely conveyed in lower dimensional space, projections are in general ambiguous. One of the first dimension reduction techniques to be proposed and still dominant in practice is Principal Components Analysis (PCA). The method conveys Euclidean distance relations in multidimensional space 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 that does not distort the data. In contrast to manifold learning, the embedding by PCA represents a genuine view of the data space that is clearly defined by a projection plane. It is this property that renders PCA as the method of choice to be combined with visual entities and interactive mechanisms, in order to facilitate the incorporation of coordinates in this projection. As introduced by Koren et al. [KC04], linear dimension reduction may also be weighted. This method is also incorporated into our approach to optimize the coordinate display.

Successful representations of complex data often utilize metaphors of commonly understandable concepts, such as topological landscapes [WBP07]. Coordinate values are especially hard to interpret in projections since projections 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 multidimensional 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. For a general overview, see Sectionsec:basics.visualApproaches. 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 multidimensional points by linear combinations, as shown in Figure 3.3. 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 that attain for clusters in the data. However, coordinates cannot be conveyed by these methods and the representation remains ambiguous.

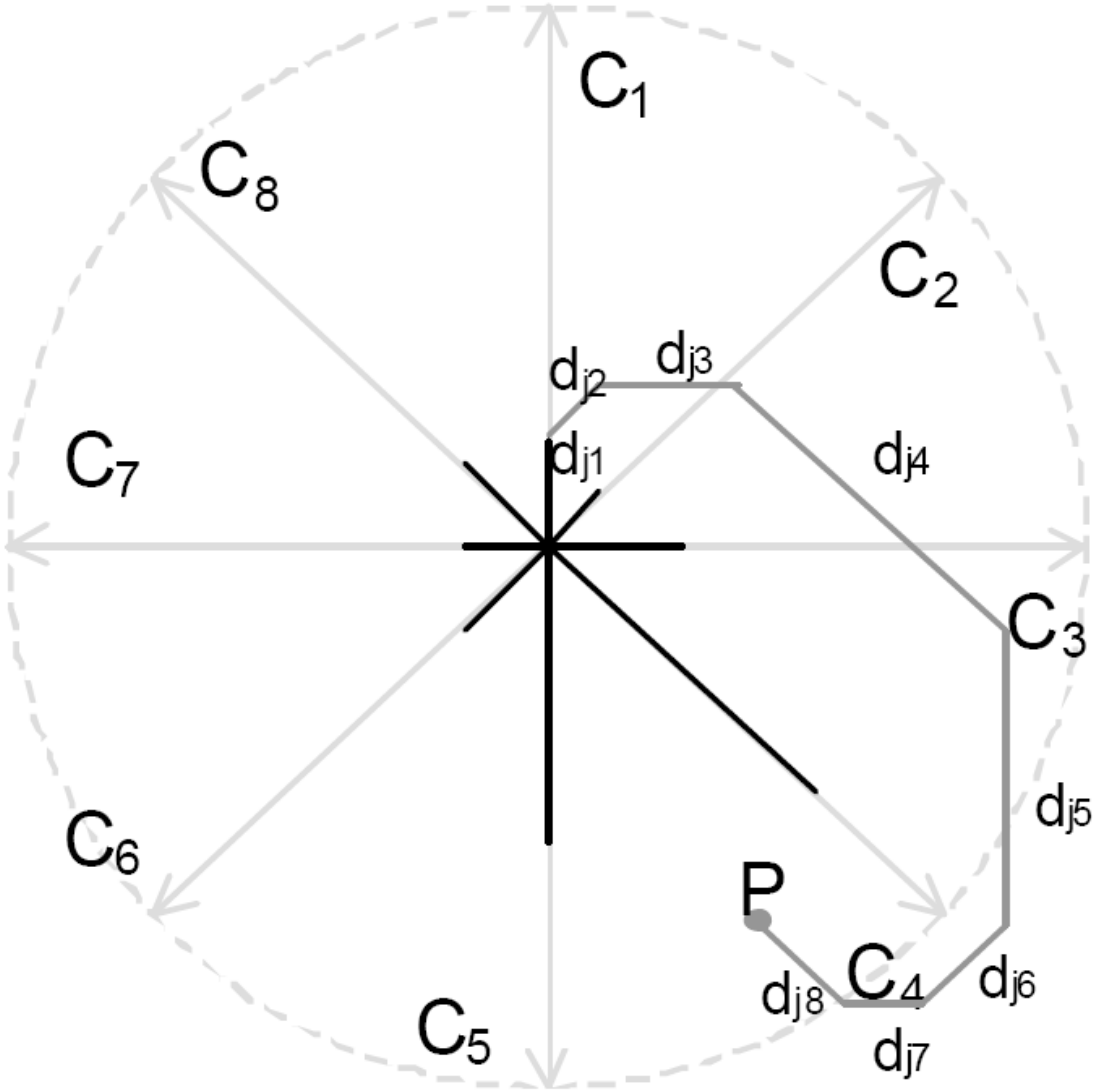


Figure 3.3: A point $P = (d_{j,1},...,d_{j,8}) \in \mathbb{R}^8$ is projected by the star coordinate system by the linear combination of its dimension anchors C_1,...,C_8 with the point's coordinates as coefficients [Kan01]. However, many points can be projected to the same location, making this representation 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.

3.2.2 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 3.3. 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 3.4). 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 configuration 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

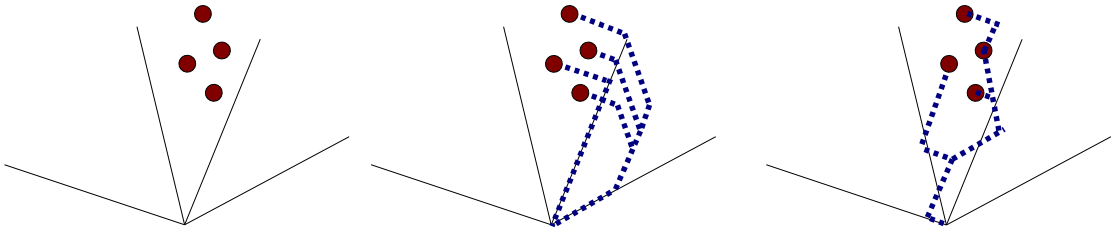


Figure 3.4: By the representation through dimensional anchors alone (left), a simple but ambiguous view is achieved. Ambiguity may be solved by the display of the point's dimension contributions (center). However, this presentation highly clutters the view due to many and redundant line segments. A tree embedding, based on the structural composition of the data, achieves a trade-off in form of an unambiguous and less cluttered view (right).

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.

3.2.3 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.

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 $C_1, ..., C_{n_c}$ containing indices of data points, i.e., $C_i \subset \{1,...,n\}, 1 \leq i \leq n_c$. We define the inter-group proximity measure δ to quantify the measure of compositional commonalities between two clusters i and j as

$$\begin{aligned} \delta_{i,j} &= |\mathbf{min}(C_i \cup C_j)|, \text{ for } \mathbf{min}(C_i \cup C_j) \in \mathbb{R}^m \\ &= \sum_{1 \leq q \leq m} \mathbf{min}(C_i \cup C_j)_q, \end{aligned} \tag{3.11}$$

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 $\mathbf{min}$ of a collection of indices C as

$$\mathbf{min}(C) = (\min_{k \in C}(x_{k,1}), ..., \min_{k \in C}(x_{k,m})), \text{ for } C \subset \{1,...,n\} \tag{3.12}$$

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 led 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

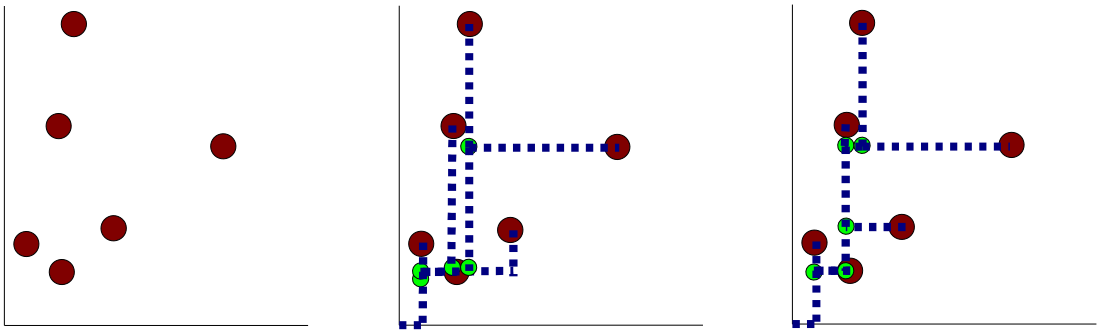


Figure 3.5: 2D data points without (left) hierarchically clustered by a shortest distance criterion (center). Spatial distortion with this metric leads to the tendency to low decomposition points and thus, to high redundancies. Clustering based on the criterion of the highest minimum commonality (right) achieves an embedding that minimizes redundancies, and shows no such spatial distortion effects.

in Figure 3.5 and shows highly favorable properties.

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 $|\mathbf{min}(C_i)| - \delta_{i,j}$ 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

$$\theta_{i,j} + \theta_{j,i} \leq \theta_{k,l} + \theta_{l,k}, for 1 \leq k,l \leq n_c, \tag{3.13}$$

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.

Initial projection

As discussed in Section 3.2.1, 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 $X^T \mathcal{L}\ X$, where $X \in \mathbb{R}_+^{n \times m}$ is centered beforehand. The pairwise weights $\mathcal{L}_{i,j}$ 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 $\mathcal{L}$ 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

$$\sum_{i<j} (dist^t_{C_i,C_j}\ dist^p_{i,j})^2, \tag{3.14}$$

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 (3.14)). 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, $a_i \in \mathbb{R}^2$, is given by

$$a_i = (\begin{pmatrix}0\\1\end{pmatrix} \gamma_{1_i} + \begin{pmatrix}1\\0\end{pmatrix} \gamma_{2_i})^T, \tag{3.15}$$

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.

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 3.2.2, 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 $\vec{v_c} = pos^p(\mathfrak{min}(C_i))^T - pos^p(\mathfrak{min}(C_k))^T$.

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 $minE(C_i) - minE(C_k)$ is drawn. As discussed with (3.13), 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 $n(m-1)$ line segments for every data set, a SDT may have any number of line segments in $[0,(2n-1)m]$, depending on the commonalities in the data. This may be of benefit for large data sets.

3.2.4 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, thus 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 $1 \leq j \leq m$:

$$\widetilde{X}_i = \sum_{1 \leq j \leq m} a_j X_{i,j}. \tag{3.16}$$

This technique, the mapping in star coordinates [Kan01], 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 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 the following, 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 (3.16) 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 $\widehat{\Gamma}$:

$$a_i = \left(\gamma_i^{(1)}, ..., \gamma_i^{(p)}\right)^T. \tag{3.17}$$

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 $1 \leq k \leq r = rank(X)$.

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

$$\widehat{S}_{i,j} = \sum_{1 \leq k \leq p} \lambda_k \gamma_i^{(k)} \gamma_j^{(k)}. \tag{3.18}$$

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: $\widehat{X}_{\bullet,i} = \sum_{1 \leq j \leq m} X_{\bullet,j} (\widehat{\Gamma}\widehat{\Gamma}^T)_{i,j}$. $\widehat{X}_{\bullet,i}$ 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 $(\widehat{\Gamma}\widehat{\Gamma}^T)_{i,j} = \sum_{1\leq k\leq p} S^{(k)}_{i,j}/\lambda_k = \widetilde{S_{i,j}}$, 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.

$$\begin{aligned} ||a_i||_2 &\overset{(3.17)}{=} \sqrt{\sum_{1\leq k\leq p} (\gamma_i^{(k)})^2} \\ &\overset{(3.18)}{=} \sqrt{\widetilde{S}_{i,i}} \ = \ \tilde{s}_i \end{aligned}$$

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.

$$\begin{aligned} \cos\angle(a_i, a_j) &= \frac{a_i^T a_j}{||a_i||_2||a_j||_2} \\ &\overset{(3.17)}{=} \frac{\sum_{1\leq k\leq p} \gamma_j^{(k)}\gamma_i^{(k)}}{\tilde{s}_i\ \tilde{s}_j} \\ &\overset{(3.18)}{=} \frac{\widetilde{S}^{(k)}_{i,j}}{\tilde{s}_i\ \tilde{s}_j} \ = \ \tilde{r}_{i,j} \end{aligned}$$

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 dominant 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. Thus, 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 section.

3.2.5 Interaction with structural decomposition trees

Although the SDT's starting projection has desirable mathematical properties, interactive analysis is usually needed to gain further insight. 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.

SDTs are most effective when used as an explorative interface within a multi-view environment. Through *selection*, point sets may be further analyzed using a high-detail value visualization. To find interesting candidates for detailed analysis, the data must first be interactively explored. Figure 3.6 gives an overview of basic interaction techniques, while more are described in the following.

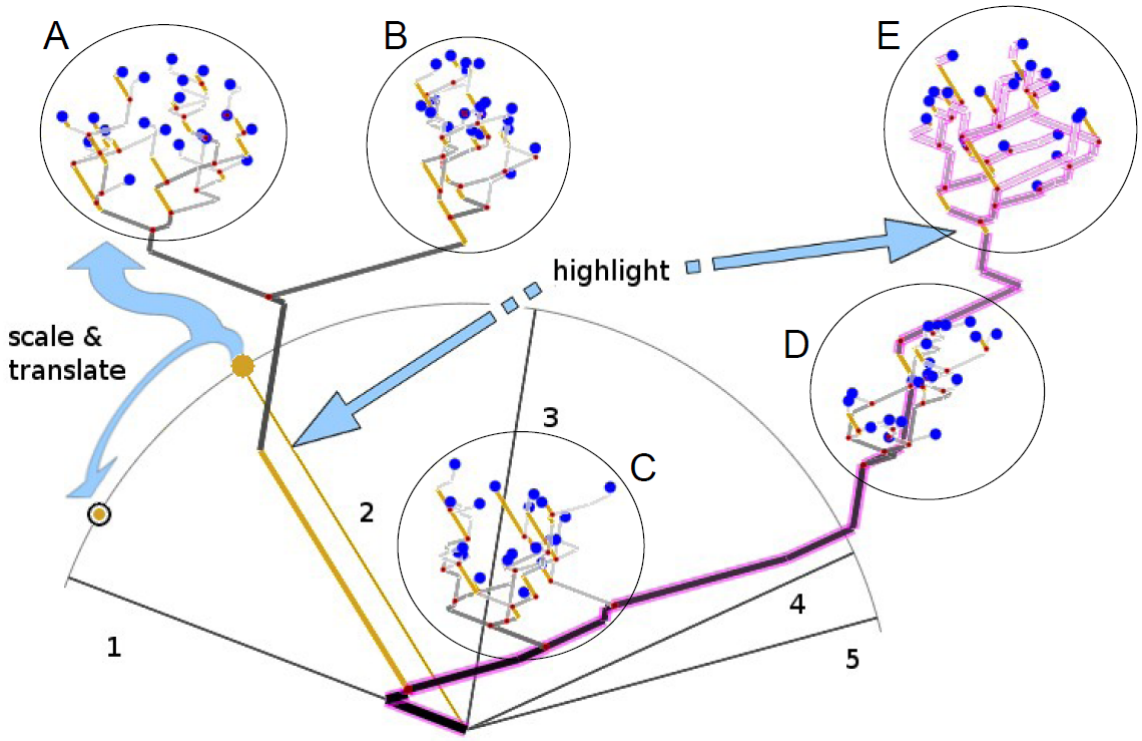


Figure 3.6: A 5-D data set with 5 point-clouds is shown. SDTs best display differences and commonalities within the structural assembly of the data. Further analysis can be conducted by adjusting the projection, highlighting, or filtering.

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 3.6) allows the user to emphasize all line segments corresponding to the coordinates of a dimension and thus helps to investigate the 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 3.6) 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.

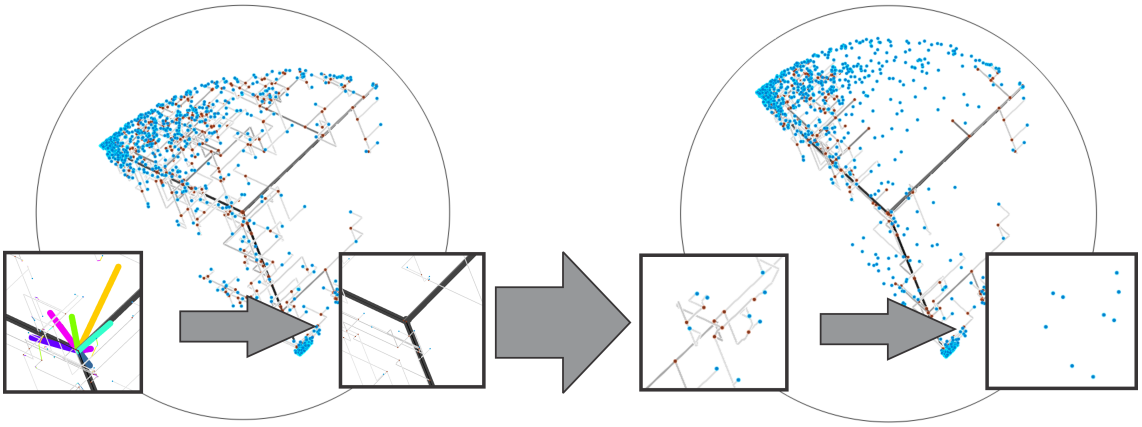


Figure 3.7: Interactive complexity and clutter reduction taking advantage of the capabilities of SDTs: dimension filtering (left) reduces the number of branch segments in the tree, node collapsing (right) the number of displayed subtrees. Additional means for zoom and pan interaction allow the users to drill-down into the presentation and data to obtain momentary detail (bottom).

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 3.7, 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.

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**
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 3.8).

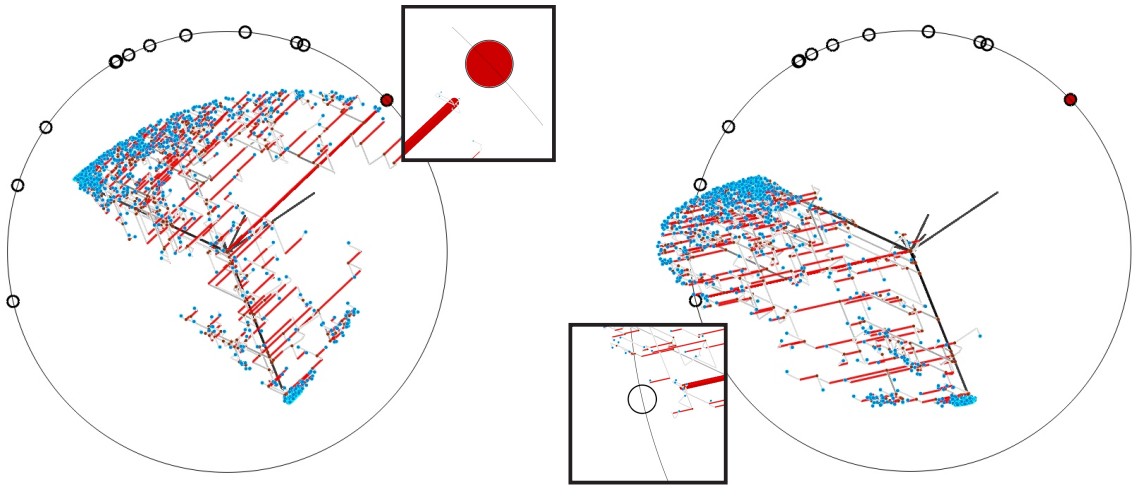


Figure 3.8: Variance points help to find other promising projections of the data. Large variance points (left) indicate projections most suited to convey the variance in the data. Opposite variance points (right), even when not accounting for much variance, often lead to strongly different projections helping to identify unexpected data properties.

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 for the associated dimensions (see Figure 3.9, 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

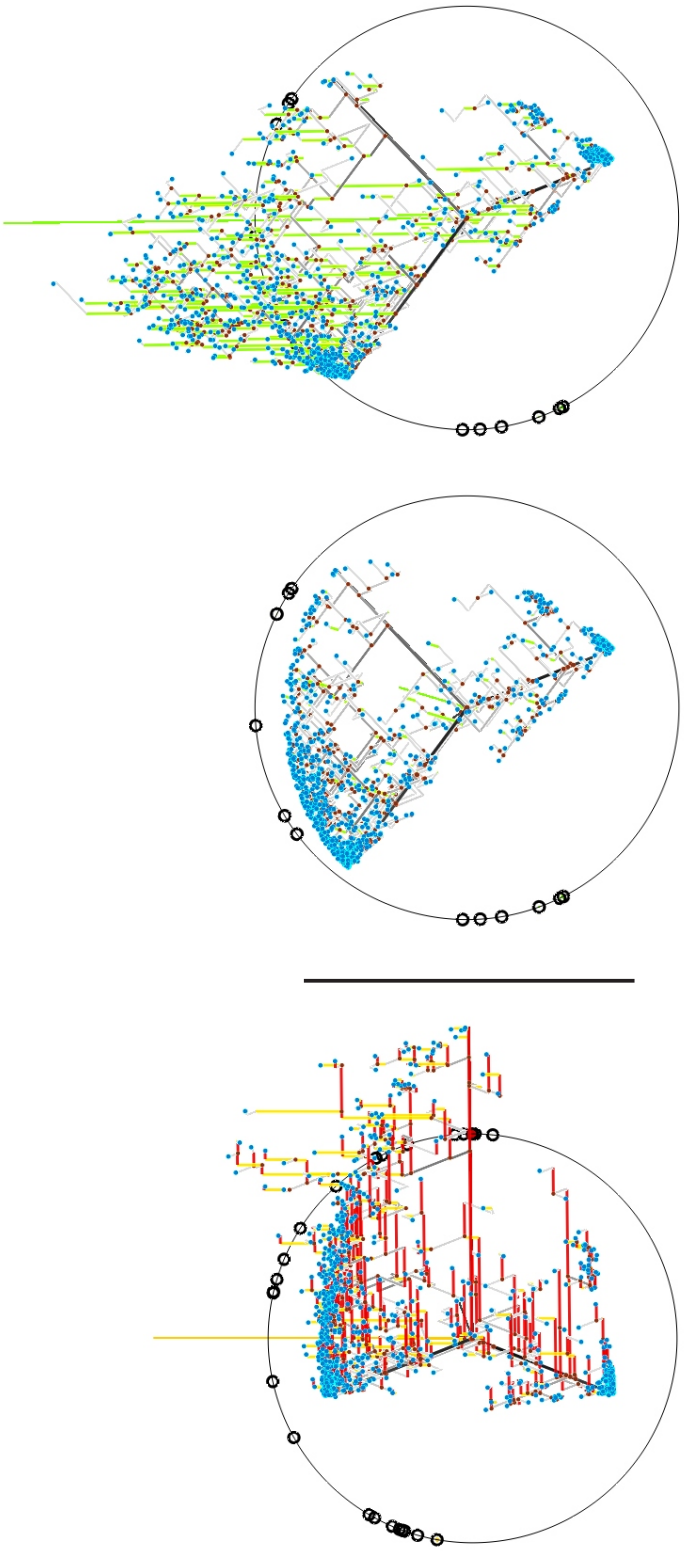


Figure 3.9: The orthogonal placement of two DAs (red and yellow color) in the presentation can simplify the evaluation of correlations between the associated dimensions (left). In order to overcome potential point cluttering within the initial projection (center), a DA (green color) can be interactively stretched (right).

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 3.9, 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 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 3.7, left). Appropriate candidates are dimensions that are correlated or show similar characteristics. They can be substituted by a single *super*-DA, whereby its angle is determined by the average and the length by the sum of all associated DAs. This changes the point projections only slightly, 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 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 3.10). 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.

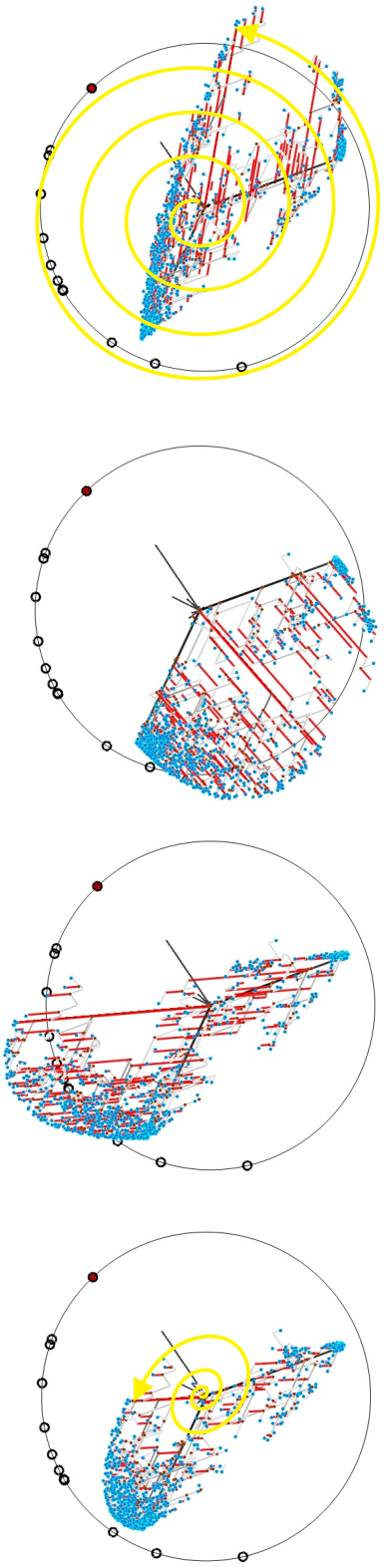


Figure 3.10: Moving a DA in circles activates the motion parallax effect of the human visual system letting the tree and the data points appear more "plastic". By providing many different coordinated projections, characteristics of the data can be identified or verified. Best results are obtained by using a DA corresponding to a dimension with high variance.

Interactions to change the view Analysis on a more granular level, such as **details** to identified clusters, can be obtained by the following interaction. This allows for the assessment of potential sub-clusters, dimension contributions, and distributions.

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 [Shn96] 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 3.7).

3.2.6 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 3.11a-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 multidimensional 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 3.11d, 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 3.11e, 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 most suited to depict a general impression of a data set and to convey an intuitive visual mapping of multidimensional data that can be remembered. The user can learn easily

- how the data is assembled, spread, where clusters are, or which pattern they follow,

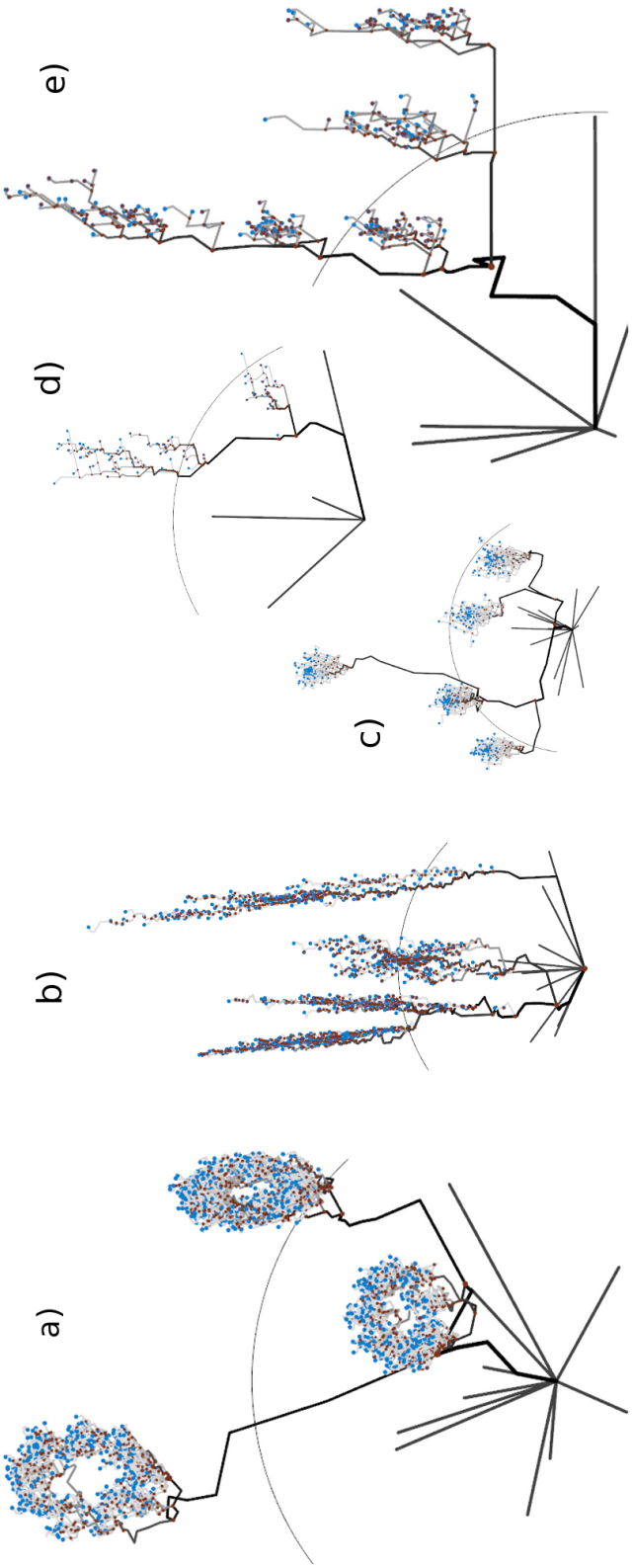


Figure 3.11: Artificial data sets: **(a)** 3 tori in $\mathbb{R}^{10}$, **(b)** 4 ellipsoids in $\mathbb{R}^{10}$, **(c)** 5 point clouds in $\mathbb{R}^{15}$; Benchmark data sets: **(d)** Iris and **(e)** Cars data set.

- 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 multidimensional 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.

3.2.7 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 3.12 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 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 3.12 b)). The opposite applies to Po. NOx is more variant and can be found in similar contributions in both sites (see Figure 3.12 c)).

While exploring the data by *projection adjustments*, it is possible to show that dimension $C24$ representing a carbon isotope cannot be found in Pittsburgh and also has only small concentrations in Fresno (see Figure 3.12 d)). Contributions of $C36$, another carbon isotope, can be found in similar concentrations at both sampling sites with either high or low values (see Figure 3.12 e)). Moving the corresponding dimension anchor to one of its *variance points* also indicates an inverse correlation of $C36$ to C (see Figure 3.12 f)).

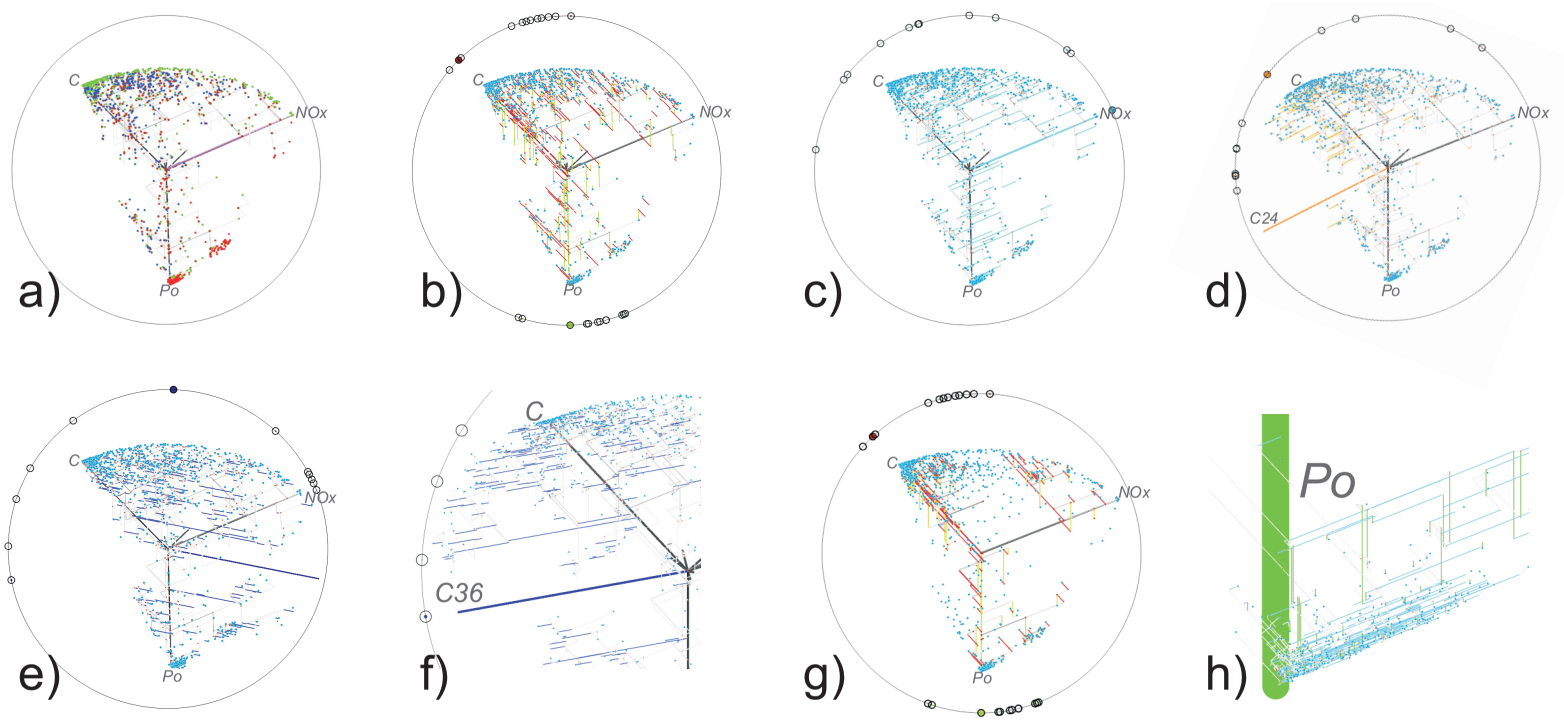


Figure 3.12: Structural decomposition tree of 1000 data points obtained from three different air particle sampling campaigns: **(a)** Initial projection. The coloring of nodes is used for illustration purposes only (red: Pittsburgh, 2002; blue: Fresno, 2007; green: Fresno, 2009). Dimension highlighting 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.

Figure 3.12 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 3.12 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 3.12 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.

3.2.8 Conclusions

The section describes a novel 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 exploration, we have developed unique interaction techniques that enhance exploratory capabilities such as projection adjusting, feature highlighting, and data filtering.

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 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.

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.

3.3 Utilizing graph abstraction for level-of-detail

The previous method is specifically designed to facilitate explorative analysis of global trends. In contrast, the method described in this section is designed to explore properties of the data in different levels of detail, both on a global and local scale. Local properties as, for example, the shape of clusters within a data set, cannot be depict optimally by dimension reduction when the available degrees of freedom are fixed on a global context. Therefore, methods focusing on the depiction of local properties of the manifold, for example, Locally Linear Embedding (LLE) [RS00a], typically apply a neighborhood graph and map distance

relationships from the local neighborhood into a global alignment of points. Thereby, the data residing on some unknown high-dimensional manifold is represented by a series of local patches that represent small locally linear approximations, projected into target space, and aligned by neighborhood distances.

While this approach works well for data that is evenly connected to resemble a single manifold, there are a number of shortcomings. The fix neighborhood size, for example in the dominant k-nearest-neighbor graph, requires not only knowledge of the sampling density beforehand but also cannot handle data with differing distribution well. This often leads to unwanted short circuiting and crossings of the graph and thus to a misrepresentation of the manifold. This disadvantage becomes especially profound for data that contains multiple disconnected clusters. For this data, the visualization often heavily distorts the shape and topology of the data set. In this setting, global projection approaches, such as Principal Components Analysis (PCA) or Isomap [TSL00a], exceed local approaches in depicting global data relationships.

We contribute to the state-of-the-art by a novel method that combines global and local projections into a level-of-detail approach for dimension reduction. Central to our approach are measures and techniques that have not yet been applied to the context of dimension reduction, although they may prove to counter a number of shortcomings. Multi-level approaches have been applied for computational speed gains but have not been made flexible enough to provide interactive level-of-detail analysis. Although dimension reduction techniques that apply neighborhood graphs generally outperform in depicting local data properties, the rigid neighborhood definition yields a number of pitfalls. Clustering methods incorporated in dimension reduction often do not preserve the geometric shape and topology of clusters.

Central to our approach and new in the context of dimension reduction is the interconnection of the following concepts:

- Relative neighborhood definition
- Hierarchical clustering of the manifold based on the Mahalanobis distance
- Visual embedding of level-of-detail projections

Thereby, the data is divided in a hierarchy by applying a clustering method designed to preserve shape and topology of the manifold. Different levels of detail are visualized by iteratively embedding projections of different compositional parts of the data set, thereby facilitating an interactive exploration of both global and local data properties.

The remainder of the section is structured as follows. Section 3.3.1 describes our method in detail, while Section 3.3.1 describes the results of our method being applied to artificial and real data sets. Finally, concluding remarks are provided in Section 3.3.3.

3.3.1 Method

We describe a method that facilitates explorative analysis of multivariate data in different levels of detail by embedding a hierarchy of linear projections. Our method consists of 6 conceptual blocks:

1. Construction of a relative neighborhood graph

2. Hierarchical clustering of neighborhoods based on Mahalanobis distance
3. Embedding of the level-of-detail structure by sets of self-embedded ellipsoids
4. Local principal components projection in each ellipsoid
5. Local projection quality for error and ambiguity
6. Interaction and parameterization of tree traversing

In the following, we describe each of these blocks in detail.

Relative neighborhood graph

The described problems with the application of k-nearest neighborhood graph have led us to the utilization of a relative neighborhood for dimension reduction. Thereby, we follow the definition of Lankford who defined two points within a set as *"relatively close"* if and only if they are "at least as close to each other than to any other points" [Lan69], i.e., the following holds for two points p and q in a set of points X:

$$\delta(p,q) \leq \max(\delta(p,r),\, \delta(q,r) \mid r \in X, r \neq p,q), \tag{3.19}$$

where ϕ is an arbitrary distance metric. Based on this definition of relative neighborhood, Toussaint defined the *Relative Neighborhood Graph* (RNG) [Tou80] that contains exactly the edges between points that are relative neighbors. It is shown that the RNG is a subset of the Delaunay triangulation and that the minimum spanning tree is a subset of the RNG. Toussaint argued that the RNG is more adaptive to the data by imposing less structure on it, while the RNG is also a better approximation of the human perception of the shape of a point set. An example of an RNG in 2D is shown in Figure 3.13

We argue that the RNG is superior to the purpose of manifold embedding than the k-nearest neighbor, as the adaptive neighborhood criterion takes into account local point densities and requires no parameterization. This eliminates the problem of specifying neighborhood size and overall better approximates the manifold in a data-driven manner.

Clustering

The Relative Neighborhood Graph serves as an approximation of the manifold the data resides in. The graph is processed further by hierarchical clustering to obtain a level-of-detail characterization of this manifold. For this, we apply a bottom-up clustering approach based on the *Mahalanobis distance*. This distance measure is defined by incorporating the distribution of neighboring points and is thus a superior approximation of distances within clusters than, for example, the Euclidean distance. The Mahalanobis distance between two groups of data points A and B is defined as follows.

$$\delta^M(A,B)_{A\cap B} = \sqrt{(\overline{A} - \overline{B}) Cov^{-1}_{A\cap B} (\overline{A} - \overline{B})^T}, \tag{3.20}$$

where $\overline{A}$ and $\overline{B}$ refer to the groups' mean, while $Cov^{-1}_{A\cap B}$ denotes the inverse of the pooled covariance matrix of A and B.

The inter-group Mahalanobis distance takes into account the (possibly non-spherical) shape of clusters, which is a highly desirable feature in our setting. However, it is not scale-

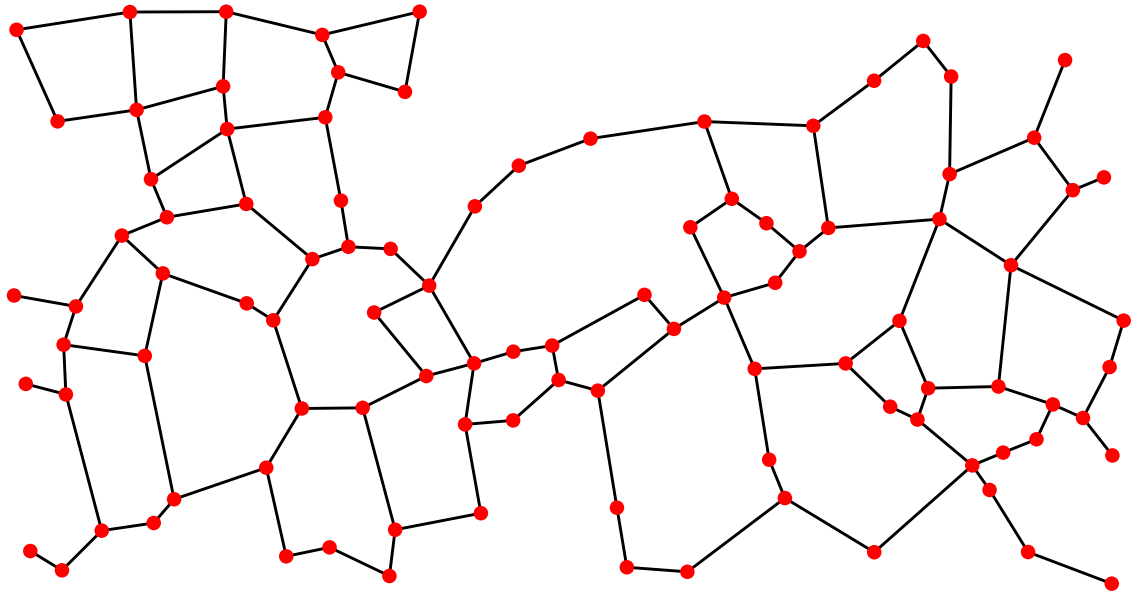


Figure 3.13: Relative neighborhood graph for an arbitrary point arrangement in 2D: The relative neighborhood criterion captures the local density of clusters without the need for parameterization. Thereby, the underlying manifold is approximated linearly. Successive hierarchical clustering of this manifold provides a level-of-detail characterization for the data.

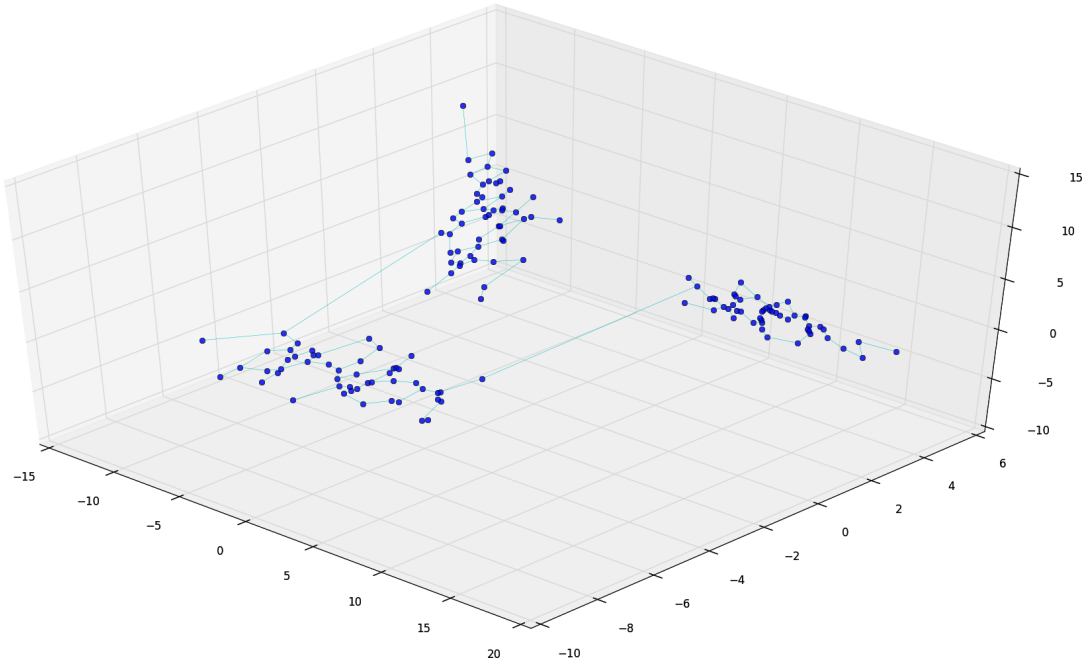
independent but increases with increasing distance of group means and, more importantly, with decreasing within-group variation. The latter represents a less desirable feature for hierarchical clustering. In our clustering method, we avoid these problems by decoupling shape from distance.

For the bottom-up binary hierarchical clustering approach of this work, we apply a new inter-group distance measure that is based on the difference in Mahalanobis distance subject to joining two groups. Further, the user can define the impact of the shape of groups to be considered in this measure by convex combination between normalized Mahalanobis distance and the shortest Euclidean distance between groups. It is defined as follows for groups $A,B \in X$.

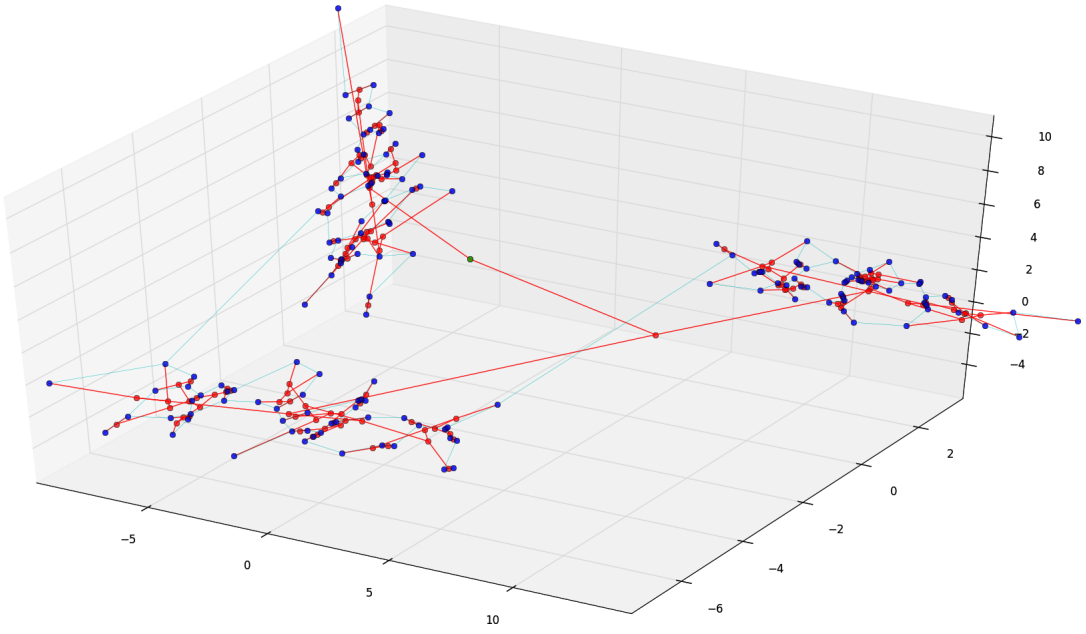
$$\begin{aligned} \phi^{M'}(A,B) &= \frac{\alpha\ d_{min}(A,B)}{\sum_{C,D\in X} d_{min}(C,D)} + \frac{(1-\alpha)\ d_{shape}(A,B)}{\sum_{C,D\in X} d_{shape}(C,D)} \\ d_{min}(A,B) &= \min(||a-b||_2 | a \in A, b \in B) \\ d_{shape}(A,B) &= |\phi^M_A(A) - \phi^M_{A\cap B}(A)| + |\phi^M_B(B) - \phi^M_{A\cap B}(B)| \end{aligned} \tag{3.21}$$

Here, d_{min} measures the minimum Euclidean distance between two groups and d_{shape} measures the difference in Mahalanobis distances for these groups by comparing before and after joining the groups, subject to the original and pooled normed covariance, respectively.

The application of a normalized Mahalanobis measure renders our inter-group shape measure scale-independent, thus allows for decreasing distances under decreasing within-group variation. The decoupled shortest distance measure works entwined with the concept



(a)



(b)

Figure 3.14: An example of the hierarchy generated from our clustering approach: (a) shows the relative neighborhood graph for three clusters in 3D. The resulting hierarchy computed by the clustering of points by our combined criterion of distance and shape is shown in (b). Our clustering captures the structure of the data by gradually accumulating compositionally fitting pieces of increasing size. Image courtesy of [Kar14].

of graph abstraction, as this measure is applied to neighboring clusters in the hierarchically abstracted Relative Neighborhood Graph. An example of the resulting hierarchy is given by Figure 3.14.

Tree embedding

The computed hierarchy of clusters on the approximated manifold is visualized in an embedding that focusses on representing the shape of patches. Patches of the hierarchically abstracted manifold are projected linearly by application of PCA, thereby approximating the manifold in the corresponding level-of-detail. The two principal components of highest eigenvalue represent (by their variance) an elliptical shape. We follow this natural interpretation to embed the projection of patches into ellipses, while sub-patches, representing higher levels of detail, are embedded recursively, as shown in Figure 3.15.

Each ellipse acts as a natural boundary to the projection by PCA and encodes information about the ratio of variance between first and second principal components by its principal axes. The projection is scaled uniformly to fit into the ellipsoid without distorting data relationships. The child ellipses are oriented by rotation such that the angle between first principal component of father and child reflect that of the main axes of the ellipses. Further, they are positioned such that they do not overlap. All applied transformations preserve local distance relationships and serve as an intuitive embedding of the locally linear abstracted manifold.

Projection quality

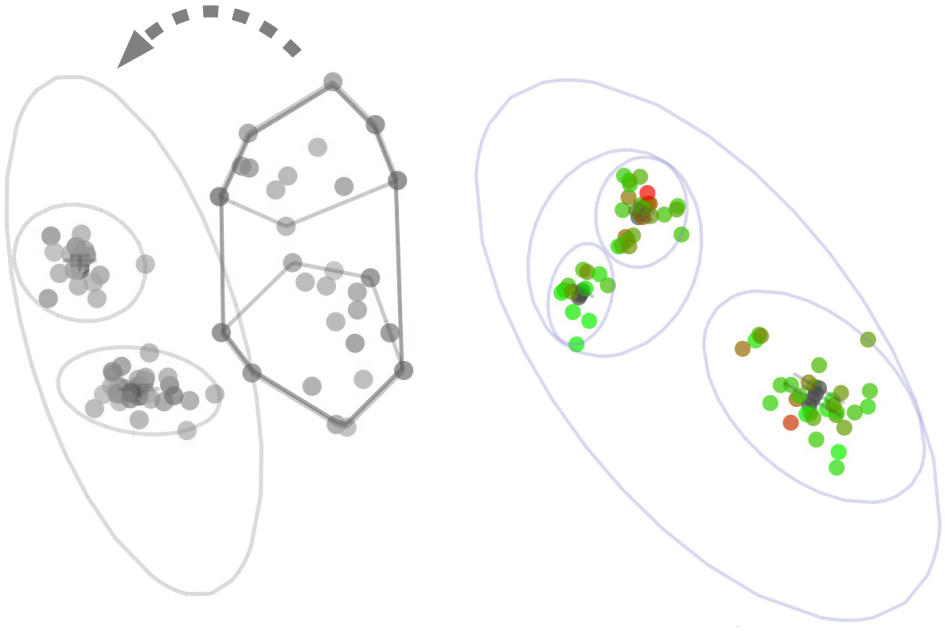
The initial projection provides a single ellipsis showing a global projection of the complete data set by PCA. The user may interact with this view and navigate through the hierarchy, going from global perspective to local representations. Thereby, piece-wise linear approximations of smaller scale of the manifold are embedded recursively in the ellipse that the user wants to subdivide.

Piece-wise linear approximation of a multidimensional manifold yields errors. While PCA minimizes this error, it does not diminish it. Therefore, it is important to provide the user with visual feedback of which part of the current projection is an inadequate representation, thus requiring further descending down the hierarchy into higher levels of detail, or if the current projection is adequate to approximate the corresponding part of the manifold, thus alarming the user that no further descending is required. For this, we provide a projection quality measure corresponding to each projected data point that reflects its distance to the linearly approximated patch of the manifold, the projection plane. Thus, we show which points are further apart in high-dimensional space than suggested by the projection.

The projection quality of a point Y_i in the two-dimensional projection is defined as follows.

$$\begin{aligned} q(Y_i) &= ||X_i - X_i'||_2 \\ Y &= X\gamma \\ X' &= X\gamma\gamma^T \end{aligned} \tag{3.22}$$

Here, γ is the matrix storing row-wise the two eigenvectors of highest eigenvalue of the



(a) (b)

Figure 3.15: The abstraction hierarchy is visualized by successively self-embedding ellipses. (a) Descending down the hierarchy and into higher levels of detail: upon user interaction, the data points that belong to the next patches are highlighted each by a bounding box. A quick animation transforms the bounding boxes into new embedded ellipses that each depict a new point projection using full degree of freedom. (b) The local projection quality, i.e., the approximation error for each point in the projection, is shown by color coding. Green marks a good fit of the projection plane, while points shown in red are ill-represented.

covariance matrix of the data X. $\gamma\gamma^T$ represents the transformation into the projection plane, embedding the points X' in high-dimensional space, while having an intrinsic dimension of two.

To visualize the quality of a projection point, color coding is applied to each point. Percentiles of the data's standard deviation may be defined by the user to mark the adequacy of approximations, thus allowing the user to model acceptable deviation from the approximation, for example by noise, in relation to the data's standard variation. Figure 3.15(b) shows an example of the projection quality.

Interaction

Our method facilitates the interactive segmentation of manifold approximations by descending the hierarchy and recursively embedding projections of the sub-patches, thereby investigating smaller patches of the manifold at higher detail. Tree traversal is facilitated by intuitive parameterization options that define the criteria for the next embedding. The following options are provided:

- Minimum projection quality
- Maximum variance

- Minimum data points

By defining one or more of these criteria, the user inherently specifies the level of detail the user is interested in. The user may define the relative difference of node's projection quality, variance, and data points in relation to that of to their root.

By defining a minimum projection quality, the algorithm descends to the topmost nodes of the subtree corresponding to patches that can be approximated linearly to a specified quality relative to the root's standard deviation, such that the sum of distances to the projection plane is less or equal to the specified ratio of standard deviation. Applied incrementally to subtrees, this leads to an incremental improvement of local projection quality. By defining a maximum variance, the algorithm descends at least so far down the tree such that the variance of the node is less than a specified percentile of variance of the node. Likewise for defining a minimum ratio for the number of data points per patch.

Starting at the root node of an arbitrary subtree within the hierarchy, the nodes closest to the root of the subtree are found that fulfill the requirement defined by the user. These nodes may not be direct children of the root but are directly embedded into the root's ellipsoid. Note that it is not our goal to show the complete cluster hierarchy, as this would unnecessarily overload the visual representation. The visual embedding reflects only the segmentation steps according to user's definition.

3.3.2 Results

In this section, we evaluate the capabilities of the described method to facilitate level-of-detail analysis of multivariate data. We assume that data resides on one or more multidimensional manifolds that can be approximated reasonably on a local scale. Further, we assume that data can be interpreted on multiple scales, i.e., in levels of detail, for example by containing multiple clusters of various distributions, which are to be investigated by locally dissecting the data into different levels. For this, we investigate benchmark data sets that meet these assumptions and compare the results of our method to those achieved by Locally Linear Embedding (LLE) [RS00a] and Isomap [TSL00a], as they represent well-known methods for local and global manifold approximation, respectively. In order to make a fair comparison, we show the two best pictures obtained by LLE and Isomap, respectively, obtained from a series of neighborhood parameterizations for k= 2%, 5%, 10%, 20%, and 40% of the total number of data points.

Orthogonal planes

Beginning our evaluation, we investigate data generated by stretched Gaussian distribution such that points reside on three non-intersecting mutually orthogonal planes in three-dimensional space. This artificial data set serves as a proof of concept for the method described. It contains three disconnected clusters of an intrinsic dimension of two that can each be perfectly represented by a local linear projection but cannot be represented well by a single global projection. In the linear subspace of each cluster, the data exhibits an elliptic shape which is the kind of distribution this method is designed to focus on.

Figure 3.16 shows that our method performs as expected on this data set. Beginning the visual exploration of the data, the user is first facilitated the global overview of the data. Enfolding of the hierarchy and going into higher detail, the three classes within the

data are identified directly. Each is outlined by a bounding box in the projection. This is followed by a quick animation that re-arranges the initial projection into three embedded ellipses. Thereby, each bounding box transforms into one ellipse and all points traverse into either one of the ellipses that each correspond to one of the generated planes. The color coding of the points (green throughout) shows the quality of the local approximations to be flawless. From this view, the user can directly infer that the data contains three classes of ellipsoidal shape, each residing in a plain. In the context of the global alignment, two of the planes can be identified to be orthogonal, although the third cannot due to the lack of degrees of freedom in the global alignment.

Figure 3.17 shows the results of LLE and Isomap for this data. While isomap manages to separate the three clusters to some extend for the initial parameterization of the neighborhood size $k = 20\%$, our method significantly outperforms for this data set.

Intersecting planes

Next, we investigate our method's capability to separate and visualize clusters in the data that are intermingled. Data points residing on two planes are generated as above but the alignment is changed such that the planes intersect each other. Each cluster is again intrinsically two-dimensional, although the intersection makes it generally hard to show

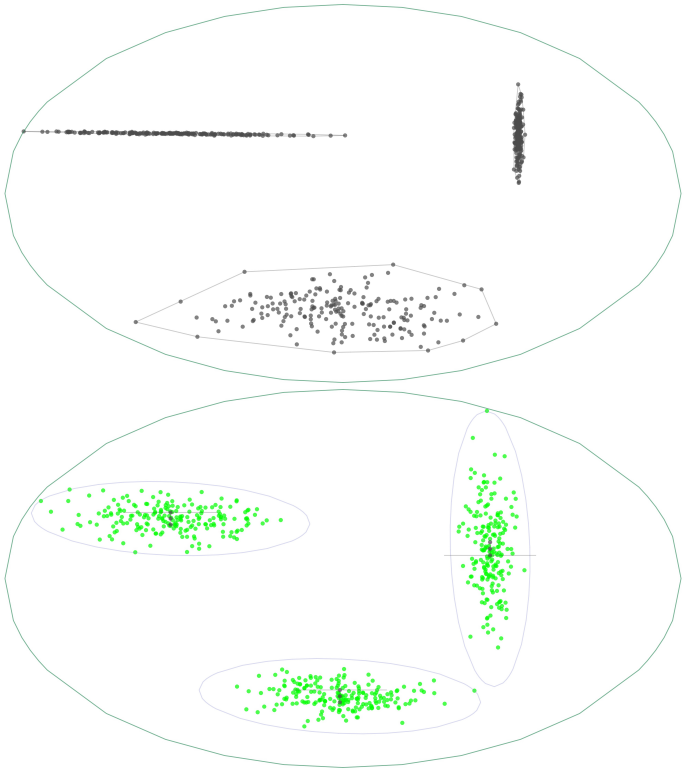


Figure 3.16: Orthogonal planes: The initial view (top) shows the global projection of the data ill-suited. Traversing into higher detail, the three embedded ellipses each facilitate an optimal sub-global view of the data. Image courtesy of [Kar14].

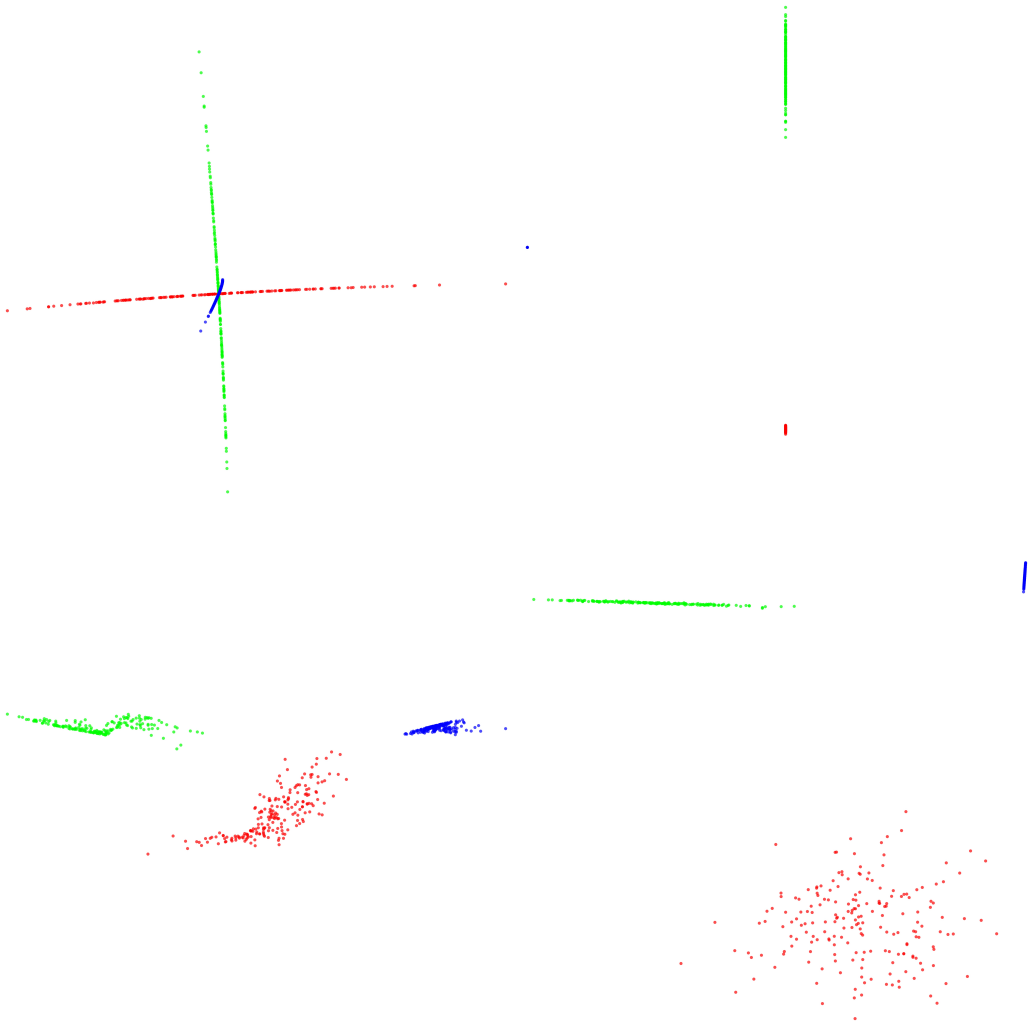


Figure 3.17: Orthogonal planes: Results obtained from Isomap (left) and LLE (right) for data points elliptically distributed on orthogonal planes. The best result is obtained with Isomap for the parameterization of neighborhood size $k = 20\%$ (bottom left). However, the planes are in general ill-represented. Image courtesy of [Kar14].

this feature in data.

Figure 3.18 shows that our method performs unexpectedly well for this data set. The initial view shows already two clusters, one of which being well represented by the global projection and one being misrepresented. Going into higher detail, the two clusters are visualized in their corresponding local projection, showing the data points to be generally represented well in two dimensions. Only a few points near the intersecting region of the planes are shown to be misrepresented. These are the exact points that have been misclassified during clustering to belong to the other cluster, respectively. The visualization alerts the user of the misfit to prevent misinterpretation. Still, the overall visualization of the compositional parts of this data set is excellent, separating the two planes almost

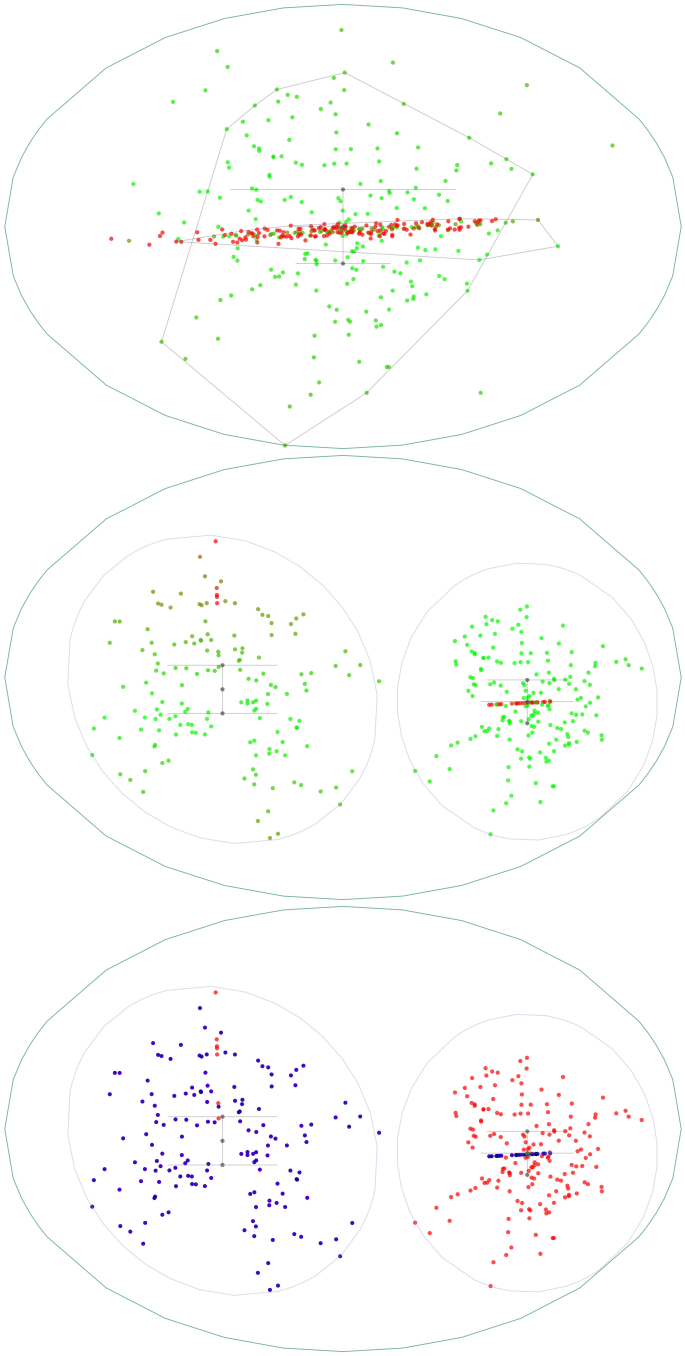


Figure 3.18: Intersecting planes: After the initial global layout (top), traversing into higher detail (middle) reveals the three planes almost perfectly (color coding by local projection quality), except for the immediate intersection region. This is shown by the bottom figure, where color coding depicts the plane classes. Image courtesy of [Kar14].

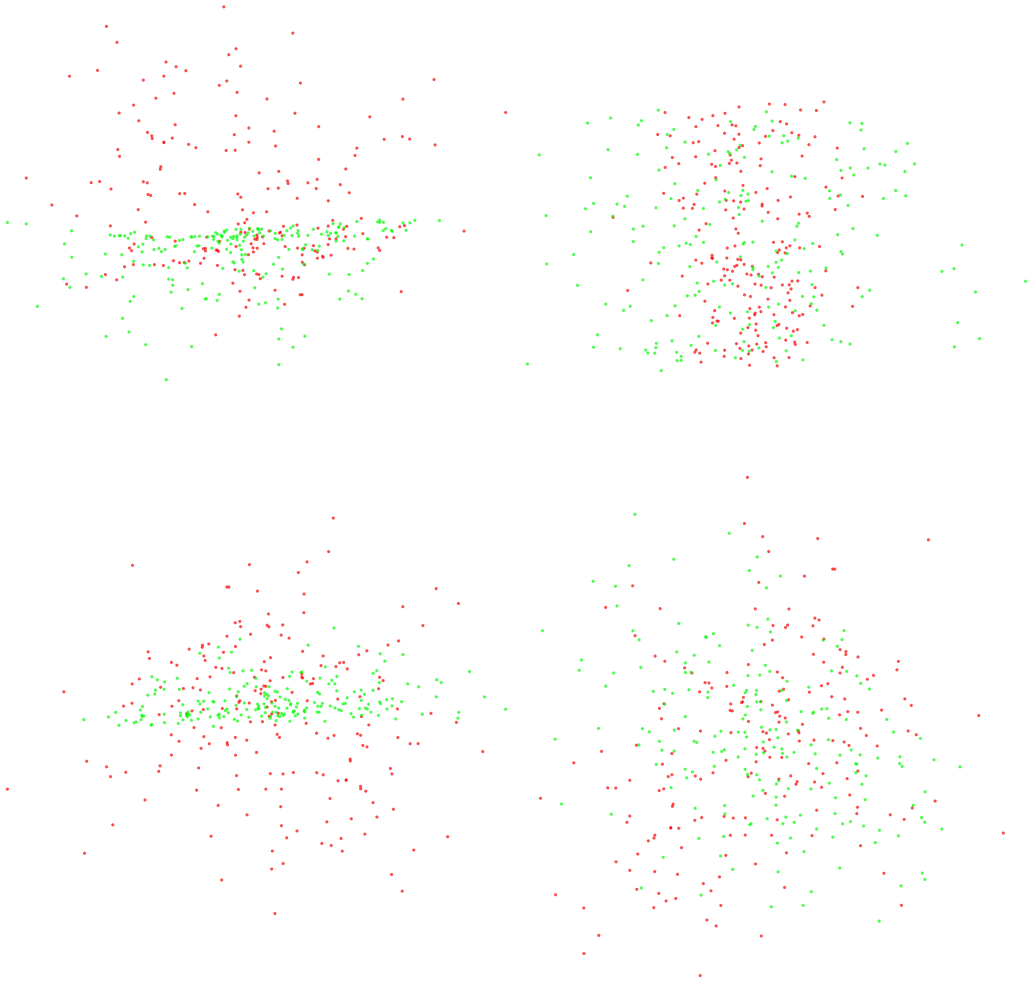


Figure 3.19: Intersecting planes: Isomap (left) and LLE (right) cannot separate the planes. Image courtesy of [Kar14].

exactly.

Figure 3.19 shows the results of Isomap and LLE for different parameterizations. Here, the representations of the data set are hardly better than that of PCA.

Entangled rings

The last artificial data set we consider is data generated along two mutually planar circles. The rings are aligned such that they represent two chain links, closely running through each other but not touching one another. Again, each cluster is intrinsically two-dimensional and aligned in three-dimensional space.

Figure 3.20 shows that our method separates and visualizes both rings perfectly. The initial view shows both classes initially but does not let on the local distance relationships of the left circle. Based on the global projection alone, the user cannot determine the

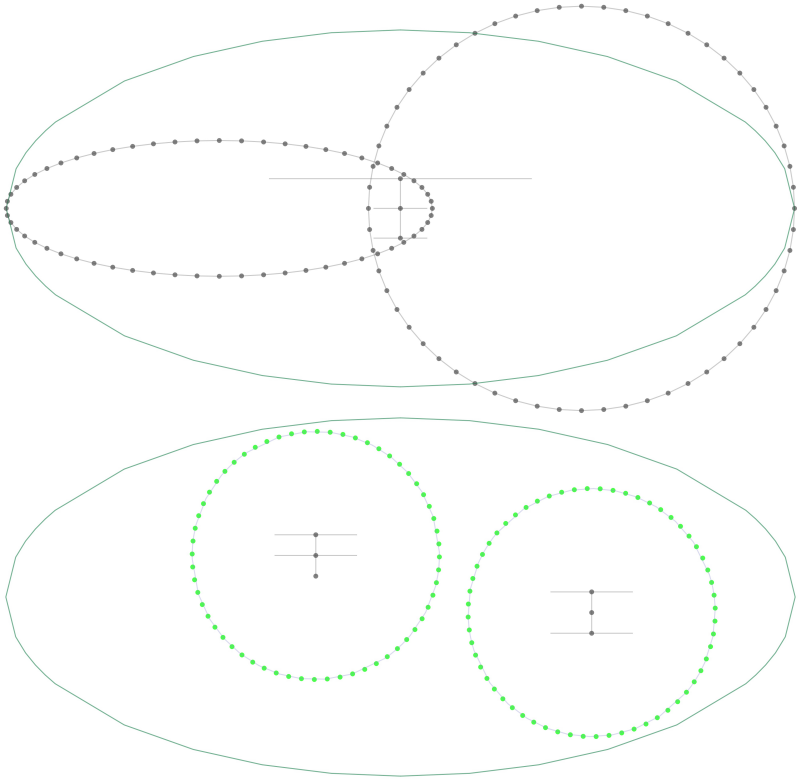


Figure 3.20: Entangled rings: Traversing beyond the initial projection (top), shows both rings unfolded perfectly (bottom). Image courtesy of [Kar14].

shape of the cluster being a circle or an ellipse. However, enfolding yields the two clusters visualized as a circle and without error.

Figure 3.21 shows the results of Isomap and LLE on this data set. The exact geometric alignment cannot be inferred from any result for the tested of parameterizations. However, when presented with results for multiple parameterizations, one can identify a general trend for the data. In one subfigure, the rings are connected, while in another they are disconnected, and in a third they are closely intermingled. However, the actual shape of the circles is generally ill represented.

Iris flower data

Having exceeded at artificial data sets described above, we finally test our method for the representation of real data. As a benchmark data set, we choose the Iris flower data that is often used in classification and pattern recognition domains. It contains three classes corresponding to different species of the iris flower, each described by four variables corresponding to measurements along four axes for each flower. While one cluster can be generally well separated from the other, two clusters are highly intermingled and are often ill represented.

Figure 3.22 shows the two results of our method for different weightings of distance and shape in our clustering approach. With 30% set to distance and 70% set to shape,

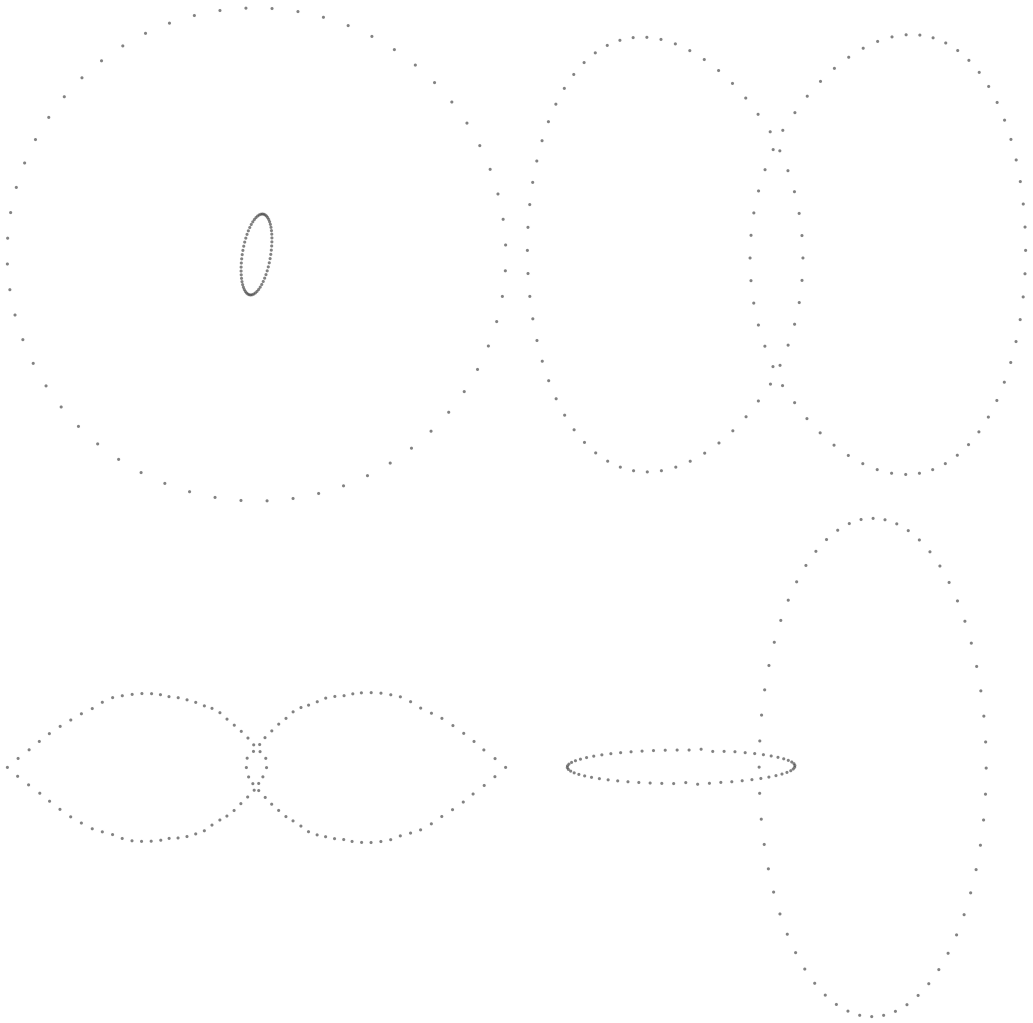


Figure 3.21: Entangled rings: Neither Isomap (left) nor LLE (right) achieve a clear separation or depict the circular geometry of the rings. Image courtesy of [Kar14].

unfolding of the hierarchy leads to the identification of four clusters, one corresponding to each of the three classes and one to the intersection region of the two intermingled clusters. However, with higher weight on shape in our clustering, the classes are visualized with almost perfect separation. The projection quality suggests that the cluster in the middle of the three is best represented by its local projection, while the others exhibit non-linear and non-elliptic shape.

Figure 3.23 shows the results of Isomap and LLE for the Iris data. Both methods generally fail to visualize the shape and alignment of clusters and distort the data. This clearly shows the benefit of the level-of-detail approach of our method.

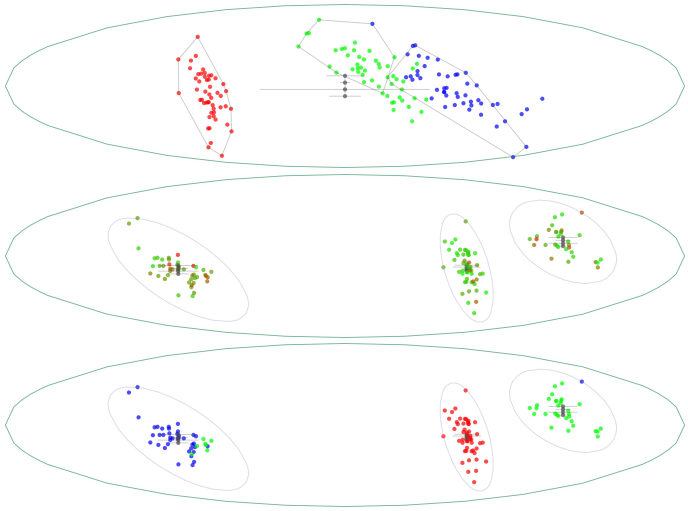


Figure 3.22: Iris flower data: After the initial view (top), the next level of detail shows the three flower classes each in a separate projection (middle), corresponding to an excellent classification (bottom). Image courtesy of [Kar14].

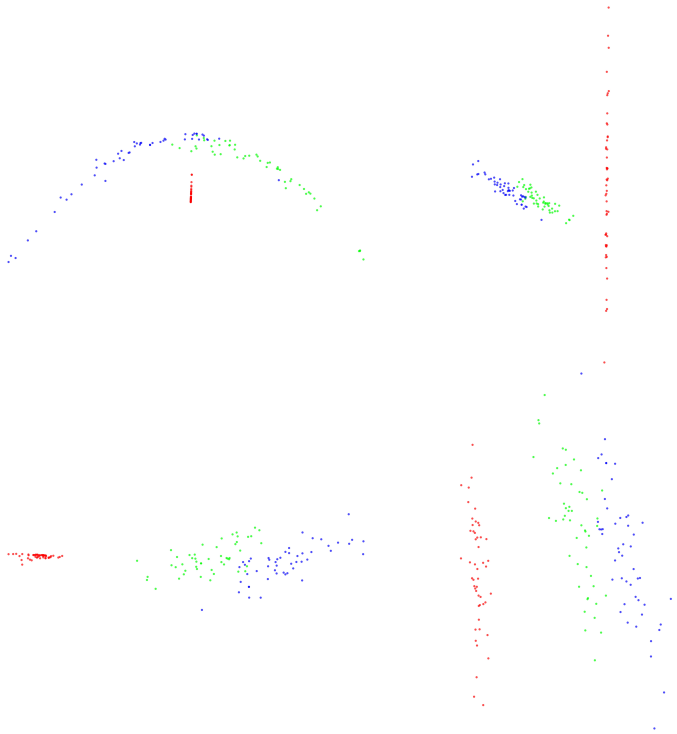


Figure 3.23: Iris flower data: left column: The results obtained from Isomap (left) and LLE (right) are highly diverse, depending on the neighborhood parameterization. Nevertheless, both methods distort the data set in all tested parameterizations, thereby, failing to convey the shape and alignment of the clusters. Image courtesy of [Kar14].

3.3.3 Conclusions

This section describes a new method for the visual exploration of multivariate data that is based on a level-of-detail approach involving dimension reduction and graph abstraction. We counter typical shortcomings of neighborhood approximation by employing a relative neighborhood graph and define a clustering method based on the Mahalanobis distance that abstracts this graph in different levels of detail. These levels are made visually accessible to the user by embedded local projections fitted into ellipses corresponding to the distribution of the data in the local approximation of the manifold. By embedding local projections, each projection utilizes full degrees of freedom, thus achieving an optimal representation for every detail level of the data being investigated.

A projection quality measure facilitates further understanding about the distance of each projected point orthogonal to the projection plane. Together with the means to parameterize the navigation between different hierarchy levels, these concepts form a powerful method for interactive exploration. Experiments show that our method captures the shape and topology of data features extremely well and significantly outperforms related methods on data sets that exhibit clusters suitable for level-of-detail analysis.

CHAPTER 4

Model-based Visual Analysis

From the beginning, this dissertation was set out to solve a single but very important application problem. The concepts of exploratory visual data analysis that have been known to succeed in numerous applications were to be applied to the specific problem of analyzing data from single particle mass spectrometry in the application of air quality research. It was not without struggle that this author has realized that these concepts alone cannot resolve the underlying analytical problem. Only a holistic scientific workflow can facilitate physically meaningful analysis of the application problem. Triggered by the problem in air quality research, this chapter describes how the workflow can be designed that not only solves the specific problem in air quality research but can also be regarded as new basic research in visualization that can be applied to a broad range of other application areas. Our main contributions are:

- Visual interface for steering and verification of non-negative matrix factorization
- Methodology to analyzing approximation errors of non-negative matrix factorization
- A general design methodology for model-based visual analysis

The application problem being solved can be summarized as follows. The data provided from single particle mass spectrometry (SPMS) contains mass spectra of individual collected airborne particles. A mass spectrum describes a distribution of ions by mass. Discretized and normalized, these spectra are stored and interpreted as points in high-dimensional space for consecutive analysis of the comprised chemical compounds of each particle. As mass is ambiguous, various sources may contribute to each dimension of the data and the spectra do not lend themselves to a straightforward deduction of chemical compounds. Instead, a basis transformation is to be found that models the observed spectra as a linear combination of the spectra arising from each of the comprised chemical compounds, such that linear combinations of this basis forms the observed mass spectra. In particular, the basis vectors are not mutually orthogonal, leading to the ambiguity in the data. Physical and chemical constraints further dictate both basis and coefficients to be non-negative.

Common methods for dimension reduction, for example, spectral decompositions like principal or independent components analysis, cannot find a basis transformation that accounts for non-negativity in both the mass spectrum and its combinatorial model. The

results are not meaningful in terms of atmospheric processing, as they show no relationship to air pollutant emission sources. Thus, atmospheric scientists are ambivalent about using dimension reduction and data analysis is mainly based on clustering techniques. However, clustering also provides unreliable results, as it fails to resolve ambiguity in the data due to its dependency on geometric distance measures. Hence, the individual mass spectra comprising the clusters are inspected, interpreted, and re-classified in a copious amount of manual effort. This is not only time-consuming but ultimately based on the subjective interpretation of an expert analyst without a fundamental mathematical underpinning.

We have studied this problem in a joint collaboration between the domains of air quality research and visualization. Dealing effectively with this problem necessitates the practical need for visualizing the analytical reasoning of a specifically designed optimization process. Thereby, we instigate a research direction for visualization in which little prior work exists. We show in this chapter that visualization is essential in providing air quality researchers with the means of finding both physically and mathematically correct interpretable lower-dimensional basis transformations of their data. Our methodology involves the visualization of a non-convex, multi-criteria, and non-negative optimization process. It entails a visual interface to this optimization that (i) allows the scientist to be "in the loop" of computations and (ii) enables verify and control the optimization process, in order to steer computations towards physical meaning while maintaining mathematical rigor. Thereby, we contribute both to the field of air quality research by providing novel means for the research of aerosol source contributions, as well as to the visualization community by laying the groundwork for further research towards enabling physically meaningful and interpretable semi-automatic data analysis.

Having applied our method to data from biomass combustion sources, our collaborators have been able to (i) reproduce established findings in a matter of hours (where prior work took months), (ii) process and analyze a hundred times more spectra than in previous studies, and (iii) gain surprising new insights enabled by the visualization. The unparallel success of our method in this application naturally prompts the question of whether the underlying concepts of our approach could be applied to other applications to achieve equal satisfactory results. We evaluate this potential in the final section of this chapter and describe the general design methodology underlying our approach. We define model-based visual analysis as an integral approach to compute parameters of a pre-defined analytical model that map to given data. In particular, our design choices are intended to facilitate analysis of mathematical ill-posed problems. Our approach involves an intensive background and requirement analysis to clearly define the analytical question of the scientist mathematically. Subsequent design of the algorithm for parameter approximation incorporates user input and steering, such that the user can be in the loop of computations. By representing analytical parameters and approximation errors in data space, we incorporate user knowledge and involve the scientist in decision making. Finally, this design involves interaction mechanisms for interactive steering, as well as approaches for verification and analysis of approximation errors.

The remainder of the chapter is structured as follows. *Section 4.1* describes the application problem that has triggered this new basic research. After ab intensive review of related work in *Section 4.2*, we describe the design of an interactive visual interface for non-negative

factorization of single-particle mass spectrometry in *Section 4.3* [EGG*12]. Further, we describe how computations can be verified and our methodology extended by the visual analysis of approximation errors in *Section 4.4* [EHH*13]. Finally, we generalize the underlying principles of our work in *Section 4.5* [EHHS14] by describing a design methodology to model-based visual analysis. We identify a broad range of applications that can benefit from our approach.

4.1 Single particle mass spectrometry in air quality research

Atmospheric particles have been shown to increase morbidity and mortality in urban areas and to alter the Earth's radiative energy balance. An important step in tackling this problem is to elucidate the chemical compounds of ambient airborne particles. A single particle mass spectrometer (SPMS) now chemically analyzes individual aerosol particles in real time, providing unprecedentedly rich data sets for air quality and climate research.

SPMS instruments collect, filter, and characterize aerosols based on their mass spectrum. The spectrum is measured by the instrument based on particle ionization. In simple terms, this process can be described as follows. Individual aerosols are filtered, accelerated through a drift tube and hit by a plasma beam. The beam causes the particle to ionize and break up into fragments of different compositional levels. The mass and relative abundance of fragment ions within the particle is then computed by a time-of-flight analyzer, providing the mass spectrum. Thus, the mass spectrum represents a function mapping *mass over elemental charge (m/z)* of fragment ions to their abundance in the particle. Discretized in bins of 1 m/z step size, the analyzer captures the first 256 m/z ratios for each aerosol. The resulting histogram data is stored as a 256-dimensional vector, where each coordinate corresponds to the abundance of fragments within the aerosol having an m/z ratio within the dimension's section of the discretized spectrum.

As mass is ambiguous, various ions may contribute to each coordinate/dimension, thus, SPMS mass spectra do not lend themselves to a straightforward deduction of chemical compounds. This phenomenon is known as *isobaric interference*. Data transformations based on geometric distance metrics are unable to resolve this ambiguity, as they rely on the comparison of coordinates which, in SPMS data, can stem from different physical sources. Consequently, data clustering results are not reliable and the state-of-the-art approach involves verifying each individual (mean) representative spectrum by the scientist. Figure 4.1 gives an example for this analysis. In the figure, individual peaks are labeled by their source contribution. Ambiguity is resolved by manual analysis and experience.

Ambiguity has created several problems for the application and has in turn triggered basic research problems for visualization. In air quality research, the task does not match the available data. SPMS data is high-dimensional, noisy and ambiguous. The application's goal is categorizing particles by their composition; however, their composition is not reflected by the mass spectrum because of its ambiguity. Consequently, a basis transformation is needed that resolves this ambiguity.

According to the underlying physics, it is assumed that each particle can be described by the linear combination of its fragment ions. As such, SPMS data $X \in \mathbb{R}_+^{(n \times m)}$, holding n particle spectra discretized in m dimensions, can be described by the discretized mass

spectra of their fragment ions as a basis $B \in \mathbb{R}_+^{(k \times m)}$ and coefficients $C \in \mathbb{R}_+^{(n \times k)}$, such that

$$\begin{aligned} X &= CB + N \text{ and} \\ X_{i,\bullet} &= \sum_{1 \le j \le k} C_{i,j} B_{j,\bullet} + N_{i,\bullet} \,. \end{aligned} \tag{4.1}$$

Here, B is the matrix storing (row-wise) basis vectors, $B_{j,\bullet} \in R_+^m$, $1 \le j \le k$, such that X is derived with the coefficient matrix C and the noise N induced by the instrument. Note that $X_{i,\bullet}$ and $N_{i,\bullet}$, $1 \le i \le n$, refer to the rows of the matrices X and N. Further, all values are positive, reflecting the combination of mass spectra of fragment ions. The problem is ill-posed and there is no unique solution because (i) C, B, and N are unknown and (ii) any change in B can be undone by changing C accordingly. In particular, adding arbitrary basis vectors to an ideal set B^* with appropriate coefficients (adding up to zero) does not change the solution. From the standpoint of numerical optimization, infinitely many bases exist with which the data can be described. Due to the complexity of the problem, the level of uncertainty involved in data collection, partially unknown machine-dependent physical models of particle fragmentation and noise induced by the instrument, previous methods have failed to deliver physically correct independent data bases. To conduct meaningful analysis, atmospheric scientists were required to inspect thousands of spectra manually in an error-prone and time consuming process.

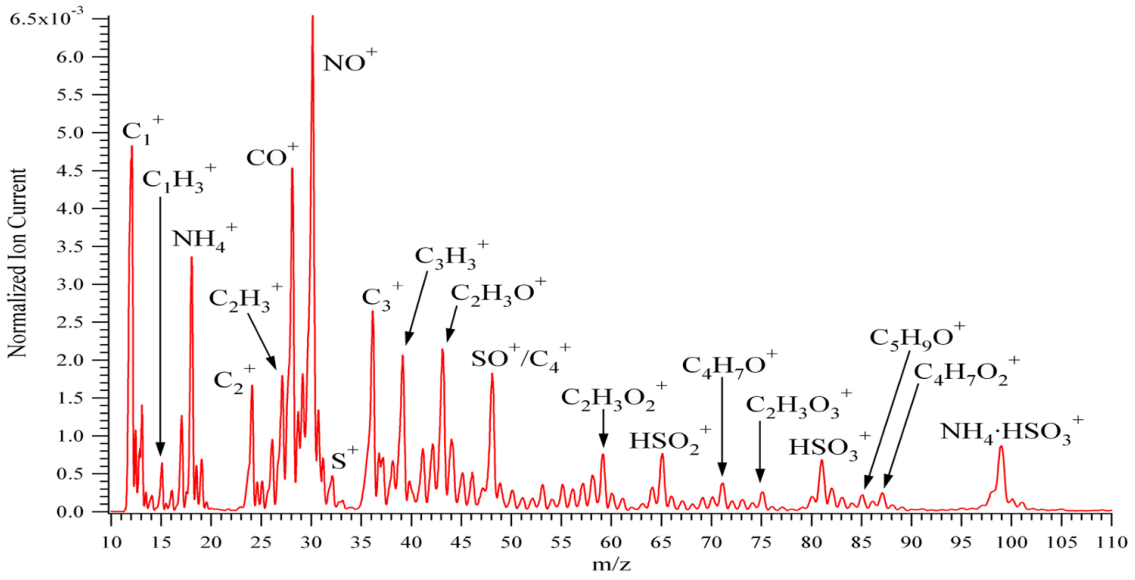


Figure 4.1: As of now, atmospheric scientists estimate aerosol composition and the sources that contribute to each m/z dimension based on their experience in investigating mass spectra. Analyzing and classifying thousands of particle spectra can take months even with the help of data clustering.

4.2 Related work

Work related to the scope of this section can be found in the fields of dimension reduction, visualization, and air quality research; we provide a brief overview in the following.

Dimension reduction

Methods for dimension reduction compute a mapping from high- to low-dimensional space. Thereby, data is decomposed into a set of new coordinates, acting as coefficients to a different basis that is more suitable with respect to data properties. In visualization, dimension reduction is commonly applied as a means of finding a lower-dimensional data embedding that best reflects distance relationships between high-dimensional points. Here, the focus lies not on the properties of the basis but on their coefficients that act as an abstraction to the high-dimensional data by mapping distance relationships. In contrast, we are also concerned with the basis of this mapping, as its interpretation is essential in air quality research.

Methods like multi-dimensional scaling [Tor58] or manifold learning [TSL00b, RS00b] define data bases in inner product space. These bases prove hard to interpret as their dimensionality equals the number of data points. While principal components analysis [Pea01] finds bases in data space, it is restricted to orthogonal basis vectors. Overcoming this restriction, independent components analysis [Com94, HO00] finds non-orthogonal independent data bases. However, this method is incapable of producing a non-negative basis transformation [CP07].

Matrix factorization (MF) methods offer more degrees of freedom for defining optimization goals. In particular, non-negative matrix factorization (NMF) has recently received great attention because it is capable of computing non-negative basis transformations. Non-negativity is an integral property for application areas that investigate physical phenomena described by non-negative measurements or mixtures, as it is the case for the type of air quality data we are concerned with. Here, we make use of the works of [KP08] and [WR10]. The former provides a framework for alternating non-negative least squares, while the latter shows how the use of a decorrelation regularization term derives independent components in non-negative data. In contrast to previous work, no matrix inversion is necessary for this computation. Other work offers a convex model to NMF but is constraint in expressing its basis as a convex combination of data points [EMO*11].

As MF methods are based on numerical optimization, their drawbacks lie with convergence speed, as well as their proneness to local minima. The latter is a seemingly insurmountable problem, as the optimization problem described in the previous section is non-convex. Finally, all methods mentioned above share the common problem of finding a basis transformation that is not only numerically correct but also physically interpretable by domain experts. With complex physical restrictions that are not (yet) well-defined, it is impossible to extract a physically correct data basis through numerical optimization alone. After all, in air quality research, as well as in most scientific applications, the motivation for data analysis is that data properties are part of open research questions. Consequently, the scientist has to be involved in interacting with computations. This insight motivates the present work.

Visualization

While visual steering of exploration [SLY*09] and simulations [KTH*02, LGD*05, WFR*10] have become well-established research areas in the field of visualization, the visual steering of practical engineering optimization has not been the focus of previous studies. However, this need is clearly documented [MC00] and recent advances from the engineering community give empirical evidence to the benefits of interactive visualization-based strategies to support engineering design optimization and decision-making [CMKS08].

Recently, novel techniques have been introduced to the visualization community that enable user interaction in dimension reduction and have demonstrated great success in application areas. For example, piecewise Laplacian-based projection (PLP) [PEP*11] allows the user to interact directly with the mapping process by providing means to adjust neighborhood and distance approximations. Thus, the user can implicitly redefine the basis for dimension reduction. However, the application of PLP does not involve analysis of this basis or the mapping error. Previous methods have used dimension reduction as a means to visualize data. In contrast, our application requires visualization as a means to steer dimension reduction toward physical correctness. In this regard, air quality research may prove to trigger a new research domain for visualization.

Techniques to visualize matrix factorization can, in part, be based upon existing research in multi-dimensional data visualization [WGK10]. Driven by applications, research focuses on better representation of specific data properties (e.g., scientific point cloud data [OHJS10]), better incorporation of domain-appropriate analysis techniques (like brushing and filtering [JBS08]), or computational speed gains [IMO09]. Opposed to visualizing data relationships by dimension reduction, value visualizations like heat maps, glyphs, scatter plot matrices, and parallel coordinates, are regarded as the dominant approach to multi-dimensional data visualization. Research in this area has focused on enhanced cluster visualization [JLJC05, ZYQ*08, AdO04], brushing techniques [EDF08, HLD02], and abstraction [MM08]. Clutter reduction through dimension ordering [PWR04, YPWR03, FR11] is one of the most promising approaches to enable data comprehension.

Visualization in air quality research

The air quality research community has recognized the need for tools to assist the interpretation of data from single particle mass spectrometers that generate many spectra of high dimensionality. Published under the synonym positive matrix factorization, NMF has been used for classification of airborne particle types [KBHH05]. However, the focus has not been on deriving an independent basis or visualizing the result. As of now, the dominant approach to mass spectrometry data analysis is to apply clustering methods. These methods are commonly based on hierarchical [MMW03], neural network [SHFP99], or density [ZHP08] schemes that utilize geometric distance measures. Available software packages for mass spectrometry data ([GAR*10], [ZIC*06]) include a variety of data mining methods that focus on clustering and the visual analysis thereof. Recently, a system allowing visual analysis and steering of data clustering has been introduced [ZIN*08]. Here, better clustering results have been obtained through the incorporation of expert knowledge into data clustering and means for refinement of prior solutions.

4.3 Model-based visual analysis of single particle mass spectrometry

Looking beyond the scope of air quality research, the underlying optimization problem described in the previous sections is not uncommon and known as blind source separation [CJ10]: given data that is derived from a combination of unknown compounds in unknown abundance and combination, the goal is to factor out both unknowns, provided only with an estimate of the number of compounds and an assumption of their mixing model. In particular, the mass spectra of the inter-mixed compounds are not mutually orthogonal which leads to ambiguity in the data. Thereby, the compounds and their combination represent the actual "hidden" information that is to be factored out from the given mixture. In the context of atmospheric processes, the unknown sources represent the latent components that appear independently in aerosols. These form the actual, unambiguous, and lower-dimensional data basis. Based on the physical model of particle ionization, each measured aerosol mass spectrum can be described by the linear combination of the mass spectra of independent compositional sources. As a consequence, the latent independent variable basis of the data and the corresponding coefficients that derive the data are, in theory, exactly the independent physical components and their occurrences in the aerosols. However, in practice, extracting these basis vectors and coefficients so that their physical composition can be interpreted proves difficult. While latent variable bases can be computed by dimension reduction, these bases are often hard to interpret, and, although numerically correct, not necessarily physically meaningful to atmospheric scientists analyzing the data. In particular, methods like independent components analysis [Com94] do not account for non-negativity in the mass spectrum and physical source mixture model. As a result, air quality researchers are ambivalent about using dimension reduction when the resulting data basis does not show a relationship to air pollutant emissions and their atmospheric processing.

In this work, researchers, from the domains of air quality research and visualization, jointly studied this problem to investigate the positive influence of visualization on solving this problem. Dealing effectively with these vast, high-dimensional data sets necessitates the practical need for visualizing the process of dimension reduction, thereby instigating a research direction for visualization in which little prior work exists. We show in this section that visualization is essential in providing air quality researchers with the means of finding unambiguous physically correct lower-dimensional basis transformations of their data. Our method involves the visualization of non-convex, multi-criteria, and non-negative matrix factorization. Further, our method entails a visual interface to this optimization process that (i) allows the atmospheric scientist to be "in the loop" of the computation, (ii) provides direct visual feedback of the optimization process, and (iii) enables controlled refinement of its solution. We introduce domain-specific visual encodings and interactive mechanisms of matrix factorization that provide the means to incorporate expert knowledge into numerical optimization and to steer this process toward physical meaning while maintaining mathematical rigor. Thereby, we contribute both to the field of air quality research by providing novel means for the research of aerosol source contributions, as well as to the visualization community by laying the groundwork for further research toward enabling physically meaningful and interpretable dimension reduction.

The remainder of the section is structured as follows. Section 4.3.1 provides the requirement analysis and task description that results from the application problem described in Section 4.1. Our framework, consisting of data factorization, domain-specific visual encodings and interaction mechanisms, is described in Section 4.3.2. In Section 4.3.6, this framework is applied to the factorization of biomass combustion particle spectra and evaluated with respect to its ability to produce new insights to the application of air quality research. Finally, concluding remarks are given in Section 4.3.9. The work is published in [EGG*12].

4.3.1 Requirement analysis

Based on the interdisciplinary research involved in this work, we have identified key requirements for implementing a basis transformation of SPMS data as described in the previous section, as well as the tasks that define the usage of such a system.

Many optimization methods are static in nature and often output their result in combination with a single error measure. For air quality research, however, this is not sufficient. An overall mapping error of "2.05", for example, provides close to no insight for analysts. Atmospheric scientists have extensive experience in interpreting SPMS data, identifying different sources, common features, and noise. Under the mathematically ill-posed problem of finding physically correct bases for the un-mixing of SPMS sources, analytical frameworks are well-advised to draw upon the scientists' knowledge in order to generate more reliable results.

As numerical error measures hardly facilitate understanding of data features or the physical correctness of the factorization basis, visualization of the basis transformation is inevitable when the scientist is to assess its correctness. Thus, visualization should permit a detailed understanding of

- the exact mapping error induced by the basis transformation with respect to the data and
- the basis vectors used in the transformation with respect to the data features.

Further, an interface is required to perform the following tasks.

- **Analyze basis:**
  In a correct factorization, each basis vector equals the mass spectrum of a stereotypical particle fragment ion that is part of the measured aerosols. The basis should therefore be depicted in a way that makes possible easy comparison between different basis vectors and how they relate to the underlying data features, i.e., the different sources in the data set. In this comparison, analysts should be able to identify the relative peak heights and sparsity of the different basis vectors and verify that the basis describes physically meaningful parts of aerosols.
- **Analyze mapping errors:**
  To verify the correctness of the basis transformation, the scientists needs to analyze how well each part of the data is captured by the factorization. In part, this error may stem from noise or outlier measurements. Both overview and detail is required such

that the overall fit, as well as the specific error with respect to certain dimensions or data points, can be assessed.

- **Adjust basis:**
  If the basis is found to be physically incorrect, intuitive means for interaction are required to adjust the basis accordingly. Also, means to account for the level of uncertainty in the basis configuration are desirable, since exact peak ratios may vary depending on measurement parameters but certain conditions of the basis configuration are known and can be specified.

The exact coefficients of the basis are not of immediate interest for verifying the factorization but for successive analysis steps. Instead, the focus lies on the error that is induced by the basis transformation, as well as on the physical correctness of the basis. Computations should be based on the physical data model, mathematically well-founded, and convey physically interpretable results. For visual user feedback, visualization is essential in conveying an understanding of the factorization's intermediate results, as well as to offer an effective interface to verify and control computations with all involved parameters.

4.3.2 Method

The goal of the method presented in this section is to provide atmospheric scientists with the means of finding unambiguous physically correct lower-dimensional basis transformations of single particle mass spectrometry (SPMS) data. Our method involves the use of non-negative matrix factorization (NMF) in combination with specific regularization terms to find a basis transformation of SPMS data that minimizes ambiguity. Further, our method entails a visual interface to this optimization process that allows the analyst to be "in the loop" of the computation, provides direct visual feedback of the optimization process, and allows controlled refinement of its solution. By introducing domain-specific visual encodings and interaction mechanisms of SPMS factorization, we provide means to incorporate expert knowledge into the numerical optimization in order to steer this process and to verify the physical correctness of the basis transformation.

4.3.3 Non-negative matrix factorization

Given n data points of dimension m with non-negative coordinates, $X \in \mathbb{R}_+^{(n \times m)}$ and a positive integer $k \in \mathbb{N}, k > 0$, methods that perform non-negative matrix factorization (NMF) find a factorization of X into a basis $B \in \mathbb{R}_+^{(k \times m)}$ and coefficients $C \in \mathbb{R}_+^{(n \times k)}$, such that

$$||X - CB|| \rightarrow min, \tag{4.2}$$

where all values in C and B are non-negative.

From the class of existing methods for NMF, we use a combination of [WR10] and [KP08], together with two regularization terms of independence and diversity that serve the objective of finding an unambiguous basis and reliable solution, respectively.

The objective of *independence* is understood by taking into account the physical process of particle ionization induced by the SPMS instrument. The contribution of fragment

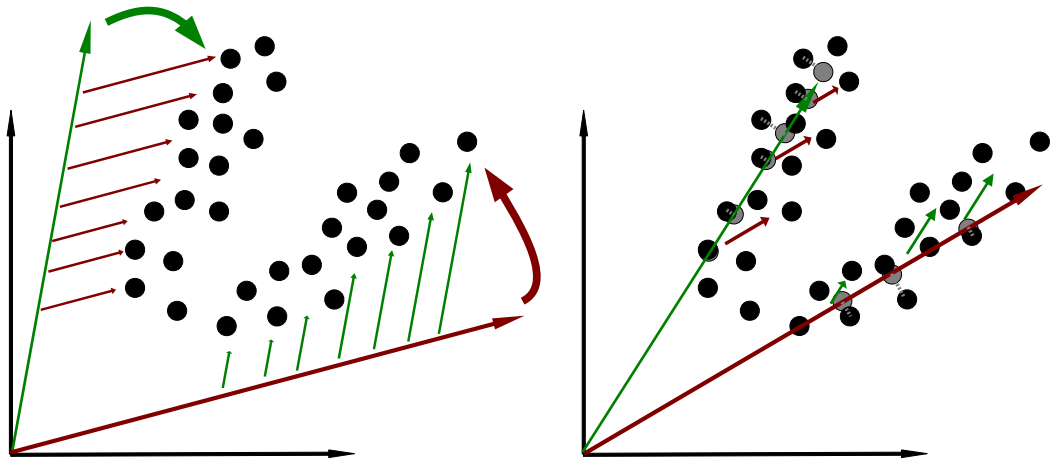


Figure 4.2: With regularization, NMF finds a non-negative factorization in coefficients of independent basis vectors (particle components). Thereby, the correctness of the mapping (errors are illustrated in gray) is balanced against the independence and diversity of the basis.

ions to the particle's mass spectrum can be modeled in a hierarchical manner, as larger fragments ionize into smaller fragments. Thereby, the molecular compositional parts of the particle contribute to the particle's mass spectrum not only by their own m/z ratio but also by those of their successive sub-fragmentations. For statistical considerations, these fragmentation patterns are constant. Consequently, SPMS data can be described by a latent variable model of independent components that represent fragmentation patterns. It has been shown that, in order to derive the independent components from a non-negative matrix, it is sufficient to find a factorization into a non-negative basis and coefficients for which the coefficients of the basis vectors are uncorrelated [WR10]. Defining the optimization goal of independence between the basis vector's coefficients, therefore, serves not only the purpose of dimension reduction (basis transformation into independent latent variables) but also leads to basis vectors of distinct molecular composition. Thus, the regularization of independence aims toward an unambiguous and physically interpretable basis.

Non-convex optimizations are prone to lead to only locally optimal results. Our experiments of factorizing SPMS data have shown that the objective of independence leads to the fact that outliers are not mapped well with this criterion alone and the optimization of the basis may become "entrapped" by the local solution for independence. Although we present no empirical evidence for this fact, the intuition behind these dilemmas is clear. From an optimization perspective, the gain in correctness by changing the basis toward faces of the bounding box of the data does not outweigh the penalty of correlation induced by this change. Consequently, the optimization terminates in a local optimum. Although independent components of SPMS data have to be *diverse*, early implementations using only the regularization of independence have not produced this result but have ended abruptly in unreliable solutions. However, steering the search for an independent basis

by including a slightly weighted regularization of diversity has produced far more reliable results.

To summarize these considerations, the non-negative matrix factorization of SPMS data has the following objectives.

- NMF: C and B define a factorization of X for which $||X - CB||$ is minimized
  $\rightarrow$ numerically correct
- Regularization w.r.t. independence: basis coefficients C are decorrelated
  $\rightarrow$ unambiguous basis
- Regularization w.r.t. diversity: basis vectors B are mutually different
  $\rightarrow$ reliable solution

We use a combination of [WR10] and [KP08] that involves a gradient-based two-block optimization scheme with multiplicative update rules according to [LS00]. The computations can be summarized as follows, where $||.||_F^2$ denotes the squared Frobenius norm.

1. **Numerical correctness** is enforced by multiplicative update rules in two successive blocks:

$$min_{C \geq 0} \qquad ||X - CB||_F^2 \text{ , by} \tag{4.3}$$

$$C_{a,b} \leftarrow C_{a,b} \frac{\left([XB^T - \alpha_C R_C]_{\geq \varepsilon}\right)_{a,b}}{(CBB^T)_{a,b} + \varepsilon} ,$$

where B is fixed and

$$min_{B \geq 0} \qquad ||X - CB||_F^2, \text{ by} \tag{4.4}$$

$$B_{a,b} \leftarrow B_{a,b} \frac{\left([C^T X - \alpha_B R_B]_{\geq \varepsilon}\right)_{a,b}}{(C^T CB)_{a,b} + \varepsilon} ,$$

where C is fixed. Here, $[.]_{\geq \varepsilon}$ denotes that values are truncated to be greater or equal to a small positive real value. The update rules employed here are inherently "normal" additive gradient updates with a relative step size. However, this multiplicative formulation yields faster computations and is currently the dominant approach to NMF [LS00]. The regularization terms R_C and R_B are weighted by α_C and α_B, respectively, producing an independent and diverse basis.[1] R_C and R_B are the partial derivatives of the cost functions $J_C(C)$ and $J_B(B)$, i.e.,

$$R_C = \frac{\partial J_C(C)}{\partial C} \text{ and} \tag{4.5}$$

$$R_B = \frac{\partial J_B(B)}{\partial B} . \tag{4.6}$$

1 Note that α_B should be distinctly lower weighted than α_C.

Detailed notations of R_C can be found, for example, in [WR10].

2. **Independence** is enforced by updating C using the partial derivative of the cost function $J_C(C)$ that defines the discrepancy between the *uncentered* correlation (normed uncentered covariance) matrix of C, $nCorr(C)$, to the $k \times k$ identity matrix, I_k.

$$\begin{aligned} J_C(C) &= || \; nCorr(C) \; - \; I_k||_F^2 \text{ , with} \\ nCorr(C) &= N_C C^T C N_C \text{ ,} \\ N_C &= diag(1/||C_{\bullet,1}||_F, ..., 1/||C_{\bullet,k}||_F) \text{ , and} \\ ||C_{\bullet,i}||_F &= \sqrt{\sum_{1 \leq l \leq n} C_{l,i}^2} \text{ .} \end{aligned} \tag{4.7}$$

3. **Diversity** is enforced by updating B using the partial derivative of the cost function $J_B(B)$ that defines the discrepancy between $cos(B)$, the $k \times k$ matrix of the cosines of the angles between all basis vectors in B, $cos(B)$, to the $k \times k$ identity matrix, I_k.

$$\begin{aligned} J_B(B) &= || \; cos(B) \; - \; I_k||_F^2 \text{ , with} \\ cos(B) &= N_B B B^T N_B \text{ ,} \\ N_B &= diag(1/||B_{1,\bullet}||_F, ..., 1/||B_{k,\bullet}||_F) \text{ , and} \\ ||B_{i,\bullet}||_F &= \sqrt{\sum_{1 \leq l \leq m} B_{i,l}^2} \text{ .} \end{aligned} \tag{4.8}$$

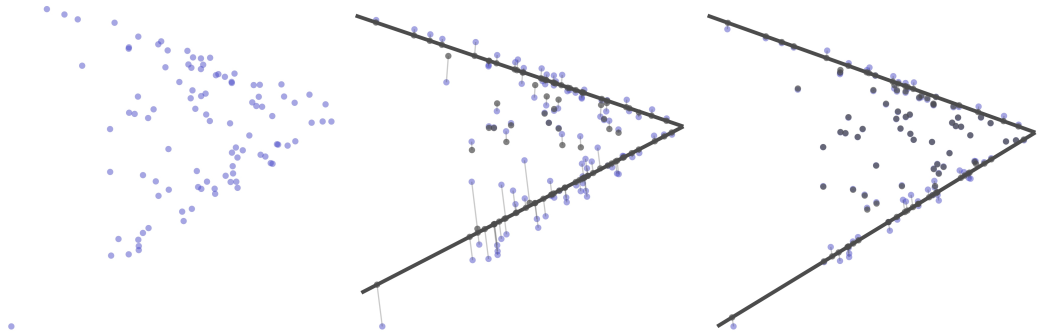
Note that computation of the partial derivative of J_B is algorithmically equivalent to that of J_C. Instead of minimizing correlation between columns in C ($C_{\bullet,i}$), R_B maximizes the cosine of the angle between rows (basis vectors) of B ($B_{i,\bullet}$). Figure 4.3 provides an example of how this NMF implementation behaves in two dimensions.

4.3.4 Visual encodings

Although the numerical optimization, as presented in the previous section, has desirable mathematical qualities with respect to factorizing mass spectra, results are not guaranteed to be physically meaningful to domain experts. Due to noise, outliers, or local optimality, the verification of the solution is required by scientists. In contrast to previous approaches using numerical error metrics, we provide domain and problem-specific visual representations of the matrix factorization process. This enables scientists to analyze the optimization result in full detail, as well as to assess its quality on an abstract level. In the following, we give a detailed account of these visual encodings and discuss their suitability for air quality research.

The factorization of a data set $X \in R^{(n \times m)}$ introduces additional entities that require visual representation to make the user aware of their properties:

- the basis $B \in R^{(k \times m)}$,
- the coefficients $C \in R^{(n \times k)}$,



(a) **(b)** **(c)**

Figure 4.3: An example of our NMF implementation. (a) shows a 2D point arrangement that reflects the geometric properties of the SPMS data model as described in Section 4.1, has two fairly independent sources, and noise added to it. Basis vectors and coefficients are computed by steepest descent (b), after which the unknown non-negative mixture of the two contributing sources is found almost exactly (c).

- the mapping error $X - CB$, as well as its metric sum $||X - CB||_F^2$,
- the correlation of the basis vectors' coefficients $nCorr(C)$, and
- the cosine of the angle between basis vectors $cos(B)$.

However, not all entities are of equal interest (or importance) in the validation of SPMS matrix factorization. Most importantly, B must be visualized, as the physical correctness of the basis defines the value of the factorization. If the basis vectors cannot be interpreted as a meaningful physical entity in the application, the factorization will hold no physical meaning to scientists. Enabling the analysis and validation of B is therefore a key requirement to be met by the visualization. Consequently, the visualization of B must show each basis vector in full detail.

As the basis is to be evaluated in relation to the data, the visualization must also involve the depiction of X in equal detail. This leads to the conclusion that X and B have to be depicted in the same visual space and form, such that the user can visually reference both entities in relation to each other more easily. Both B and X are histogram data for which several alternative visual representations exist. However, mass spectrometry data has already an established visual representation in the engineering community which novel visualizations have to conform to. As Figure 4.1 shows, mass spectrograms are visually represented as piecewise linear functions over mass. The effectiveness of our method depends on the scientist's ability to investigate patterns in these spectra, as well as on the experience in identifying diverse sources contributing to SPMS data. The basic geometric form for the visualization of X and B must therefore relate to the existing representation of the histogram data. If X and B are seen as multidimensional, then piecewise linear functions equal the representation by parallel coordinates [Ins09b] with the following exceptions:

- No vertical lines are drawn for the dimensions.

- Every dimension has equal scale.
- The order of dimensions is not arbitrary but given by mass ratios.

These exceptions render the application of state-of-the-art visualization techniques for parallel coordinates (see Section 4.2) very limited. For visualizing SPMS data, dimension reordering, edge bundling, or other forms of abstraction are generally less accepted by air quality researchers. However, transfer functions, alpha blending, as well as coloring schemes, that slightly affect the visual appearance but not the spatial presence of data features, are accepted degrees of freedom for the visual representation.

Next to the basis, the mapping error $X - CB$ is of equal importance in the verification of SPMS factorization. Even if the basis is verified to be physically correct, if the factorization does not hold for the data, it will be of little worth to successive analysis. While the overall quality of the factorization can be assessed by the norm $||X - CB||_F^2$, a detailed visualization of $X - CB$ gives clues as to how the basis can possibly be adjusted to achieve a better factorization result. NMF optimization is prone to local minima and the computed basis is most likely not optimal. However, adjustment of B with respect to the mapping error can improve results and lead to finding a global optimum. Consequently, a detailed visualization of the mapping error is equally crucial to the effectiveness of our method.

There are two possibilities for defining the mapping error: absolute ($X - CB$) and relative (CB/X), both are interesting for atmospheric scientists; even small measurements in specific m/z dimensions can be important. As numerical optimization based on least sum of squares tends to neglect small values, the visualization of the relative error, showing the factorized data in relation to the original data, is required for verification. On the other hand, the absolute error is of equal importance as it allows for the assessment of information loss and shows the patterns of features that are not captured in the factorization. Consequently, our visualization has to be able to depict both absolute and relative factorization error in detail, while the absolute error should be displayed in relation to the basis in order to give insights to its adjustment. Therefore, we depict $X - CB$ as separate polylines in the same axis as X and B. To reduce confusion between these three visual entities, we use distinct colors for each of them. We also allow for the option to hide the absolute error in case it should clutter the view.

The relative error should be visually distinguishable from the absolute error and, therefore, have a distinct representation. In our context, it shows the user how the factorization fits to each coordinate of the data, i.e., whether the mapping accounts for higher or lower values with respect to each value in X. To avoid further cluttering the visualization, we exploit the yet unused degree of freedom of color coding X by this error measure. Although this is intuitive to analysts, the drawback is that color may not be used for other visualization goals, for example, to depict different clusters. Here, we use alpha blending to bring out some of the data's structure. To easily distinguish under- and over-representation, and since relative errors CB/X generally show different distributions in these ranges ([0,1] and [1,2]), two distinct colors are required for the definition of the color map. For $CB/X = 1$, the user's attention is not needed as no error is present.

To this point, we have established the necessity of a single high-detail visualization showing B relative to X (colored by the CB/X) and $X - CB$. However, for analysis and

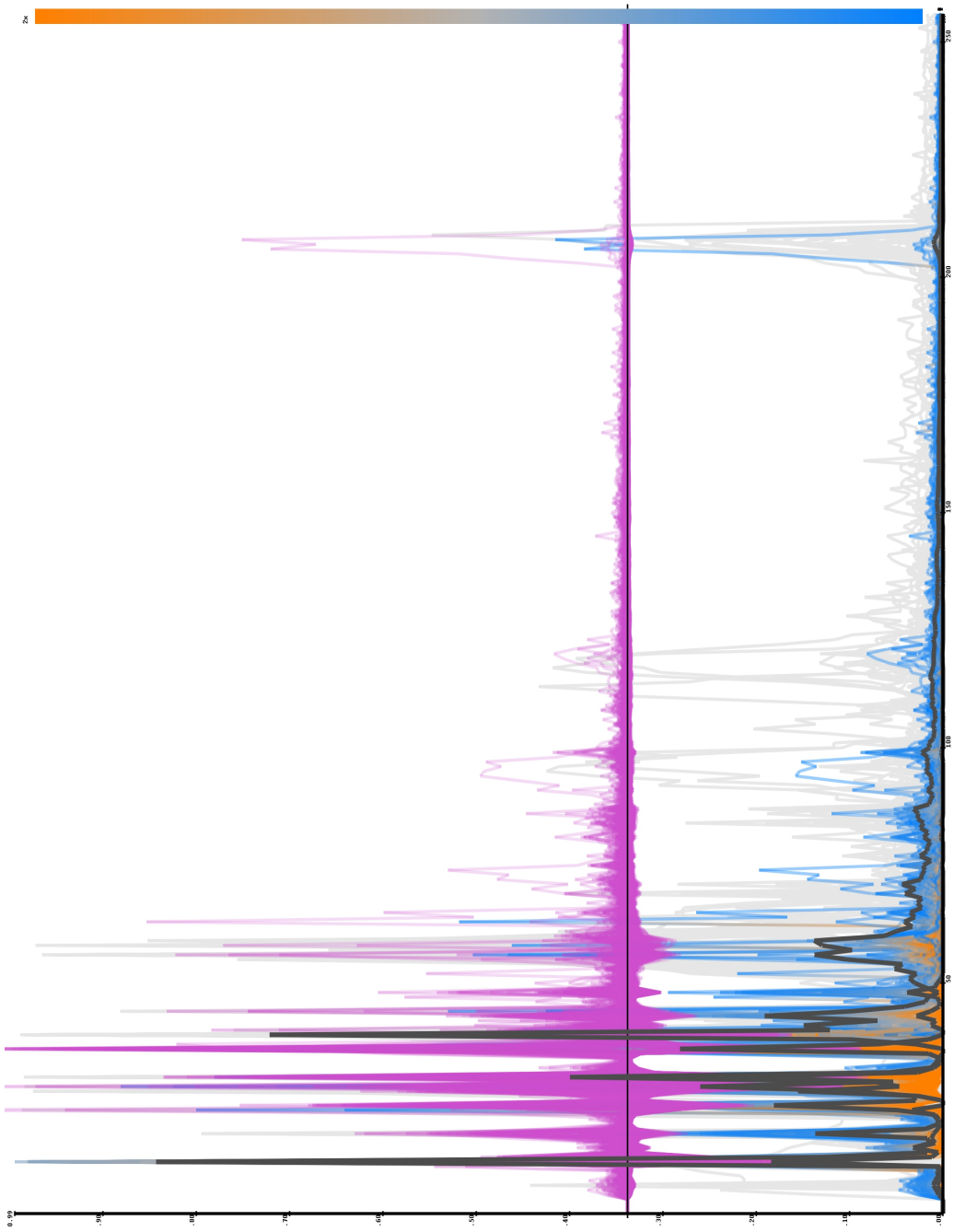


Figure 4.4: Overview of particle spectra in 256 dimensions showing absolute (magenta) and relative error (orange-blue) of the computed factorization by two independent basis vectors (dark gray). Spectra are represented as piece-wise linear functions over m/z and exhibit the same patterns as established representations.

verification of the factorization, the user also needs to quickly gain an overview of the general mapping quality. This overview should facilitate an abstract comprehension of the overall fit of the factorization and serve as a platform for interaction and navigation. We provide this overview by a linear projection of X, CB, and B, defined by the two principal components of the covariance matrix of X, as they are the orthogonal axes of maximal variance in X. Further, we connect each point in X with its corresponding mapping in CB. In this projection, the user can quickly identify outliers that are not mapped well by the factorization, as well as its general quality.

A third plot accounts for the visual representation of $nCorr(C)$ and $cos(B)$, as independence and diversity are important aspects of the optimization. However, their depiction is not as essential to data analysis as the projection or graph view. As heat map representations of correlation matrices are well-established in the engineering community and are also space efficient, we join both symmetric matrices by their upper and lower triangular half, respectively, and display their values by color coding in gray scale. Note that the basis has no inherent order and reordering of this matrix is semantically possible, however, we find no need to do so as our NMF implementation generally performs well with respect to independence of the basis.

4.3.5 Interaction

Effective visual analysis and manipulation of matrix factorization requires interaction. In this section, we introduce the means for interaction provided by our visual interface and comment on their eligibility for interactive visual verification of SPMS factorization. Figure 4.5 illustrates these techniques.

As established in previous sections, numerical error measures do not facilitate understanding of complex data features. The visual evaluation of the factorization is a necessary step in air quality research. Thereby, effective interaction techniques are required for atmospheric scientists to analyze the mapping error in each dimension with respect to outliers, noise, and data features. We distinguish between two interaction classes: *analysis* and *manipulation*. For the analysis part, we apply interaction techniques from visual analytics to our visualization of matrix factorization. However, few techniques are available that focus on interfacing dimension reduction methods. Here, we introduce novel and intuitive interaction mechanisms that interface matrix factorization.

The first step in assessing the quality of a factorization result is its visual analysis. Given the level of restrictions for visual encodings, parallel coordinates are inevitably prone to visual clutter with increasing number of data points and dimensions. Therefore, interaction is usually required for effective visual analysis. *Zooming and panning* allows for detailed analysis of specific parts of the factorization. In order to analyze a group of data points or basis vectors, we allow for problem-specific, semantic *selection and filtering* mechanisms. Thereby, the projection plot acts as the selection interface that induces filtering operation in the graph plot. Upon selection, the alpha values of all polylines are adjusted according to their analytic connection to the selection. The selection of points

- hides other points in X,
- hides absolute errors in $X - CB$ not stemming from the selected points, and

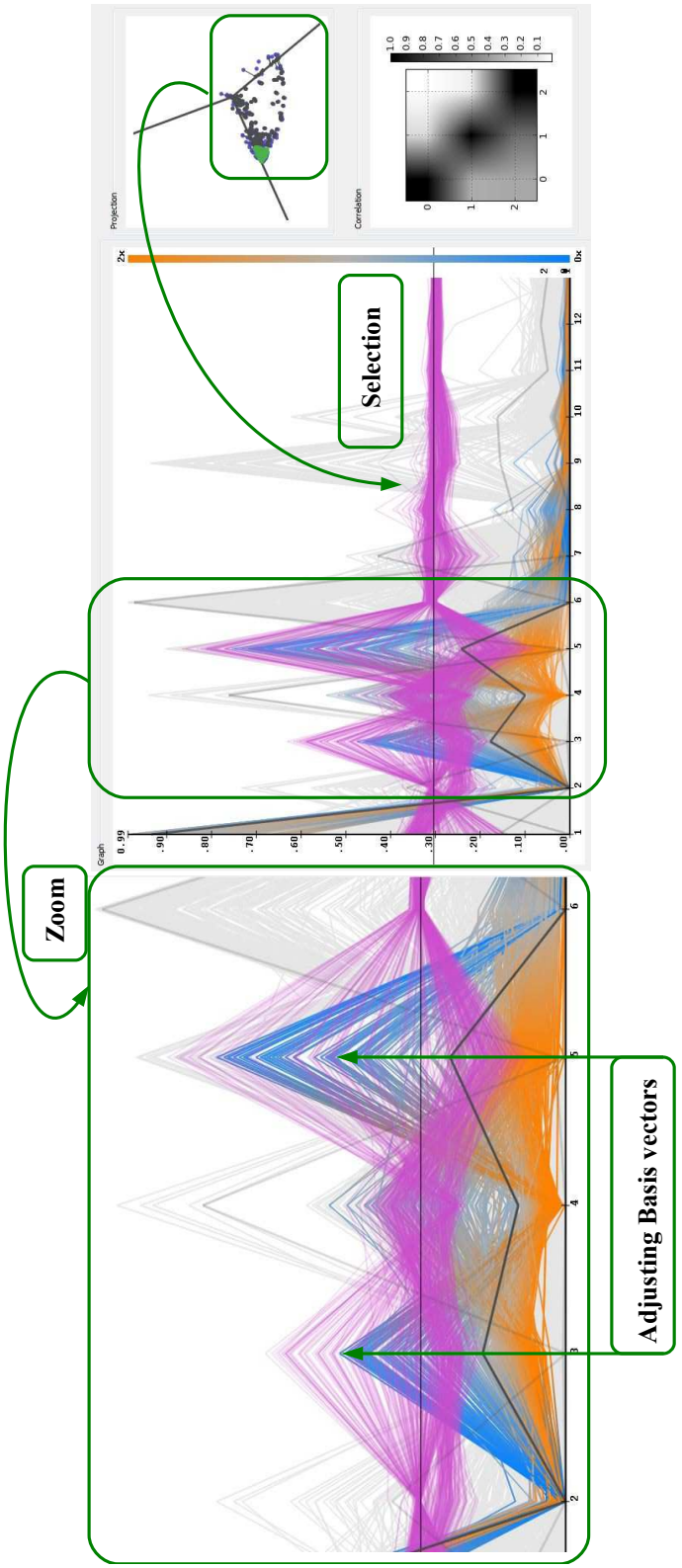


Figure 4.5: Using interaction techniques and filtering, atmospheric scientists can analyze the factorization error in specific dimensions, investigate the contribution of basis vectors to the mapping of the spectra. Initial results of the factorization can be adjusted by setting new starting parameters or thresholds for the optimization. Expert verification and steering aims toward physically sensible factorizations and interpretable results.

- shows basis vectors according to their coefficients in the mapping of the selected points, i.e.,
  $\alpha_b = \sum_{p\ selected} C_{p,b}\ /\ \sum_{p\ selected} ||C_{p,\bullet}||_1$ [1],

while the selection of basis vectors

- hides other basis vectors in B,
- leaves absolute errors in $X - CB$ untouched[2], and
- shows points according to their coefficients in the mapping of the selected basis vectors, i.e., $\alpha_p = C_{p,b}\ /\ ||C_{p,\bullet}||_1$ [3].

After thorough analysis of the factorization and its error, the atmospheric researcher may refine parameters for a successive optimization step, either from an adjusted or random starting point of parameter space. By interacting with the factorization, the analyst can *manipulate the basis* by adjusting coordinates of basis vectors via left-mouse dragging. As basis vectors are normalized by the optimization, the unadjusted coordinates are updated such that the norm holds for the adjustment. By permitting this manipulation, the user can iteratively define new starting points for the gradient-based search of an optimal data basis for factorization.

Non-convex gradient-based optimization can be unpredictable, especially when applied to high-dimensional data. As the definition of different starting points for optimization may not necessarily imply a different result after steepest descent, setting the starting point is not sufficient to assure physical correctness. Additionally, the user can set threshold values for the basis optimization. Relative from the position of each basis coordinate, positive and negative thresholds can be set by right-mouse dragging. These thresholds act as strict boundary limits for the steepest descent which are guaranteed to be met by the optimization. Setting boundary levels for only a few coordinates can change the entire basis in a way that the configuration is optimal regarding to the given restrictions, while the scientist can decide the exact degree of freedom for every part of the basis optimization.

When, after restarting the factorization with a refined basis, the result is still unsatisfactory with respect to the mapping error, the researcher can investigate whether the number of basis vectors is ill-set for the data's factorization by *adding or removing basis vectors*. The complete *randomization* of basis vector coefficients can also provide the necessary means for overcoming local optima. Often, this scenario can be observed while the online visualization is running. Therefore, the optimization can be *stopped or resumed* at any time.

1 For selected points p, the opacity of each basis b is adjusted to α_b, where $||C_{p,\bullet}||_1$ refers to the sum of the p'th row of the coefficent matrix C.

2 Note that the mapping errors cannot directly relate to any selection of basis vectors. Highlighting parts of the errors might lead to the semantically incorrect conclusion that the selected basis vectors are accountable for these errors.

3 For basis vector b selected, the opacitiy of each spectrum (m-D point) p is adjusted to α_p, where $||C_{p,\bullet}||_1$ as above.

4.3.6 Results

Collaborating researchers from the domain of air quality research have applied our method to the factorization of biomass combustion sources. Here, we give excerpts of the study and evaluation of our visualization framework with respect to speed, accuracy, and ease of use. We also give a glimpse into the insights generated; however, these findings will be published in a different forum. In summary of what is presented here, our collaborators have been able to (i) reproduce established findings in mere a fraction of the time than it was possible before, (ii) process and analyze ten-times more spectra than in previous studies, and (iii) gain surprising insights enabled by the visualization.

4.3.7 Factorization of biomass combustion sources

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. Among the largest inadequacies in quantifying emission factors of biomass combustion is the general paucity of methods identifying and quantifying particle classes in ambient measurements for a wide range of ecosystems and combustion conditions, including anything from naturally occurring, large wildfires to woodstoves and fireplaces used for residential heating [RKEE05]. This is largely a result of the physical and chemical complexity of particulate matter (PM). PM is the least understood factor in almost all issues ranging from human health to global climate change and provides the impetus for studying ambient particles in increasing detail to close these knowledge gaps.

In prior work, high-resolution clustering algorithms were used to characterize particle classes of biomass combustion emission factors [BZWJ05]. The particle class depictions in Figure 4.6 are the averages of all single particle mass spectra within their class and the listed percentages represent the fraction of the total detected particles belonging to that class. The relevant carbon cluster ions C_x^+ (typically attributed to elemental carbon (EC)), hydrocarbon fragment ions $C_xH_y^+$ (organic carbon (OC)) and isotopic K^+ ions are labeled. However, the apparent distinction between these particle classes, or compositional discretization, is somewhat arbitrary and largely a result of the parameters chosen to control the data clustering algorithm. In reality, there is a "continuous distribution" of particle compositions ranging from those with mass spectra dominated by K (Figure 4.6.a) to purely carbonaceous aerosol with mass spectra dominated by EC and OC ions (Figure 4.6.d). The fundamental issue in correctly distinguishing these particle types is isobaric interference at m/z 39. Values in m/z 39 can represent K^+ ions, $C_3H_3^+$ ions or some mixture of the two. $C_3H_3^+$ ions are fairly ubiquitous in the mass spectra of carbonaceous aerosol, and thus our ability to accurately separate biomass combustion from other sources of EC/OC particles in ambient mixtures resides almost exclusively in our ability to accurately determine the relative contribution of these two ions to the signal intensity observed at m/z 39. The progression of particle compositions shown in Figure 4.6 was designed in attempts to capture this issue. The presence of K is unambiguous in Figure 4.6.a, and even Figure 4.6.b. As the ion signal at m/z 39 decreases and becomes comparable to, and eventually less than, the C_3^+ signal (in Figures 4.6.c and 4.6.d), it is increasingly difficult to unambiguously specify the presence of K in the particle. This is a predominant issue as

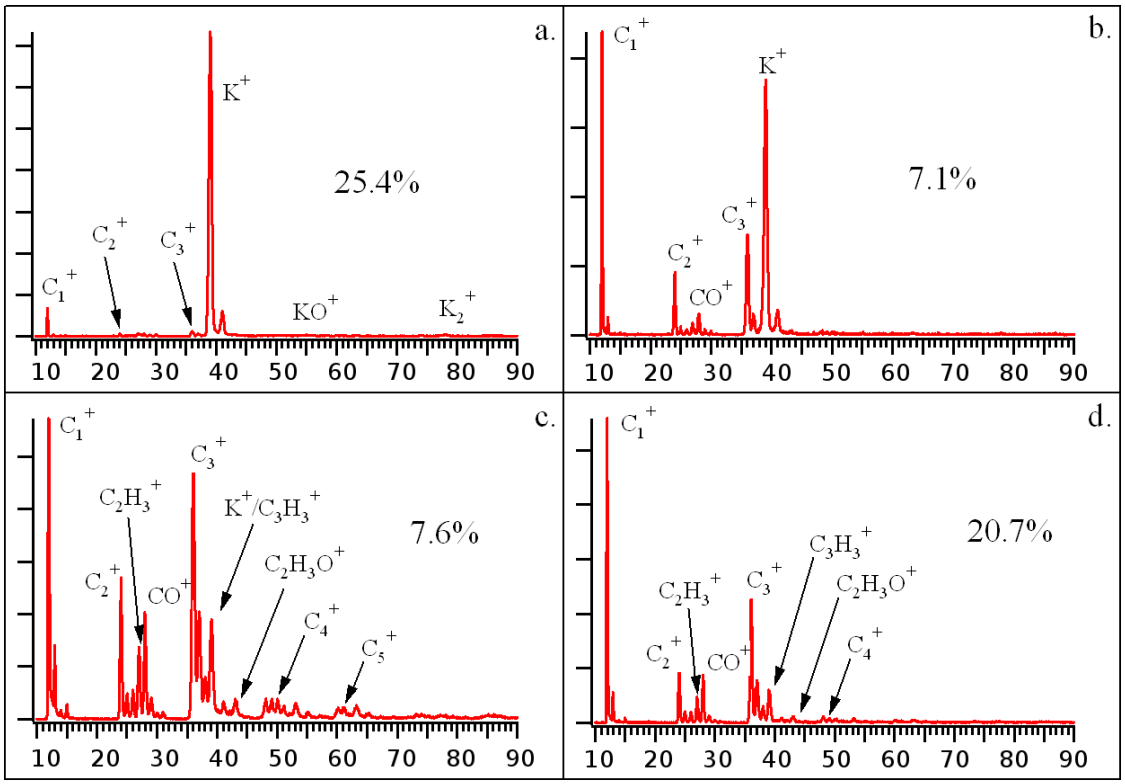


Figure 4.6: Single particle mass spectral representations of particle types observed from biomass combustion in a wood stove during source sampling experiments. Due to isobaric interference at m/z 39 (K+ / C3H3+), a clear separation could not be achieved.

Figure 4.6.d is characteristic of what is generally observed for carbonaceous particles from a variety of sources. As a result, a significant amount of manual effort is expended making the appropriate peak assignments necessary to distinguish between purely EC/OC and biomass combustion particles. Unfortunately, this cannot be done using the cluster-average mass spectra, such as those shown in Figure 4.6, since the relevant clusters commonly contain a mixture of both particle types and clustering obscures important details in the spectra that assist the analyst when making the assignments. Instead, the individual mass spectra comprising the clusters must be inspected, interpreted and classified manually to obtain more accurate source separations. This is extremely time-consuming and ultimately based on the subjective interpretation of an expert analyst without a fundamental mathematical underpinning.

In the following, we discuss the application of our method to (1) the woodstove source sampling data from Pittsburgh, Pennsylvania, discussed above and (2) ambient Rapid Single-ultrafine-particle Mass Spectrometer (RSMS) data collected during a sampling campaign in Fresno, California [BZW09].
We evaluate this method based upon (i) accuracy in resolving K-containing particles from

EC/OC particles, (ii) efficiency in reducing overall analysis time and (iii) the contribution of the visualization and interactive elements in conducting, interpreting, and gleaning new insight from the overall process. The woodstove source sampling experiments offer a good basis to evaluate performance since the data are already well-characterized while the Fresno campaign provides a relatively unexplored and complex ambient data set to test the visualization and interactive elements. The Fresno data is particularly well-suited for this study since the two largest sources of particle pollution in the area are vehicular emissions from local traffic and biomass combustion emissions from residential heating and agricultural burning. As a result, the composition of the air shed is a large external mixture of internally mixed EC-, OC- and K-containing particles, as well as components formed in the atmosphere by gas-phase photochemistry and condensed on the particles, providing a very challenging environment for resolving sources.

Results from the interactive analysis of the RSMS data collected during the wood stove source sampling experiments are shown in Figure 4.7. Only those dimensions, or m/z values, relevant to resolving the ambiguity in the presence of particulate K were included in the analysis: m/z 24 (C_2^+), 27 ($C_2H_3^+$), 36 (C_3^+), 39 ($^{39}K^+$ / $C_3H_3^+$), and 41 ($^{41}K^+$ / $C_3H_5^+$). This ability to select dimensions of interest, rather than analyzing the full m/z range of the data, is a strong feature of the visualization interface and reduces both computational burden and analysis time tremendously. The top panel of Figure 4.7 is a screen shot of the visualization interface showing all of the data (fine lines) and basis vectors (bold lines) identified during the factorization. It is immediately clear in this figure that the algorithm does an excellent job factoring out the K^+ and $C_3^+/C_3H_3^+$ signals and that these two basis vectors model the data well. Also apparent is the fact that the solution is slightly over-determined and the elements of the C_2^+ basis vector could have been incorporated into the C_3^+ basis vector without any loss of information. The major advantage of the visualization interface is that this can be done interactively by removing the C_2^+ basis, adjusting the C_3^+ basis vector lines to match the C_2^+ and $C_2H_3^+$ signal evident in the data and then performing the optimization again. This is a very useful and efficient way of analyzing these data. An unexpected and highly interesting result is the apparent irrelevance of the m/z 41 dimension. Previous efforts have been made to separate biomass combustion and EC/OC particles based on the ratio of integrated ion signal at m/z 39 to 41, and thus its inclusion in the analysis, but with limited success. The underlying assumption is that $C_3^+/C_3H_3^+$ ratios are small compared to the larger values associated with the natural isotopic abundances of ^{39}K and ^{41}K. Further research shall be conducted in this regard.

The bottom left panel of Figure 4.7 shows the projection of all data points onto the K^+ and C_3^+ basis vector space. Again, the separation in the data is strikingly clear with K dominant particles clustered in the upper left-hand corner and EC/OC dominant particles in the bottom right; these areas are circled in the figure. This interactive visualization framework is very resourceful and the ability to interact with the projection by highlighting individual data points, or clusters of points, and inspecting the relative basis contributions to the selected points, is invaluable to interpreting the results and understanding the structure of the data.

To quantitatively separate K-containing particles from purely EC/OC particles, the

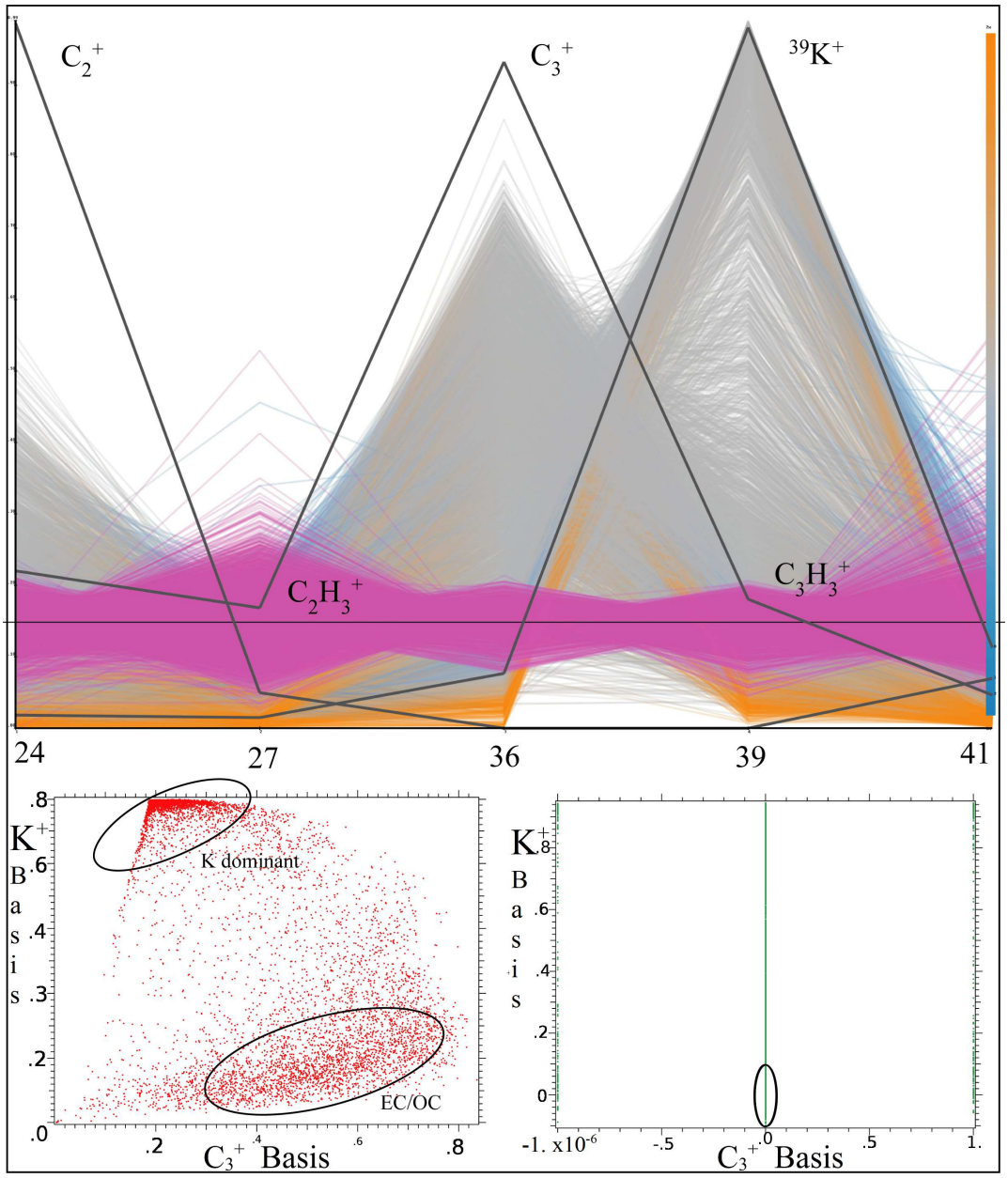


Figure 4.7: Top: Screen shot of the visualization interface showing results from factorization of the RSMS data collected during the PAQS woodstove source sampling experiments. Bottom left: Projection of the RSMS data onto the K^+ and C_3^+ basis space identified during factor analysis. Bottom right: Projection of the RSMS data onto the K^+ and C_3^+ basis space after factoring out the C_3^+ basis vector from the data; As can be seen in the top figure, the focus for dimension reduction lies with information in C_2^+, C_3^+, and K^+, and not in $C_2H_3^+$ or $C_3H_5^+$. The factorization minimizes the error in mapping K^+ and C_3^+ signals which corresponds nicely to existing research stating that K^+ is the major classifier in these dimensions. Further, the spectra of K^+ and $C_3^+/C_3H_3^+$ particle classes are separated automatically and accurate. In prior work, atmospheric scientists have required months to obtain this result.

C_3^+ basis vector was factored out of all data points and the resulting data re-projected onto the K^+ and C_3^+ basis vector space, as shown in the bottom right panel of Figure 4.7. The idea is that those data points showing near-zero contribution from the K^+ basis vector are purely EC/OC particles while the above-zero points are K-containing particles. Clearly, there is error in the overall fit of the basis vectors to the data, as evidenced by points with negative K^+ contribution, and this must be incorporated into the analysis. Using the average residuals between the data points and basis vectors at m/z 39 shown as bright pink lines in the top panel of Figure 4.7 as an error estimate for the factorization, a window centered about zero has been drawn in the projection designed to separate purely EC/OC particles from K-containing particles; note that the window fully encompasses those points with negative contributions. Summing all points within the window yields a value of 0.29 for the fraction of the total number of particles sampled that are purely EC/OC particles. This result is in nearly perfect agreement with the value of 0.32 obtained during the manual analysis of all 7000 mass spectra [BZWJ05].

Figure 4.8 shows results from the interactive analysis of the RSMS data collected during the field campaign in Fresno. The same five dimensions were used and over 70,000 single particle mass spectra were analyzed. Quite notably, this would not have been possible by manual analysis. The optimization was initiated with three basis vectors but interactively reconfigured to only two during analysis. The projection of all data onto the K^+ and C_3^+ basis vector space is shown in the lower left panel and the clustering of the data points is very distinct and clear. A majority of the data appears to fall roughly along a positively sloping line (circled in the figure), where the upper cluster represents K-dominant particles and the K^+ contribution decreases down the line toward the lower cluster. A snap shot of the visualization interface when the lower cluster of data points is highlighted is shown in the top panel. The apparent weak contribution of both the K^+ and C_3^+ basis vectors to this cluster is actually due to the prevalence of NO^+ ions (m/z 30) in these mass spectra. The framework does an excellent job resolving this particle class, especially since m/z 30 was not included in the analysis. Purely EC/OC particles cluster in the lower right-hand corner of the projection but are very sparse relative to the other particle types, and even the results of the woodstove source sampling. This is also a highly interesting and informative result but will not be addressed any further here. Similar to the analysis above, the C_3^+ basis vector was factored out of all data points and the resulting data re-projected onto the K^+ and C_3^+ basis vector space (lower right panel). Again, an error estimate based on residuals was used to create a threshold range about zero to calculate a value of 0.06 for the fraction of the total number of particles sampled that are purely EC/OC particles. A notable strength of this particular interactive exercise was the ability to robustly differentiate the presence of K^+ versus $C_3H_3^+$ even in spectra where NO^+ dominates the ion signal.

4.3.8 Expert evaluation

When processing high-dimensional data, paying attention to all the dimensions is challenging, so it is crucial to provide the user with a mechanism for appropriately reducing the dimension of the problem at a minimum loss of information, as well as showing the user both which dimensions are important and where information is lost. In the example

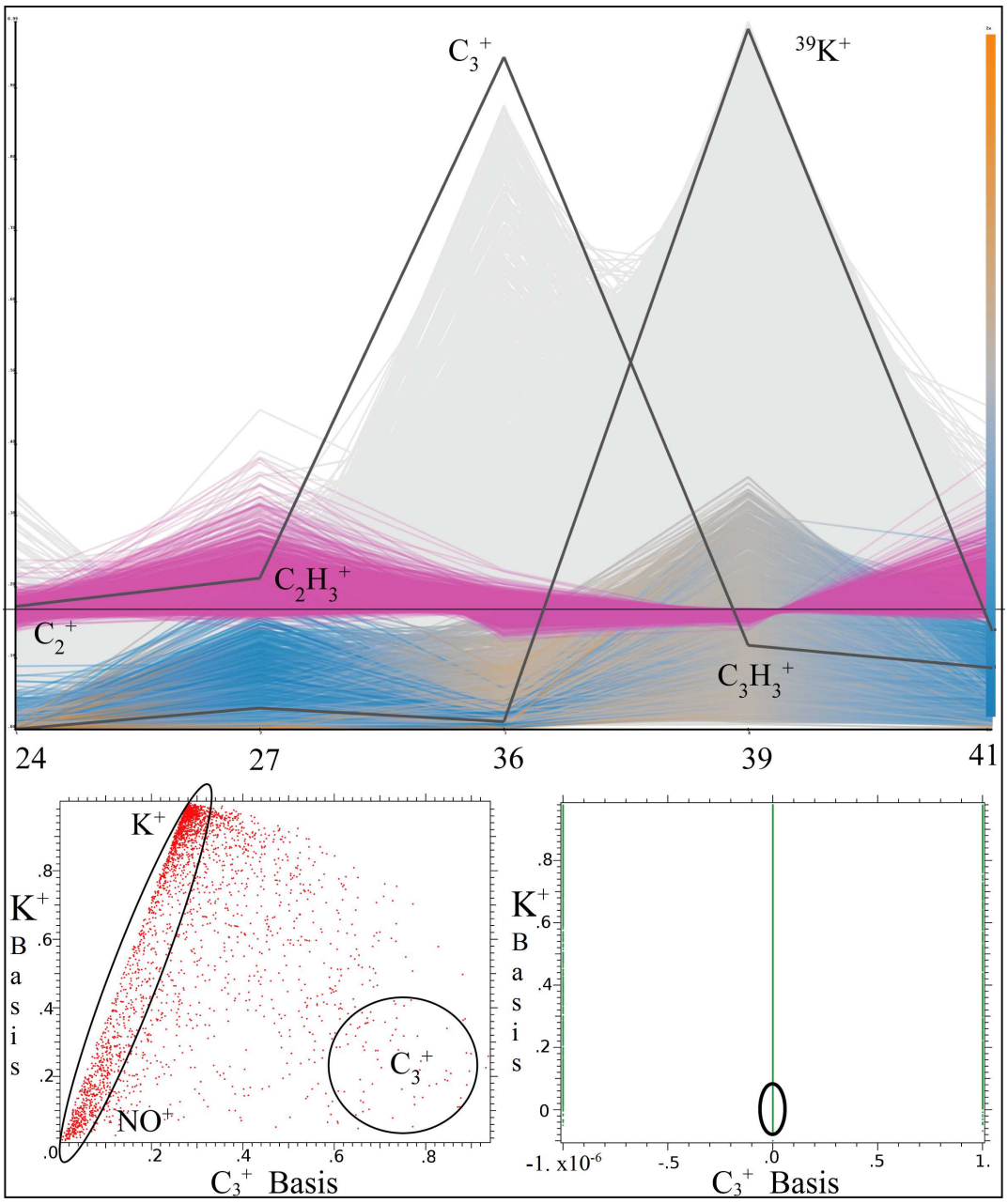


Figure 4.8: Top: Screen shot of the visualization interface showing results from factorization of the RSMS data collected during the field campaign in Fresno, CA. Bottom left: Projection of the RSMS data onto the K^+ and C_3^+ basis space identified during factor analysis. Bottom right: Projection of the RSMS data onto the K^+ and C_3^+ basis space after factoring out the C_3^+ basis vector from the data; While K^+ and C_3^+ are captured and separated well in the factorization, the projection by the basis suggests a dominant cluster in the data that is not part of these two particle classes. Further analysis shows that this is due to the prevalence of NO^+ ions (m/z 30) in these mass spectra.

here, subtle differences between mass spectra can provide crucial guidance for assessing and quantifying the source contributions to a given air shed. Clustering algorithms may obscure these subtleties reducing their usefulness. At some level, the involvement of an expert analyst is unavoidable but the burden of manually inspecting the several hundred thousand mass spectra acquired during typical field campaigns, or even a subset thereof, is unreasonable, time-consuming, and largely subjective. A highly visual and interactive computational platform for analyzing, characterizing and manipulating these data is essential to these efforts. The emerging data visualization community requires interdisciplinary collaborations to develop effective platforms – the impetus for the current work.

The interactive visual framework developed here provides educated, mathematically rigorous suggestions, while leaving full control and physical verification to the analyst. In this regard, both visualization and interaction build trust in both the factorization method and its implementation. These computations may identify large errors, which must be conveyed to the analyst. While information loss may be unavoidable at some level, it is crucial to visualize exactly where this loss occurs. Both the color coding and the error lines are helpful in qualitatively interpreting the basis transformations related to potential errors. Filtering spectra and basis vectors is intuitive and by merely performing a few interactions, the dominant clusters of spectra are easily highlighted. This astonishing level of visual interaction with mass spectrometry data has not been possible before, thereby introducing both new and exciting research potentials.

While previous methods did not recognize important features in the spectra, the factorization is much more facile at identifying important commonalities and subtleties in the spectra. We were able to reproduce established findings from the woodstove biomass combustion measurements in a matter of hours, where prior work took months. Also we were able to gain new insights from data collected in Fresno that will be the focus of future investigations.

4.3.9 Conclusions

This section describes a framework for dimension reduction of single particle mass spectrometry data that entails the use of visualization and interaction in order to steer computations. Our work contributes to the community of air quality research by providing novel means to reduce data dimensionality by physically correct and unambiguous basis transformation. Thereby, we overcome limitations of previous methods that use dimension reduction as a black-box scheme and move toward more physically sensible computing. By the utilization of specifically tailored regularization terms, the presented non-negative matrix factorization is capable of resolving ambiguity of the mass spectrum, thereby, laying the groundwork for more reliable data analysis in air quality research. As the expert evaluation shows, analysts are able to reproduce known research results with great ease and speed by using our method. To the visualization community, we contribute well-justified visual encodings, as well as novel interaction techniques that aid in the visual analysis and verification of non-negative matrix factorization. Going further in this direction, the following section describes how approximation errors in non-negative matrix factorization can be analyzed and refined.

4.4 Analysis of approximation errors in non-negative matrix factorization

The previous section describes how to characterize air parcels in terms of abundance of characteristic sources: a basis transformation is to be found that models the observed spectra as a linear combination of the spectra arising from each of the chemical compounds, such that linear combinations of this basis forms the observed mass spectra. Physical and chemical constraints further dictate both basis and coefficients to be non-negative, rendering spectral decompositions inapplicable and the optimization problem non-convex. Thus, approximations can only be computed iteratively by gradient-based approaches such as non-negative matrix factorization (NMF) and currently known algorithms can only produce sub-optimal results. The approximation error, defined as the discrepancy between data and its lower-dimensional approximation, of mere locally optimal approximations deviates significantly from that of a globally optimal (correct) solution. As a consequence, the quality of approximations needs to be assessed by the scientist. The visual communication of errors in non-negative matrix factorization, however, has not yet been studied in visualization research and common visualization tools are not applicable to this problem.

In this section, we introduce a new methodology to the visual analysis of approximation errors in non-negative matrix factorization that includes (i) an approach to assess the quality of a factorization basis, (ii) a visualization of factorization errors, and (iii) means to interactively minimize specific errors. During analysis, the scientist can compare the numerical benefit in introducing a basis vector that minimizes error features selected in the visualization against the benefit of each vector currently in the basis. Following this methodology, the scientist can improve the factorization quality and consequently overcome "being stuck" in local minima of non-convex optimization. Due to the high degree of interactivity in this analysis, our method also provides awareness about the information loss associated with the dimension reduction process and allows for an educated decision regarding the degree of freedom needed to approximate high-dimensional data. Thereby, we contribute to applications including, but not limited to, air quality research, by providing novel means to elucidate chemical species content of MS spectra from any wide range of sources. The core of our methodology can further be generalized and applied to other settings of non-convex linear approximation of high-dimensional data.

The remainder of the section is structured as follows. Section 4.4.1 provides the necessary requirement analysis and task description for our effort. Section 4.4.2 describes our method, entailing a description of our methodology, the projection of factorization errors, our approach to interactive analysis and refinement of the factorization, as well as implementation remarks. Section 4.4.3 demonstrates how this method is effectively applied in the factorization of SPMS data and evaluated with respect to its ability to produce new insights to the application of air quality research. Finally, concluding remarks are given in Section 4.4.4. The work is published in [EHH*13].

4.4.1 Requirement analysis

In the following, a brief account of the problem definition is given that involves a description of errors in SPMS factorization, as well as our terminology. Finally, we describe the tasks and requirements arising from this problem for the application of air quality research.

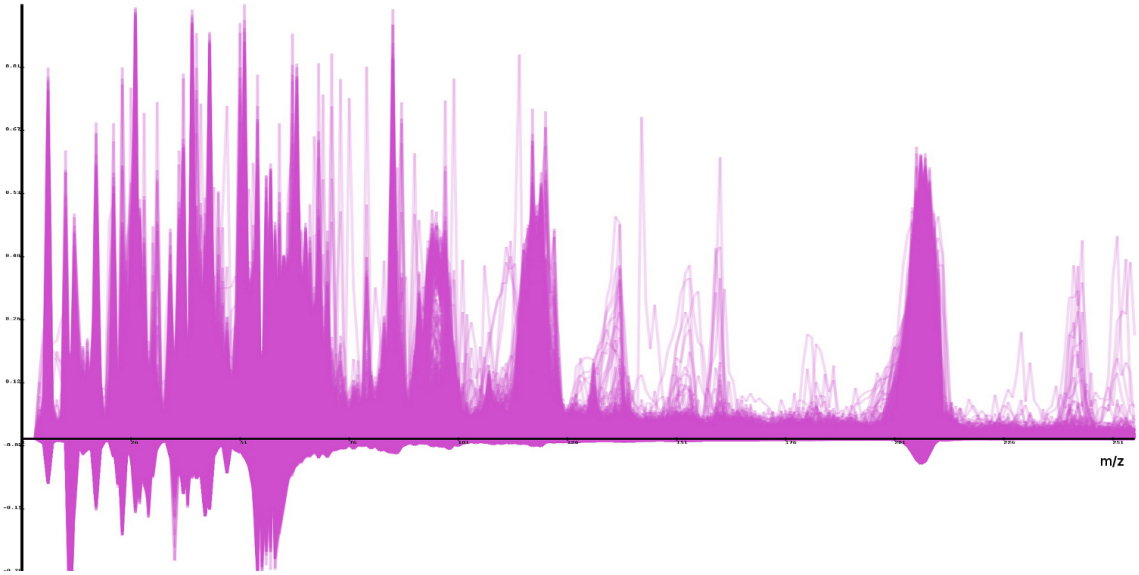


Figure 4.9: Previous work visualizes the errors produced by SPMS data factorization in high detail. Due to data complexity and dimensionality, this representation is prone to visual clutter and fails to provide an overview to analysts who are faced with the problem of identifying, classifying, and analyzing error features.

Errors in SPMS data factorization

Several errors are involved in the various stages prior to SPMS data analysis including, but not limited to, data acquisition, sensor measurements, bit noise, integration of the mass spectra, dimension reduction, gradient descent, and visual mapping. While many of these errors are marginal or cannot be determined, the errors introduced by dimension reduction can be both considerably large and determined based on the original data as ground truth. Given the complexity of high-dimensional SPMS data (that is almost of complete rank), any approximation to lower-dimensional space produces errors. However, given the non-convex nature of our optimization, for which globally optimal results cannot be expected, analyzing these errors becomes a necessity.

Consider a factorization for n data points of dimension m, $X \in \mathbb{R}_+^{(n \times m)}$, in coefficients $C \in \mathbb{R}_+^{(n \times k)}$ and basis $B \in \mathbb{R}_+^{(k \times m)}$ for $k \ll m$. For the purpose of this work, we define *the error of a factorization* as the discrepancy between the original data and its factorization: $X - CB \in \mathbb{R}^{(n \times m)}$. Hence, errors are high-dimensional residuals, given by the misfit for each point in the data. We impose no restrictions on the errors, as they may be both positive or negative and of arbitrary magnitude, as depicted in Figure 4.9.

In addition to the errors introduced by dimension reduction, a SPMS factorization largely exhibits noise that is assumed to follow a Gaussian distribution (for example, due to gradient descent optimization and sensory noise). For the analysis of a suitable factorization basis, these error contributions are of relatively low interest to analysts, as they are both unavoidable and practically independent of the factorization basis. In contrast, specific error features that are of interest are those that significantly deviate from a Gaussian

distribution. If these specific error features occur in abundance, it indicates that the factorization basis does not allow the depiction of these features in the data. This may be either due to the dimensionality of the basis being set too low, or due to a sub-optimal factorization basis that does not cover significant parts of the data.

In this section, we make use of terms as significance and optimality. We resort to this terminology with respect to the quantity of information (variance), as the quality of information cannot be assessed numerically. As such, we define the overall error of a factorization by a norm of its errors ($||X - CB||$) and define a factorization to be optimal that produces a minimal overall error. However, at no point during analysis do we dismiss any solution due to numerical inefficiency. To determine what may serve as adequate to the current purpose of analysis is left to the analyst.

Tasks and requirements

In order to determine an adequate factorization of SPMS data, atmospheric scientists have the ultimate task to minimize both dimensionality and error of the approximation. Thereby, the goal is to choose a trade-off between dimensionality and error, admitting identified errors that have been minimized and are unavoidable due to dimension reduction for the sake of having a lower-dimensional representation. However, for mere locally optimal solutions, it is unclear whether errors are truly minimized and unavoidable. Therefore, a methodology is needed to assess both (i) the error and (ii) the quality of a factorization. While the overall approximation error can be computed as described in the previous section, the quality of a factorization relates to the efficiency of a basis in approximating the data. Basis efficiency quantifies how much information from the data is represented in the factorization in relation to a globally optimal solution given the same degree of freedom. As knowledge of a globally optimal solution is unknown in general, analyzing factorization quality requires a human-in-the-loop approach and the tools to aid in visual analysis.

It is only by the conveyance of both properties (error and efficiency) that scientists can determine the "right" dimensionality for the basis and an adequate approximation of the data. Analysis to ascertain basis efficiency must be tightly coupled with the visualization of error features (and their significance) to aid the scientist in determining an admissible trade-off, deciding which errors to admit as a consequence of dimension reduction and weighting errors against dimensionality of the approximation. Finally, this methodology to error-based analysis should include the means to systematically refine factorizations towards minimizing errors. In summary, the key tasks and requirements for the visual analysis of errors in SPMS data factorizations are:

1. **Visualizing error features:**
   Error visualization should convey a classification of errors by importance and type, and serve as a basis to conduct detailed analysis. A major requirement is the visual separation of noise from specific error features described in the previous section. The visualization must convey how much of the data is factorized with (less significant) small errors following a normal distribution over all dimensions, as opposed to how much of the data is not well represented, producing errors of (significant) specific features, as described in the previous section.

2. **Analyzing basis efficiency:**
   In assessing the quality of a factorization, it is important to understand where errors originate from, as they may stem from either (i) due to shortcomings of the optimization process (local minima) or (ii) due to a necessity in dimension reduction defined by basis dimensionality. Visualization should help to answer this question and, when possible, uncover inefficiencies of the factorization basis with respect to approximating the data.
3. **Refining factorizations:**
   Once errors are identified during the analysis that are unacceptable, an analytical system should entail the refinement of the factorization towards eliminating these errors. A key requirement is interactivity of the data factorization and providing visual feedback concerning the benefit of adjustments.

Our method aims at satisfying these requirements.

4.4.2 Method

We describe how factorization quality can be analyzed, sub-optimality assessed, and the factorization be improved. Essential to our approach is a highly visual and analytical framework that involves the analyst in several key steps.

Assessing optimality

Visualizing optimality of a factorization is a challenging task, as there exists no method that can spot local minima or quantify their (sub-)optimality effectively. However, considering the following concept leads to the conclusion that local minima in non-negative matrix factorization can in fact be revealed with the help of visualization and interaction.

An optimal data basis must consist of basis vectors that are all optimal. Consequently, the exchange of one vector in the basis set must not produce a (numerically) better approximation. Further, for a sub-optimal basis must hold that there are better basis vectors that produce less overall error, with respect to lowering error magnitudes in their abundance. Conversely, the presence of similar errors of high magnitude and abundance directly corresponds to a basis vector candidate that is not part of the current basis, while being numerically beneficial to be included. This leads one to conclude that optimality of the basis can be assessed by comparing the amount of information currently conveyed by each basis vector against the amount of information that could be conveyed by basis vector candidates. Further, these candidates are directly reflected by and can be identified based on similar errors of high magnitude and abundance. Consequently, local minima in the factorization can be revealed by (i) identifying candidates based on visually assessing error magnitudes, similarity, and abundance, and (ii) comparing the amount of information conveyed by the current basis vectors in relation to that of the candidate. If the candidate allows for the conveyance of more information than one of the current basis vectors, then the basis is sub-optimal. In this case, the candidate can be introduced into the basis, possibly by exchanging one of the other basis vectors of lower benefit.This concept requires three aspects: (i) a comparative measure of conveyed information per basis vector, (ii) an error visualization focusing on error magnitudes, similarity, and abundance, and (iii)

interactive probing of the error to visually compare the benefit of each basis vector against that of selected candidates. We introduce this measure in the following.

In NMF, coefficients are exclusively non-negative. Consequently, each basis vector $b_i \in \mathbb{R}^m$ only adds to the total approximation of $X \in \mathbb{R}^{(n \times m)}$ according to its coefficients $c_i \in \mathbb{R}^n$ and does not delimit other basis vectors' contributions. Thus, the overall approximation is decomposed into each basis vector's contribution, such that $CB = \sum_{1 \leq i \leq k} c_i \otimes b_i$. For each basis vector's contribution, we quantify its "gain" by a norm of the residuals to X. It is possible that such contributions cover more variance than present in X. Therefore, the gain of b_i must be based on how its contribution matches the data. Consequently, our gain measure is defined as follows:

$$gain(b_i) = ||X||_1 - ||X - c_i \otimes b_i||_1 \ . \tag{4.9}$$

Analysis of this measure facilitates insight into the amount of information each basis vector introduces in the factorization. Thus, spotting a local optimum reduces to identifying the basis vectors of small gain and comparing them to the gain of the basis vector candidate that corresponds to the largest error cluster.

Projection of factorization errors

In the following, we describe the design of a visualization that focuses on providing an overview of factorization errors, while highlighting error classes for identifying possible basis vector candidates. Thereby, we rely on two major classifiers for factorization errors: **magnitude** and **irregularity**. *Error magnitudes* classify error severeness per data point by a norm. While different norms may be suitable for this task depending on the application, we apply the Euclidean norm to quantify the error magnitudes of SPMS factorization, since it emphasizes larger misfits over smaller ones. Additionally, we classify errors by a measure of *irregularity*, similar to Hoyer's sparsity measure [HR09], that is orthogonal to error magnitudes and suggests a misfit in the factorization basis, as opposed to inadequate numerical computations. With $e \in \mathbb{R}^m$ referring to one of n errors, each consisting of m residuals, this measure of error irregularity is defined as follows:

$$\begin{aligned} \alpha(e) &= 1 - \frac{cos\angle(abs(e), \mathbf{1}) - \frac{1}{\sqrt{m}}}{1 - \frac{1}{\sqrt{m}}}, \text{ where} \\ cos\angle(abs(e), \mathbf{1}) &= ||e||_1 / (||e||_2 \sqrt{m}) \ . \end{aligned} \tag{4.10}$$

The dominance of a (sparse) feature in the error is defined by the cosine of the angle between its absolute and $\mathbf{1} \in \mathbb{R}^m$, the vector of ones in all coordinates. Independent of the error's magnitude, it holds a measure of irregularity for $0 \leq \alpha(e) \leq 1$, where an error of equal absolute coordinates leads to a value of 0 and a unit vector to 1. Figure 4.10 illustrates how this measure is interpreted to SPMS factorization errors. Based on this measure, our projection ϕ, depicting error magnitude and irregularity, is defined as follows:

$$\begin{aligned} \phi : \ \mathbb{R}^m &\to \mathbb{R}^2 \\ e &\mapsto (\alpha(e), \ ||e||_2) \end{aligned} \tag{4.11}$$

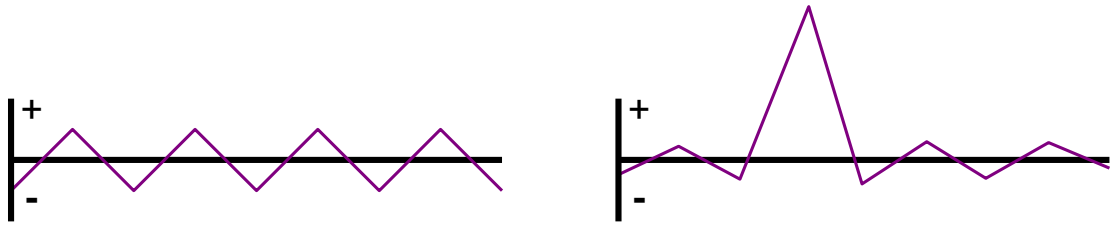


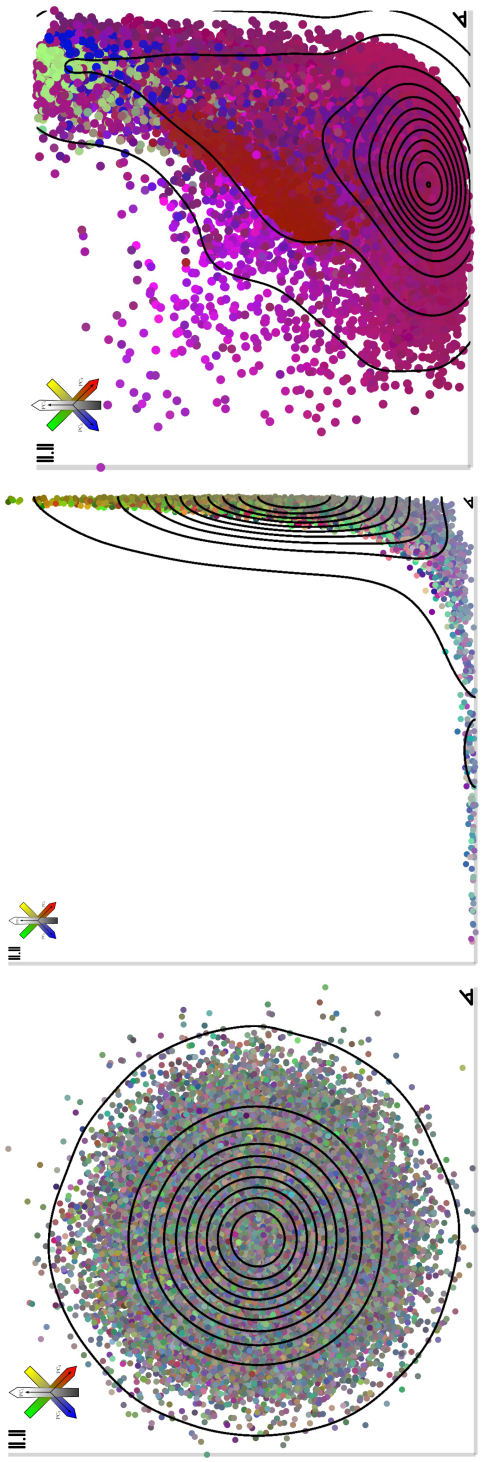
Figure 4.10: By utilizing a measure of error regularity (left: regular $\mapsto$ 0, right: irregular $\mapsto$ 1), the presence of dominant features in errors can be quantified, allowing for a visual assessment of noise level.

The y-axis of this projection maps the magnitudes of the factorization errors, while the x-axis maps to $\alpha(e)$, which indicates the dominance of a specific feature in the residuals of an error, as opposed to showing uniform residuals.

By this mapping, errors of the same magnitude and regularity are mapped to the same locations, regardless of their coordinates being identical. This problem is inherent to dimension reduction and impossible to overcome. However, it can be at least partially alleviated by a color scheme that shows additional differences. We define a projection to three dimensions that assigns color values to each error according to its spatial configuration in $\mathbb{R}^m$. First, the errors $X-CB$ are normalized to unit scale in order to render the projection independently of error magnitudes and centered in order to utilize the full color range. Second, the covariance matrix is built from this normalized centered matrix. Finally, the eigenvectors associated with the three largest eigenvalues of this covariance matrix define the projection into color space. A suitable color space is, for example, CIE*lab*, as it is uniform and of orthogonal basis.

For effective error investigation, the abundance of errors within ranges of specific magnitude and irregularity must be accounted for in the visualization. In order to convey information about the quantity of errors belonging to the same classifiers, the visualization must make aware of the concentration of points within regions of the projection. However, given limited resolution, the specific concentration of points in a projection is visually impossible to assess for large data sets. Although interactively zooming into a projection can unclutter the point configuration, this does not provide quantitative insight into the point concentrations within a region. While assigning opacity values to points, either by the use of alpha blending or by application of a non-linear transfer function, can help convey point density, this approach does not scale well with increasing number of data points.

In order to convey point concentrations within the projection, we use density field contouring. Thereby, a high-resolution 2D scalar field is computed that holds, for each pixel, the number of points projected to this location. Subsequently, the field is processed via a convolution step using a Gaussian filter kernel, which is scaled to have a peak height of 1 that decreases to 0 over its bandwidth. The Gaussian filter smoothes the field and accumulates density values in the locality of its bandwidth, producing a density field. A texture of contours can be computed, for example, by thresholding for isovalues in



(a) (b) (c)

Figure 4.11: An overview of factorization errors is achieved by projecting errors based on magnitude (vertical axis) and irregularity (horizontal axis). Further classification of error types is provided by color (similarity) and density contours (abundance).

the density field. Contours of equal width in image space can be realized by setting the threshold dependent on the local gradient of the density field. For further information on kernel density estimation, we refer to [WJ95].

To summarize the properties of the error visualization defined above, we list the main features in the following:

- **Horizontal axis:** irregularity of errors (feature dominance)
- **Vertical axis:** magnitude of errors (Euclidean norm)
- **Color:** similarity of errors (in $\mathbb{R}^m$)
- **Contours:** local quantity of errors (point density)

Figure 4.11 shows examples for different data factorizations.

Interaction

The selection of errors in a specific magnitude-distribution range (regional selection) and/or (sub-)selection of errors based on their spatial relationship (color selection) in this visualization can be linked and act as a filtering mechanism for different high-detail views. Further sub-selection in high-detail views can effectively identify error features of a factorization. These features correspond to a potential basis vector candidate that eliminates these errors. In the following, we describe how the factorization quality can be assessed based on these candidates and how the basis can be interactively refined.

Interactive refinement

After the selection of errors that are of interest in the interactive analysis process, our methodology entails the visualization of the potential gain produced by the addition of the corresponding basis vector candidate. This candidate, the optimal basis vector that eliminates the selected errors, is given by the mean of the data points producing it, weighted by the absolute mean of the errors per coordinate. As such, the basis vector is introduced that has the exact features of the data points that are not covered by the factorization. The coefficient matrix is adjusted projecting all data points onto the candidate vector and adjusting coefficients of the other basis vectors in relation to how the candidate allows for a better representation, while the coefficients for the candidate vector are generated conversely based on the best fit.

Using this adjusted starting configuration, our NMF model is performed for several iterations to produce an adequate estimate of the factorization quality that is achievable by including the candidate. Due to the linear nature of the approximation, the gain that can be expected by the addition of a vector to the basis depends on the magnitude and abundance of spectra covered by the vector minus the variance between the spectra. Consequently, the gain is highest for introduction of a basis vector corresponding to an abundance of large errors showing similar features. However, by defining a basis candidate, the analyst does not restrict the basis. While the two-block optimization scheme will first optimize the coefficients to the initial basis, consecutive iterations will also update the basis vectors if they are not optimal. Without delimiting the optimization, this methodology can be used to overcome local minima, as well as to analyze and refine the basis.

Subsequently to the NMF optimization with adjusted basis, the gain of the basis prior to adjustment is visualized in relation to the gain post adjustment in the chart, while the differences are highlighted in distinct colors. Figure 4.13 shows an example of this separate view in our framework. If the gain of the analyst's candidate is larger than that of a basis vector from the previous basis configuration, then this candidate contributes more information to the approximation and, consequently, a local minimum in the computation has been uncovered. On the other hand, equally high gain values for all basis vectors, in spite of high errors, suggest that the degree of freedom is set too low for the basis. By selection in the bar chart, the analyst can flag any basis vector to be added or removed from the basis and subsequently trigger the optimization to be performed again for the desired configuration. Thus, the basis vector that minimizes the selected errors can be added to the basis, other basis vectors of low gain can be deleted, or the candidate can be forfeited in order to continue probing of the errors. As interactivity is an integral part of this analysis, performing optimization methods on the GPU is inevitable for large data sets. We describe our implementation in the following.

Independence regulation on the GPU

In [WR10], Wilson et al. described a term for regulating mutual independence between the coefficients of basis vectors in non-negative mixtures. Although being very robust, their formulation requires no matrix inversion, making it more flexible than previous approaches and fast to compute on the CPU. The update of the coefficient matrix C, applicable to multiplicative NMF update schemes, that regulates independence is based on the derivative of a cost function J_C measuring correlation coefficients, as described in (4.3.3).

We note that the formulation given in [WR10] of the partial derivative $\partial J(C)/\partial C_{a,b}$, is not easily realized on a GPU and can be reformulated more efficiently. By exploiting the fact that the partial derivatives of the correlation matrix terms are symmetric and non-zero only in a single row and column, we can greatly simplify the formulation as follows:

$$\frac{\partial J(C)}{\partial C_{a,b}} = 4 \left|\left| Corr_{b,\bullet} \otimes \frac{(\mathbf{n}_c \mathbf{n}_c^T)_{b,\bullet} \otimes C_{a,\bullet} - \frac{C_{a,b}}{\mathbf{n}_{c_b}} \mathbf{n}_c \otimes (C^T C)_{b,\bullet}}{\mathbf{n}_c \mathbf{n}_c^{T^2} + \varepsilon} \right|\right|_1 \tag{4.12}$$

Here, $\otimes$ denotes the element-wise multiplication between two matrices of the same dimensions, analogously to the division of $\mathbf{n}_c \mathbf{n}_c^{T^2}$ which is understood as element-wise division of the element-wise squared outer product matrix of $\mathbf{n}_c$. The correlation matrix $Corr$ and norm vector $\mathbf{n}_c$ are given by

$$\begin{aligned} Corr &= N_C C^T C N_C \ , \\ N_C &= diag(\mathbf{n}_c^{-1}) \ , \text{ and} \\ \mathbf{n}_c &= (||C_{\bullet,1}||_F, ..., ||C_{\bullet,k}||_F) \ . \end{aligned} \tag{4.13}$$

Our formulation (4.12) requires no index evaluations and only k accumulations for updating each entry in C, as opposed to k^2. Consequently, computations are significantly faster, while being solely based on general operations, lending itself towards a straightforward implementation on the GPU.

4.4.3 Results

The following case study and domain-expert feedback provided by atmospheric scientists demonstrates the utility of our method. We have been able to (i) produce factorizations of considerably higher quality than it was possible before, (ii) process and analyze ten times more spectra than in previous studies, and (iii) gain surprising insights enabled by the visualization.

Case study

The data we use as an example was collected from wood stove exhaust using a single particle mass spectrometer [LR05]. Factorizations of this data are used to quantify emission sources of biomass combustion. This aspect is of interest to atmospheric scientists, as biomass combustion is ubiquitous, while being suspected to play a key role in present day environmental concerns including health effects and climate change. The Pittsburgh June-July data (X) contains roughly 70k particle spectra in 256 dimensions and was factorized (in C and B) using an eight-dimensional basis. The error in this factorization can be quantified in relation to the data, $||X - CB||_F$ / $||X||_F$, producing a value of 31.1%. This magnitude of information loss is typical for SPMS factorization, making the need for analysis apparent. In our investigation, we first gain an overview of these errors in the projection shown in the center of Figure 4.12 based on error magnitude (y-axis) and irregularity (x-axis). Snapshots from a detailed view of selected errors are shown on the sides in the figure. The depth contouring in the projection shows that the majority of the data is factorized with good quality (low error magnitude/irregularity). However, large amounts of spectra are not well approximated. The contours of the projection depict two local maxima in error abundance, reflecting the spectra that are factorized by low and high error magnitude, respectively, while irregularity increases with magnitude.

These results support the initial assumption that there are important features in the data that are not covered by the factorization. Coarse classification of these error classes is provided by the coloring of points in the projection. There are three major error clusters visible in the projection, shown by the local abundance of green, blue, and red points. Selection of these points allows for detailed investigation of the corresponding residuals to be conducted in a high-level view. This reveals that the error types are characterized by major misfit of the factorization in the following features: (i) Pb+-predominant error in green cluster (372 spectra), (ii) NO^+, SiO^+ and Fe^+ in blue cluster (151 spectra), and (iii) $C_xH_y^+$-predominant in red cluster (7,851 spectra).

Having identified dominant error clusters, we investigate the gain in minimizing these errors. Figure 4.13 shows the estimated improvement that can be gained by introducing a basis vector that minimizes each of the error features. While the (numerical) gain in reducing the error feature outlined by the blue cluster is relatively low, it is considerably higher for the green and red clusters. Noticeably, the gain in introducing a basis vector for these clusters is higher than for other basis vectors (noted by index 0 and 5 in the figure), as computed by the initial factorization. Consequently, we have shown that this basis is sub-optimal and have found alternatives that improve the factorization.

As the initial factorization basis is shown to be sub-optimal in this analysis, the overall error of the factorization can be decreased, while keeping the same dimensionality of the

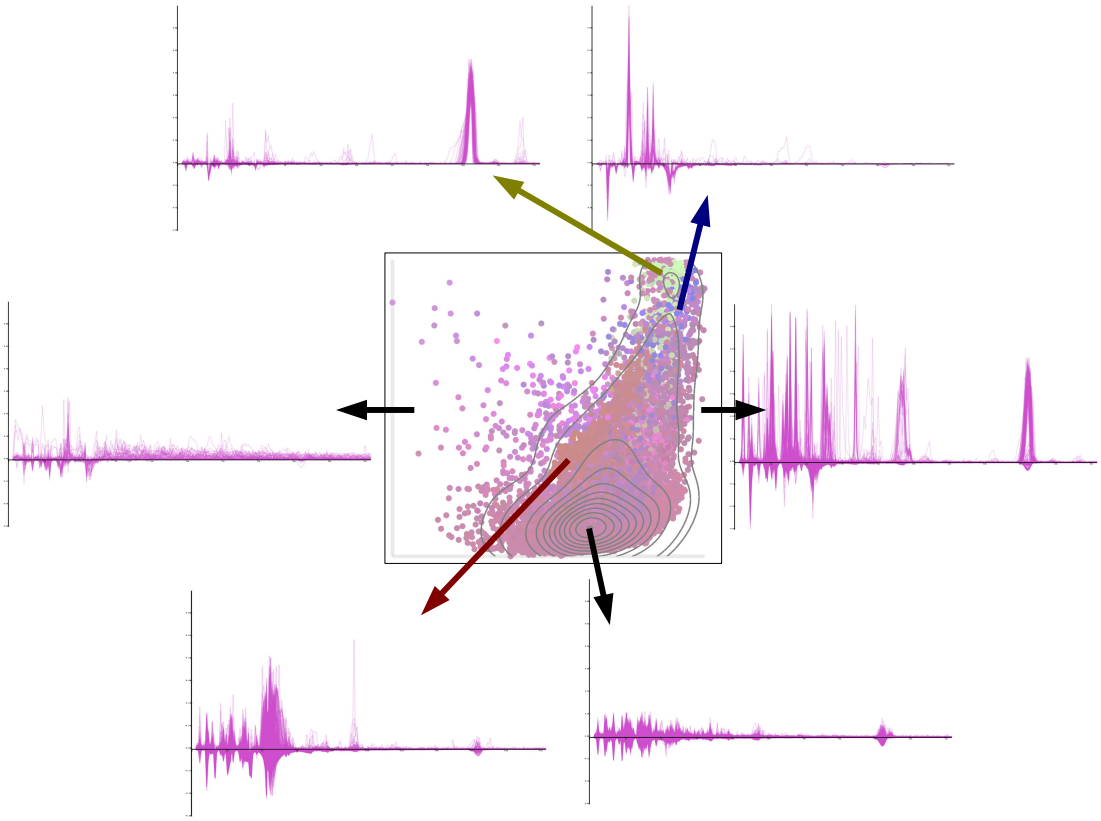
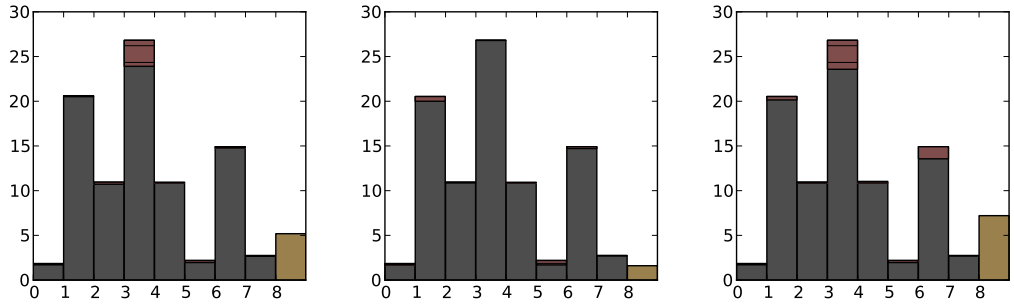


Figure 4.12: Errors of the factorization of Pittsburgh source sampling data, June-July, 2002. Selecting errors by color and/or region in the projection (center, also shown in Figure 4.11(c)) effectively filters high-level views and, thereby, makes possible a detailed data analysis by uncovering errors of high (right) or low (left) irregularity, magnitude, maxima of abundance (bottom right), and provides further classifications by color. Red (bottom left), green (top left), and blue (top right) error clusters are selected.

basis. With respect to refining the factorization, the sub-optimal parts of the basis can be deleted and/or the more suitable vectors (for the red and green error classes) added to the basis. Subsequently, the factorization is recomputed with the adjusted basis. In this experiment, we have deleted the sub-optimal parts and introduced the two candidates of higher gain instead. After convergence, the refined factorization features an overall error of 24.7% in relation to the original data. While being restricted to the same dimensionality of the basis as the initial factorization, these results represent an improvement of the overall error by 21.5%. An overview of the remaining error is depicted in Figure 4.14(a). Noticeably, both error features that were minimized in our refinement are not apparent in the projection. However, there are two new error clusters distinguishable at the top right corner of the projection, in addition to the blue cluster. These new clusters correspond to the two basis vectors that have been deleted in our refinement. Although of high magnitude and irregularity, the clusters contain only a small number of spectra.

Our experiments have shown that significant additional improvement of the factorization



(a) Gain in minimizing green errors

(b) Gain in minimizing blue errors

(c) Gain in minimizing red errors

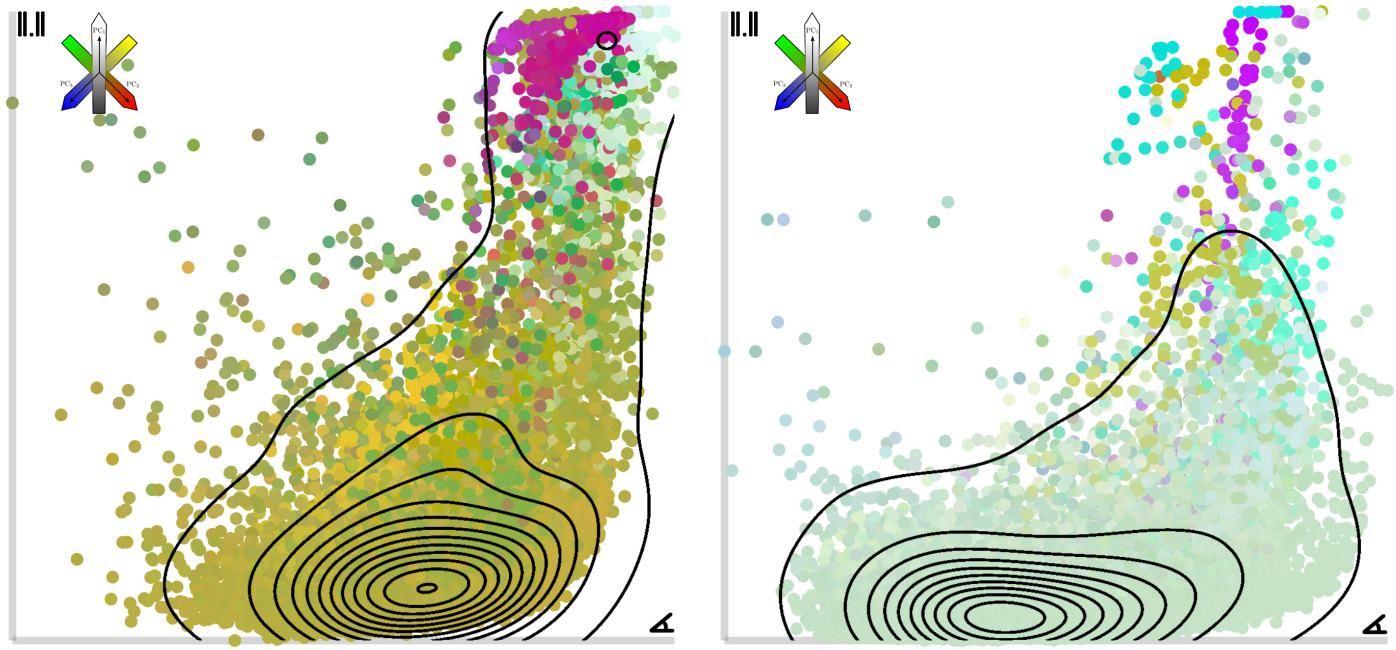
Figure 4.13: The numerical gain in introducing basis candidates minimizing specific errors is depicted in relation to the previous basis configuration (red = decrease). Sub-optimal parts of the factorization exhibit a smaller gain than the analyst's candidate ((a) and (c)). The analyst can add the candidate to the basis, delete existing parts, or continue analysis.

for this data set can only be gained by increasing the dimensionality of the basis. However, the amount of information that is consequently added decreases rapidly. Figure 4.14(b) shows the error projection for a factorization of this data using a 24-dimensional basis. By increasing basis dimensionality, an overall error of 14.8% with respect to the original data was achieved. These results make apparent the need for visual analysis in data factorization. Looking beyond the scope of this work, results also indicate that more research needs to be conducted to support application domains. As such, actively searching for specific error features may provide analysts with the ability to query factorization errors and to quantify the quality of the approximation with respect to these features.

Expert feedback

The recent advent of single particle and related real time techniques in atmospheric science has increased the quality and quantity of available data, so that improvements in data visualization and comprehension techniques are increasingly desired. Single particle mass spectrometers and other similar instruments that collect spectra in real time generate a tremendous amount of data of high dimensionality. These huge, complex data sets pose challenges for atmospheric scientists that need to analyze the data for various endpoints such as emissions source, atmospheric transformations and toxicity. The high dimensionality of the data also confounds comprehension by the atmospheric scientist because so few dimensions can be readily observed.

The methods presented here reduce the dimension of the data set by discovering the bases that underlie the data and visually present the resulting information to the scientist in a way that elucidates the factors that establish the basis as representing significant pollutant sources or atmospheric transformations. In typical studies, the common bases are hundreds or thousands of times more prevalent than the uncommon ones so techniques for identifying the bases must also take into account that bases with infrequent spectra may have lower variability so appear more significant. Data analysis must not arbitrarily



(a) Errors of factorization using an 8-dimensional basis

(b) Errors of factorization using a 24-dimensional basis

Figure 4.14: (a) Controlled refinement of the factorization produced a decrease of the overall error by 21.5% in relation to the initial solution. (b) Further decrease was achieved by increasing the basis dimensionality, here accounting for an overall error of 14.8% in relation to the original data.

exclude this important information but instead communicate important basis properties, such as efficiency, local minima, and information loss, to the scientist. The system described here supports this objective and enables more accurate and verifiable data analysis. The visualization makes it possible to analyze and classify different basis sets with respect to information loss and different objectives. Alternative basis configurations can be readily identified, by a cluster in the projection, and then selected for analysis. Visually comparing the efficiency of basis vectors enables one to explore alternatives and identify new bases, ultimately producing factorizations of higher quality. The interactive nature of this new tool enables ready exploration of hypotheses and discovery of aspects of such large data sets that one might not be able to discover otherwise.

4.4.4 Conclusions

It is important and difficult to address the issue of "error" in any data factorization method and application setting. In our case, error can be associated with the result of approximating original data in a lower-dimensional space. The error is directly influenced by the number of chosen basis vectors and the efficiency of the basis transformation. This multi-criteria and non-convex optimization problem cannot be solved in an optimal way by known algorithms. It is therefore crucially important to have the data analyst play an integral role in the entire process of factorization: by specifying the number of dimensions needed for lower-dimensional approximation, specifying individual basis vectors, and determining what is and what is not a "good approximation." Error quantification and visualization, combined with the ability to interactively influence the data factorization/approximation process, is thus a highly desirable and crucially important component of any system aimed at dramatically reducing the dimensionality of a complex and high-dimensional data set to assist effectively with understanding. Our approach is exactly supporting this objective.

4.5 Generalizing model-based visual analysis

The described methods and approaches to analyzing single particle mass spectra have been shown to prevail and to clearly succeed previous approaches in the application of air quality research. Thereby, they represent a significant contribution to the application area. At this point, the present work could be concluded. However, the research triggered by this application prompts the question of whether the concepts can be applied to other applications to achieve equal satisfactory results and a wider contribution. In this concluding section, we evaluate this potential in answering where and how the described specific concepts can be generalized to be made applicable to other application areas.

One key characteristic of the present work is the close link between machine learning, visualization, and interaction. As such, this work affiliates to the relatively young research field of Visual Analytics. The scope of Visual Analytics can be described as an integral approach combining data analysis, visualization, and human factors [KMS*08]. The process of Visual Analytics is defined by the key concepts: data, hypothesis, visualization, and insight; as well as the transition functions from one concept to another. Broadly speaking, the process involves data management and transformation, visualization and exploration, as well as methods to support hypothesis formulation and verification aimed towards the ultimate goal to generate new insights from data. In general, the process does not focus on

a specific analytical task, but is rather designed to illustrate the scope of visual analytics, covering the intersecting set between visual and analytical approaches. In contrast, the work described here is also located at this intersection but represents a very specific sub-setting that is far more linear than the full scope of this process by its aim to facilitate a single analytical task. To this end, the goal is also not to derive insight in general but to produce results for a very specific analysis problem. The analytical task is well defined by the application, although in general, cannot be solved automatically. The task is to produce parameters of a pre-defined analytical model that map to the data as observed. In this setting, the transformation from model to data space is known and well defined. Verification and error analysis can be facilitated by comparing the observed data to the generated data from the model and showing the physical entities generated by the model parameters in data space.

The core idea of our approach to solving the application problem described in the previous sections is to divide and conquer the visual analysis process by solving each of the following sub-problems: mathematical modeling of the analysis problem, algorithm design for solving the mathematical model, visual representation of the analytical parameters, interaction for steering computations, and verification for physical and numerical correctness. Each of these conceptual building blocks are adjustable according to the application at hand. While some of the visual representations described are designed for single particle mass spectra, the interface and verification methods are designed for non-negative matrix factorization, not restrictive to any specific application. The analysis approach as a whole is not restricted to matrix factorization or to any particular application but is designed to facilitate a task as defined above. In the following, we describe the conceptual building blocks behind our approach to derive a more general concept of model-based visual analysis. Further, we discuss applications that can benefit from this concept.

4.5.1 Design methodology for model-based visual analysis

We define model-based visual analysis as an integral approach to compute parameters for a pre-defined analytical model that map to given data. In particular, the design choices that are described in this section are intended to facilitate analysis of mathematical ill-posed problems. The core of our approach is to incorporate the scientist in every step of both the design and the usage of the final system. Scientists have extensive experience in interpreting complex data, identifying different sources, common features, and noise. Under a mathematically ill-posed problem setting, analytical methodologies are well advised to incorporate this knowledge as best as possible into every step of the analysis process. Our goal is to provide guidelines for the design of a system that can help scientists solve these ill-posed problem in a semi-automated manner, where the computer does the “heavy lifting” and maintains unsubjective mathematical rigor, while the scientists oversees and steers computations towards physical correctness and interpretability.

Although any particular solution to this analytical task must be specific to data and application, we provide general design hints for implementing such a methodology that are independent of data and application. Thereby, we pursue an interdisciplinary and integral approach involving mathematical modeling, machine learning, visualization, and human-computer interaction. In particular, our approach involves an intensive background and

requirement analysis to clearly define the analytical question of the scientist mathematically. Subsequent design of the algorithm for parameter estimation incorporates user input and steering, such that the user can be in the loop of computations. By representing analytical parameters and approximation errors in data space, we incorporate user knowledge and involve the scientist in decision making. Finally, this design involves interaction mechanisms for interactive steering, as well as approaches for verification. We provide our guidelines in the corresponding thematic blocks.

Mathematical modeling

Beginning the design of a method to facilitate model-based visual analysis, we must first define the intended analysis mathematically. Based on extensive background research and discussions with application scientists, the basic mathematical conditions of question and answer to the application problem must be clear and well defined. Thereby, we mathematically define the analytical model space M and data space D, both being high-dimensional, as well as the mapping functions between these spaces. While these definitions may be given directly by the applications, they may also be given implicitly, leading to the following design tasks.

- Define the data space D, as well as the parameters of the analytical model.

$$p_1, ..., p_n \in M \tag{4.14}$$

  These parameters are the final answer to the application problem and the result of the intended analysis. It is important to note that the analytical model space is defined (explicitly or implicitly) by the application and that it is not necessarily the ideal space with regard to mathematical properties or optimization. As such, it may be ambiguous or over-defined. Parameters may not necessarily be independent of each other and the space they reside in may not be metric. Nevertheless, it is important to model these properties and to define, for example, measures for the similarity of parameters, boundary conditions, and restrictions. Likewise, these properties must be defined for the data space, including the similarity measure δ between data points, as well as metric properties of the space.

- Model the mapping from analytical model to data space.

$$f: \quad M \quad \rightarrow D \tag{4.15}$$

  One of the aspects that renders model-based visual analysis unique in a sense is the clear definition of the transformation between analytical model to data space. This is also the essential basis for designing an optimization for the parameter fit.

- Describe physical complement of the model parameters in data space.

$$f: \quad p \subset M \quad \rightarrow d \subset D \tag{4.16}$$

  In order to render optimization results interpretable and verifiable by domain experts, it is essential that the corresponding object or effect is visualized in a way that

relates the fit for each parameter to features in the data. Therefore, the physical complements of the model parameters must be defined.

Machine learning design

Having mathematically defined the analytical problem, the next step is to design an appropriate algorithm to estimate the parameters in M such that they map to given data in D. For this, an inverse transformation f^{-1} is defined as an optimization algorithm minimizing a cost function.

$$\begin{aligned} f^{-1}: \quad D &\rightarrow M \\ X &\mapsto (p_1,...,p_n) \text{ , s.t.} \\ ||X - f(p_1,...,p_n)|| &\rightarrow \min \end{aligned} \tag{4.17}$$

Considerable care should be placed on the definition of a domain-appropriate error measure for the misfit of parameters. Generally, this involves comparing the original data to the data generated by the estimated parameter in data space using the defined mapping f. Based on the Euclidean distance, a general choice can be the squared Frobenius norm $||.||_F^2$. Domain-specific measures are based on the similarity measure δ.

The design of the parameter optimization is of course not merely based on the cost function but also largely influenced on the data- and domain-specific properties noted in the mathematical modeling phase. These properties can render the optimization problem non-convex, ultimately leaving a limited amount of choices for the optimization approach. While gradient descent type methods have the flexibility to incorporate such restrictions, they are often prone to slow convergence and sub-optimal solutions. This also holds for the matrix factorization method applied in this work, although we have described how these problems can be alleviated by multiplicative update rules, implementation on the GPU, and the careful design of mechanisms for steering and verification. To facilitate these mechanisms to the scientist, parameters of the algorithm must be made adaptable during run time, such that inter-mediate solutions can be adapted and refined.

Visualization

In our design methodology, visualization takes the key role to render the parameter approximation visually accessible to the scientist. Due to noise, outliers, and sub-optimal solutions, this process cannot be automated. Therefore, visual verification is required by the analyst to ensure physically meaningful results. This in turn requires the visualization to represent three analytical entities in a manner that allows for exploration and comparative analysis. These entities are (i) the data $X \subset D$, (ii) the analysis parameter $P \subset M$ (sub-results of the optimization), and (iii) the error $X - f(P) \subset D$ of the parameter mapping. In particular, knowledge about the total mapping error is not enough for this task but instead the mapping error must be made interpretable corresponding to approximate solutions for the analysis parameters and to data features. This amounts to the following sub-tasks.

- Design the explorative data visualization.
  The first step in assessing the results provided by the optimization algorithm is to first

understand the data. The design of an explorative data visualization involves first defining an appropriate visual representation of the data that adheres to application requirements and constraints. As our methodology relies on the scientist's ability to identify patterns and features within the data, the geometry of domain-typical data representations should not be neglected. Instead, appropriate representations must be discussed with and agreed upon by the intended user. In some domains, transfer functions can be a good compromise to decrease visual clutter while maintaining the generally accepted basic geometric form. Further, methods for interaction should be provided to explore data effectively and to identify salient features.

- Design the visualization of model parameters and approximation error.
  Designed in correspondence with the data visualization, the visualization of model parameters and approximation errors must be similar enough to facilitate direct comparison, while being distinctly identifiable. Thereby, the goal is to facilitate comparative analysis between parameters, data features, and error. The error should be incorporated both by absolute and relative values. While absolute values are given by $X - f(p_1,...,p_n)$, relative values, interpreted by an over- or underfitting, are given by $f(p_1,...,p_n)/X$. If the data visualization allows the necessary degree of freedom, color coding the relative error values in the data representation can facilitate quick visual assessment of misfitted regions.

  Where possible, the visualization should highlight the dependencies between data, parameters, and error. The user must gain an overview by being able to assess quickly which dimensions are important for the parameter solution and where errors are. Parameters may have physical complements in data space that can be visualized when triggered by interaction. This visualization should facilitate knowledge to the scientists on how the parameters map to the data, how things relate, and how well which part of the data are captured by the parameters. Thus, the goal is to visualize the analytical reasoning of the optimization process, as well as the mapping to the data, in terms of the function f.

Interaction

Effective visual analysis and manipulation requires interaction. Here, the goal of the scientist is to visually evaluate and to interface the optimization algorithm. If the current solution of the algorithm is inappropriate from a physical or semantical point of view, the analyst should be able to change the optimization parameters by interacting with the visualization. This requires intuitive operands that serve as an interface for interaction. If possible, the scientist should be able to redefine and adjust parameters directly in the data view. General interaction mechanisms to facilitate this methodology are methods for exploration and refinement.

- Design interactive exploration of data features, parameter space, and error.
  General mechanisms to facilitate explorative analysis are overview and detail, as well as zooming and panning that can be embedded in any visualization. With increasing number of dimensions and data points, these methods represent absolute prerequisites

for exploration. Further, methods for selection and filtering aid the user in more directed analysis. Other options include to hide or highlight parameters, as well as the corresponding data features, or to trigger the visualization of physical complements for select parameters. Additionally, options to exclude or include data dimensions, data points, or parameters can further aid the scientist in specific analysis tasks.

- Design interactive refinement of parameters.
  If parameter misfit is identified during explorative analysis, our methodology entails an interactive refinement of the parameters. With this, the scientist can guide the algorithm semi-automatically out of local minima and to find more adequate solutions. In a monitoring and steering environment, the first interaction is to stop the optimization process, in order to change the parameter setting and to later resume computations with adjusted parameters. Adding or removing certain parameters can be useful in settings where the solution is over- or under-determined.

  In order to adjust parameters, appropriate methods facilitate direct interaction with the visual primitives of these parameters to intuitively and quickly change parameter values. In a computational steering context, methods to interactively define boundaries for certain parameter values can aid the algorithm to traverse into regions of the parameter space that would otherwise not be reachable. This method can also facilitate the definition of physical restrictions for parameters after the design of the optimization algorithm. Finally, the option to save and load (sub-)results of the parameters can aid scientists to compare solutions or save analysis time.

Verification

Next to the visual verification of parameter values and errors, our methodology entails assessment of local optimality of approximation errors by exploration of the error space and comparison between different alternatives. Thereby, the goal is to facilitate analysis of the quality of results from non-convex optimization. This can also aid in the interpretation of results with respect to uncertainty. Depending on the application, it may be beneficial to analyze the optimization landscape, that is the scalar-valued error field for the parameter space (or a proximity around possibly sub-optimal parameters).

The core of our approach to assessing the quality of optimization results is based on the following principle. Approximation errors (in $X - f(P)$) that follow a small normal distribution can be regarded as noise and should thus be visually filtered out, while specific (non-gaussian) error features represent a misfit of the parameter optimization to account for corresponding data features. In Section 4.4.2, we describe a method to visually characterize errors by abundance, similarity, sparsity, and magnitude. This is generally applicable to any verification context in model-based visual analysis. The projection of the error landscape acts as an interface for exploration and selection in a multi-view environment. Upon selection of error features within this projection, a more detailed visual representation allows analysis of the corresponding absolute errors in relation to the data and parameters.

Our methodology for finding more optimal results, as described in 4.4.2, is applicable to settings where the analysis model entails linear projection of the data. This is achieved by introducing the data feature corresponding to select errors in the projection, thus

minimizing the identified error. Although, this is not applicable in any analysis model, the general idea can be transferred. Generally, any salient error feature is a feature in the data that is not mapped by the analytical parameters. The goal is therefore to characterize the salient features in the error, compute the corresponding data features that are misrepresented, and to compute an optimal set of parameters that minimize this specific error feature. Our methodology further entails a comparison of the benefit from minimizing this specific error feature with respect to the previous solution. Thereby, the user can adapt the parameter set and compare different locally optimal solutions, with respect to the overall approximation error or specific parts of the data being captured.

4.5.2 Applications

The task of model-based visual analysis is to produce parameters of a pre-defined analytical model that map to given data. This represents a setting where the focus lies not on exploration of the unknown but on analysis of the known. Additionally, the goal of analysis is to produce specific analytical parameters that do not reside in data space. For example, this setting frequently occurs to classification problems. In the following, we list applications that could benefit from our approach and describe examples of domain-specific classification problems fitting to the scope of our methodology.

- Physics:
  The evaluation of experiments or simulation with respect to pre-defined parameters, physical restrictions, and analytical model maps directly to our methodology.
- Biology:
  We have applied our methodology successfully in biology where it was used to study conformational changes of proteins in molecular biophysics [EGF*13]. To facilitate model-based visual analysis of circular dichroism spectra, we have applied a different optimization and new visual encodings, while following our general design methodology. The results show that our method can be transferred and analogously applied in the analysis of protein conformational changes, while leading to parallel success.
- Geology:
  In geochemistry, an example would be the classification of isotope and trace metal geochemistry of the shells of recent and fossil organisms.
- Environmental science:
  Next to atmospheric science, our methodology could facilitate a way to better understand and tune algorithmic parameters in climate modeling.
- Engineering:
  There are arguably numerous optimization problems in engineering that cannot be automated but require the knowledge and experience of an expert. Adding to the scope of model-based analysis, specific engineering disciplines that include such problem settings are civil, chemical, electrical, and mechanical engineering. A close link between optimization and human expert is especially desirable in settings where important design decisions are made, for example, in urban planning, water management, and construction.

4.5.3 Conclusions

Visual Analytics can be defined by the interlinking of methods from machine learning, visualization, and user interaction to a holistic scientific workflow. Here, we define model-based visual analysis as a subset of this field by interlinking methods of these domains to facilitate a single and specific analytical task, that is based on a well-defined analysis model but not straight-forward to solve algorithmically. Thus, it involves the user with the help of visualization, facilitating means for interaction and verification. Hereby, the goal is neither primarily to explore, nor to find hypothesis, but to find parameters in a pre-defined analysis model that map to given data. Our methodology entails verification and error analysis by comparing the observed data to the data generated by the model parameters. Thereby, we pursue an interdisciplinary and integral approach that involves mathematical modeling of the analysis problem, algorithm design for solving the mathematical model, visual representation of the analytical parameters, interaction for steering computations, and verification for physical and numerical correctness.

Model-based visual analysis is of interest in applications that involve analytical problems where automated methods are not sophisticated enough to find reliable and physically correct solutions. These problem classes include non-convex optimization problems, such as non-negative matrix factorization, operating under application-specific requirements and constraints, and therefore have to involve the scientist to draw from his knowledge and intuition. With our generalization to model-based visual analysis, we contribute a design basis to develop a holistic scientific workflow by means of which application scientists are able to conduct exactly this type of analysis effectively.

Although our approach was originally developed to solve a specific analysis problem in one application, we have shown that the underlying design principles do in fact extend to a far greater field of application. As we have discussed, the analytical tasks solved by model-based analysis are abundant in physical sciences and engineering disciplines. In these applications, there is a need for visual aid in analyzing increasingly vast amounts of data and facilitating user intervention where the machine does not produce adequate solutions. Model-based visual analysis addresses exactly this need.

CHAPTER 5

Conclusion

Due to enhanced data acquisition and analysis methodologies in almost all application domains, progressively larger and inherently more complex (multivariate) data sets are being produced. This dissertation describes new approaches to visually explore and analyze these data sets. Our main contributions are:

- Visualizing values in dimension reduction
- Level-of-detail in dimension reduction
- Visual interface for steering and verification of non-negative matrix factorization
- Methodology to analyzing approximation errors of non-negative matrix factorization
- A general design methodology for model-based visual analysis

In the subject of data exploration, we focus on incorporating methods from dimension reduction to facilitate more visually scalable and intuitive representations of multivariate data. We describe how the incorporation of data values within a lower-dimensional embedding enhances its analytical capabilities and effectively counters ambiguity of the mapping [ERHH11, REM*12, EHHR13]. Our approach encodes data values into a tree metaphor that illustrates the structure of data sets and helps to reinforce a mental mapping of inherent relationships. The computation and display of the tree structure is optimized with regard to depicting minimal redundancies, while unique mechanisms for interaction enhance exploratory capabilities and support analysis of global trends.

Further, we describe how dimension reduction can be embedded to facilitate analysis and exploration on different levels of detail, from global trends to local properties of the manifold [EKHS14]. By utilizing criteria such as relative neighborhood and Mahalanobis distance, the manifold underlying the data is approximated hierarchically. In order to render these levels visually assessable, local projections are recursively embedded into ellipses, each utilizing full degree of freedom to produce an optimal representation corresponding to the level of detail. Means to interactively traverse these levels enable the user to explore local properties gradually, thereby enabling effective analysis and exploration of the shape and topology of clusters comprised in high-dimensional data.

In Addition to our contributions in explorative data visualization, we describe work which may prove to instigate a new basic research direction in visual analytics. Working closely with atmospheric scientists, we have developed a fundamentally new algorithmic and analytical methodology for visualization, by which it is now possible to characterize aerosol source contributions [EGG*12]. Our methodology entails a visual mapping of the application's analytical model and the visualization of an optimization algorithm for approximating the model's parameters. As this model is mathematically ill-posed and the corresponding optimization problem non-convex, approximation results are not guaranteed to be numerically the best solution or physically meaningful with respect to their semantical interpretation. Therefore, our methodology entails the use of visualization and interaction to steer the approximation algorithm and analyze the errors that results from it towards producing results of physical meaning and mathematical rigor. Thereby, we overcome limitations of previous methods that use machine learning algorithms as a black-box scheme and move toward more physically sensible computing.

Further, we describe how errors in non-negative matrix factorization may be quantified, visualized, and methodically analyzed, in order to provide scientists with the ability to interactively improve the results [EHH*13]. Our methodology for user-guided error-aware data factorization entails an assessment of the amount of information contributed by each dimension of the approximation, an effective combination of visualization techniques to highlight, filter, and analyze error features, as well as novel means to interactively refine these errors, both to physical meaning and to numerical optimality. Following this methodology, the scientist can improve the factorization quality and consequently overcome "being stuck" in local minima of non-convex optimization. Due to the high degree of interactivity in this analysis, our method also provides awareness about the information loss associated with the dimension reduction process and allows for an educated decision regarding the degree of freedom needed to approximate multivariate data.

Our methodology has been implemented in a single framework for visual analysis, tested, and evaluated by domain scientists, serving as a proof of concept. Using our framework, atmospheric scientists have been able to (i) reproduce established findings in a fraction of the time, (ii) process and analyze a hundred times more spectra than in previous studies, and (iii) gain surprising new insights enabled by the visualization. The outstanding success in the application of air quality research naturally prompts the question of its applicability to other domains. To answer this question, we derive the key design choices that underlie and transcend beyond application-specific solutions and describe a general design methodology to computing parameters of a pre-defined analytical model that map to multivariate data [EHHS14]. Thereby, we describe how a data- and application-specific analysis task can be made visually accessible by representing analysis parameters and error to the scientist and providing further means to interactively verify computations with the scientist in-the-loop. Core application areas in natural sciences and engineering disciplines are identified that can benefit from our approach to model-based visual analysis.

Bibliography

[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.

[AC91] Alpern B., Carter L.:
The hyperbox.
In *VIS '91: Proceedings of the 2nd conference on Visualization '91* (Los Alamitos, CA, USA, 1991), IEEE Computer Society Press, pp. 133–139.

[AdO04] Artero A. O., de Oliveira M. C. F.:
Levkowitz h.: Uncovering clusters in crowded parallel coordinates visualizations.
IEEE Symp. on Information Visualization (2004).

[And72] Andrews D.:
Plots of high dimensional data.
Biometrics (1972).

[AP98] Ascher U., Petzhodl L.:
Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations.
SIAM, 1998.

[BBKY06] Bronstein M. M., Bronstein A. M., Kimmel R., Yavneh I.:
Multigrid multidimensional scaling.
Numerical Linear Algebra with Applications (NLAA) 13 (March-April 2006), 149–171.

[BG05] Borg I., Groenen P.:
Modern Multidimensional Scaling: Theory and Applications.
Springer, 2005.

[Bur10] Burges C. J. C.:
Dimension reduction: A guided tour.
Foundations and Trends in Machine Learning 2, 4 (2010).

[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.

[BZW09] Bein K. J., Zhao Y., Wexler A. S.:
Conditional sampling for source-oriented toxicological studies using a single particle mass spectrometer.
Environmental science technology 43, 24 (2009), 9445–9452.

[BZWJ05] Bein K. J., Zhao Y. J., Wexler A. S., Johnston M. V.:
Speciation of size-resolved individual ultrafine particles in pittsburgh, pennsylvania.
Journal of Geophysical Research 110, D7 (2005), 1–22.

[CC80] Chatfield C., Collins A. J.:
Introduction to Multivariate Analysis.
London: Chapman and Hall, 1980.

[CC94] Cox T., Cox M.:
Multidimensional Scaling.
Chapman & Hall, London, 1994.

[Che73] Chernoff H.:
The use of faces to represent points in k-dimensional space graphically.
Journal of the American Statistical Association (1973), 361–368.

[CJ10] Comon P., Jutten C.:
Handbook of Blind Source Separation: Independent Component Analysis and Applications, 1st ed.
Academic Press, 2010.

[CMKS08] Carlsen D., Malone M., Kollat J., Simpson T. W.:
Evaluating the performance of visual steering commands for user-guided pareto frontier sampling during trade space exploration.
ASME Conference Proceedings 2008, 43253 (2008), 499–509.

[Com94] Comon P.:
Independent component analysis, a new concept?
Signal Process. 36, 3 (Apr. 1994), 287–314.

[CP07] Chen D., Plemmons R. J.:
Nonnegativity constraints in numerical analysis historical comments on enforcing nonnegativity.
Office (2007), 1–32.

[DDRvdV00] Demmel J., Dongarra J., Ruhe A., van der Vorst H.:
Templates for the solution of algebraic eigenvalue problems: a practical guide.
Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[Dij59] Dijkstra E. W.:
A note on two problems in connexion with graphs.
Numerische Mathematik 1 (1959), 269–271.

[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.

[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.

[EGF*13] Engel D., Gillmann C., Fiedler S., Keller S., Scheler I., Hagen H., Garth C.:
Visual analysis of circular dichroism spectra in molecular biophysics.
In *Jankun-Kelly, T.J., Andrienko, G., Maciejewski, R., Tory, M., Lee, B., Leitte, H., eds., extended abstract and poster at the IEEE conference on visualization (IEEE VIS)* (2013).

[EGG*12] Engel D., Greff K., Garth C., Bein K., Wexler A. S., Hamann B., Hagen H.:
Visual steering and verification of mass spectrometry data factorization in air quality research.
IEEE Trans. Vis. Comput. Graph. 18, 12 (2012), 2275–2284.

[EHH12] Engel D., Hüttenberger L., Hamann B.:
A survey of dimension reduction methods for high-dimensional data analysis and visualization.
In *Garth, C., Middel, A. and Hagen, H., eds., Visualization of Large and Unstructured Data Sets - Applications in Geospatial Planning, Modeling and Engineering, OpenAccess Series in Informatics (OASIcs), Schloss Dagstuhl, vol. 27* (2012), pp. 135–149.

[EHH*13] Engel D., Hummel M., Hoepel F., Bein K., Wexler A. S., Garth C., Hamann B., Hagen H.:
Towards high-dimensional data analysis in air quality research.
Comput. Graph. Forum 32, 3 (2013), 101–110.

[EHHR13] Engel D., Hagen H., Hamann B., Rosenbaum R.:
Structural decomposition trees: Semantic and practical implications.
Computer Vision, Imaging and Computer Graphics - Theory and Applications, Communications in Computer and Information Science (CCIS) Series 359 (2013), 193–208.

[EHHS14] Engel D., Hamann B., Hagen H., Scheler I.:
A generalization to model-based visual analysis.

Submission in progress: IEEE Transactions on Visualization and Computer Graphics (TVCG) (2014).

[EKHS14] Engel D., Karer B., Hagen H., Scheler I.:
Level-of-detail by self-embedding local projections.
Submission in progress: IEEE Transactions on Visualization and Computer Graphics (TVCG) (2014).

[ELL01] Everitt B. S., Landau S., Leese M.:
Cluster Analysis, 4th ed.
Arnold, London, 2001.

[EMO*11] Esser E., Möller M., Osher S., Sapiro G., Xin J.:
A convex model for non-negative matrix factorization and dimensionality reduction on physical space.
Arxiv preprint arXiv11020844 stat.ML (2011), 14.

[ERHH11] Engel D., Rosenbaum R., Hamann B., Hagen H.:
Structural decomposition trees.
Comput. Graph. Forum 30, 3 (2011), 921–930.

[FBF77] Friedman J., Bentley J., Finkel R.:
An algorithm for finding best matches in logarithmic expected time.
ACM Transactions on Mathematical Software 3, 3 (1977), 290–226.

[FH08] Farin G., Hansford D.:
Mathematical Principles for Scientific Computing and Visualization.
AK Peters Ltd, 2008.

[FR11] Ferdosi B. J., Roerdink J. B. T. M.:
Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis.
Comput. Graph. Forum 30, 3 (2011), 1121–1130.

[GAR*10] Gross D. S., Atlas R., Rzeszotarski J., Turetsky E., Christensen J., Benzaid S., Olson J., Smith T., Steinberg L., Sulman J., Ritz A., Anderson B., Nelson C., Musicant D. R., Chen L., Snyder D. C., Schauer J. J.:
Environmental chemistry through intelligent atmospheric data analysis.
Environmental Modelling & Software 25, 6 (2010), 760 – 769.

[GTC01] Grinstein G., Trutschl M., Cvek U.:
High-dimensional visualizations.
In *Proceedings of Visual Data Mining workshop, KDD'2001* (2001).

[Han05] Han J.:
Data Mining: Concepts and Techniques.
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[Har08] Harville D. A.:
Matrix Algebra From a Statistician's Perspective.
Springer, 2008.

[HG97] Hoffman P., Grinstein G.:
Visualizations for high dimensional data mining - table visualizations, 1997.

[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.

[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.

[Hin88] Hinterberger H.:
Using graphical information from a grid file's directory to visualize patterns in cartesian product spaces.
In *Proceedings on International Workshop on Computational Geometry on Computational Geometry and its Applications* (New York, NY, USA, 1988), Springer-Verlag New York, Inc.

[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)* (2002), pp. 127–130.

[HO00] Hyvärinen A., Oja E.:
Independent component analysis: algorithms and applications.
Neural Networks 13, 4-5 (2000), 411–430.

[HR09] Hurley N., Rickard S.:
Comparing measures of sparsity.
IEEE Transactions on Information Theory 55, 10 (2009), 4723–4741.

[HS07] Härdle W., Simar L.:
Applied Multivariate Statistical Analysis, 2nd edition ed.
Springer, 2007.

[HSM01] Hand D. J., Smyth P., Mannila H.:
Principles of data mining.
MIT Press, Cambridge, MA, USA, 2001.

[HW09] Heinrich J., Weiskopf D.:
Continuous parallel coordinates.
IEEE Transactions on Visualization and Computer Graphics 15, 6 (2009), 1531–1538.

[ID90] Inselberg A., Dimsdale B.:
Parallel coordinates: a tool for visualizing multidimensional geometry.

IEEE Visualization (1990).

[ID94] Inselberg A., Dimsdale B.:
Multidimensional lines ii: Proximity and applications.
SIAM Journal on Applied Mathematics (1994).

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

[Ins85] Inselberg A.:
The plane with parallel coordinates.
Visual Computer (1985).

[Ins09a] Inselberg A.:
Parallel Coordinates.
Springer, 2009.

[Ins09b] Inselberg A.:
Parallel Coordinates: Visual Multidimensional Geometry and Its Applications.
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2009.

[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, 6 (2008), 1459–1466.

[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.

[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–.

[Jol02] Jolliffe I. T.:
Principal Component Analysis, second ed.
Springer, 2002.

[KAK95] Keim D. A., Ankerst M., Kriegel H.-P.:
Recursive pattern: A technique for visualizing very large amounts of data.
In *VIS '95: Proceedings of the 6th conference on Visualization '95* (Washington, DC, USA, 1995), IEEE Computer Society, p. 279.

[Kan00] Kandogan E.:

Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions.
In *In Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics* (2000), pp. 9–12.

[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.

[Kar14] Karer B.:
Manifold learning and projection of high-dimensional data using graph abstraction.
Bachelor thesis. University of Kaiserslautern, Germany, 2014.

[KBHH05] Kim E., Brown S. G., Hafner H. R., Hopke P. K.:
Characterization of non-methane volatile organic compounds sources in houston during 2001 using positive matrix factorization.
Atmospheric Environment 39, 32 (2005), 5934–5946.

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

[KMP10] Koutis I., Miller G. L., Peng R.:
Approaching optimality for solving sdd linear systems.
In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science* (Washington, DC, USA, 2010), FOCS '10, IEEE Computer Society, pp. 235–244.

[KMS*08] Keim D. A., Mansmann F., Schneidewind J., Ziegler H., Thomas J.:
Visual analytics: Scope and challenges.
In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics, Springer, Lecture Notes In Computer Science (lncs)* (December 2008).

[KP08] Kim J., Park H.:
Fast nonnegative matrix factorization: an active-set-like method and comparisons.
Science (2008).

[Kru64] Kruskal J.:
Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis.
Psychometrika 29 (1964).

[KTH*02] Kreylos O., Tesdall A. M., Hamann B., Hunter J. K., Joy K. I.:

Interactive Visualization and Steering of CFD Simulations.
Eurographics Association, 2002, pp. 25–34.

[KTT98] Kearsley A., Tapia R., Trosset M.:
The solution of the metric stress and sstress problems in multidimensional scaling using newton's method.
Computational Statistics 13, 3 (1998), 369–396.

[Lan69] Lankford P. M.:
Regionalization: Theory and alternative algorithms.
Geographical Analysis 1, 2 (1969), 196–212.

[LGD*05] Laramee R. S., Garth C., Doleisch H., Schneider J., Hauser H., Hagen H.:
Visual analysis and exploration of fluid flow in a cooling jacket.
In *In Proceedings IEEE Visualization 2005* (2005), pp. 623–630.

[Loh95] Lohninger H.:
Multivariate exploratory data analysis by means of inspect.
In *SoftwareDevelopment in Chemistry* (1995), pp. 91–98.

[LR05] Lipsky E., Robinson A.:
Design and evaluation of a portable dilution sampling system for measuring fine particle emissions from combustion systems.
Aerosol Science and Technology 39, 6 (2005), 542–553.

[LS00] Lee D. D., Seung H. S.:
Algorithms for non-negative matrix factorization.
In *In NIPS* (2000), MIT Press, pp. 556–562.

[LV10] Lee J. A., Verleysen M.:
Unsupervised dimensionality reduction: Overview and recent advances.
In *IJCNN* (2010), IEEE, pp. 1–8.

[Man86] Manly B. F.:
Multivariate statistical methods: a primer.
Chapman & Hall, Ltd., London, UK, UK, 1986.

[MC00] Messac A., Chen X.:
Visualizing the optimization process in real-time using physical programming.
Engineering Optimization 32, 6 (2000), 721–747.

[Mit98] Mitchell M.:
An Introduction to Genetic Algorithms.
MIT Press, Cambridge, MA, USA, 1998.

[MM08] McDonnell K. T., Mueller K.:
Illustrative parallel coordinates.
IEEE-VGTC Symposium on Visualization 2008 (2008).

[MMW03] Murphy D. M., Middlebrook A. M., Warshawsky M.:

Cluster analysis of data from the particle analysis by laser mass spectrometry (palms) instrument.
Aerosol Science and Technology 37, 4 (2003), 382–391.

[MW02] Moustafa R., Wegman E. J.:
On some generalizations of parallel coordinate plots, seeing a million.
In *In Proceedings of 2002 Data Visualization Workshop (Rain am Lech (nr* (2002).

[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.

[PB95] Pal N., Bezdek J.:
On cluster validity for the fuzzy c-means model.
IEEE TFS 3, 3 (1995), 370–379.

[Pea01] Pearson K.:
On lines and planes of closest fit to systems of points in space.
Philosophical Magazine 2, 6 (1901), 559–572.

[PEP*11] Paulovich F., Eler D., Poco J., Botha C., Minghim R., Nonato L.:
Piece wise laplacian-based projection for interactive data exploration and organization.
Computer Graphics Forum 30, 3 (2011), 1091–1100.

[PG88] Pickett R. M., Grinstein G. G.:
Iconographics displays for visualizing multidimensional data.
In *Proceedings of the 1988 IEEE Conference on Systems, Man and Cybernetics* (1988).

[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.

[PS82] Papadimitriou C. H., Steiglitz K.:
Combinatorial optimization: algorithms and complexity.
Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1982.

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

[PWR04] Peng W., Ward M. O., Rundensteiner E. A.:

Clutter reduction in multi-dimensional data visualization using dimension reordering.
In *In INFOVIS 04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS04* (2004), IEEE Computer Society, pp. 89–96.

[RC94] Rao R., Card S.:
The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information.
In *In Proceedings of ACM SIGCHI 94* (1994), ACM Press, pp. 318–322.

[REM*12] Rosenbaum R., Engel D., Mouradian J., Hagen H., Hamann B.:
Interpretation, interaction, and scalability for structural decomposition trees.
In *GRAPP/IVAPP* (2012), pp. 636–647.

[Ren02] Rencher A.:
Methods of Multivariate Analysis, 2nd ed.
Wiley, New York, 2002.

[RKEE05] Reid J. S., Koppmann R., Eck T. F., Eleuterio D. P.:
A review of biomass burning emissions part ii: intensive physical properties of biomass burning particles.
Atmospheric Chemistry and Physics 5, 3 (2005), 799–825.

[RS00a] Roweis S. T., Saul L. K.:
Nonlinear dimensionality reduction by locally linear embedding.
SCIENCE 290 (2000), 2323–2326.

[RS00b] Roweis S. T., Saul L. K.:
Nonlinear dimensionality reduction by locally linear embedding.
Science 290 (2000), 2323–2326.

[Sam69] Sammon J. W.:
A nonlinear mapping for data structure analysis.
IEEE Trans. Comput. 18, 5 (1969), 401–409.

[SHFP99] Song X.-H., Hopke P. K., Fergenson D. P., Prather K. A.:
Classification of single particles analyzed by atofms using an artificial neural network, art-2a.
Analytical Chemistry 71, 4 (1999), 860–865.

[Shn96] 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.

[SLY*09] Stump G., Lego S., Yukish M., Simpson T. W., Donndelinger J. A.:
Visual steering commands for trade space exploration: User-guided sampling with example.
Journal of Computing and Information Science in Engineering 9, 4 (2009), 044501.

[SSM98] Schölkopf B., Smola A. J., Müller K.-R.:
Nonlinear component analysis as a kernel eigenvalue problem.
Neural Computation 10, 5 (1998), 1299–1319.

[ST03] Silva V. D., Tenenbaum J. B.:
Global versus local methods in nonlinear dimensionality reduction.
In *Advances in Neural Information Processing Systems 15* (2003), MIT Press, pp. 705–712.

[STDS95] Spence B., Tweedie L., Dawkes H., Su H.:
Visualization for functional design.
In *INFOVIS '95: Proceedings of the 1995 IEEE Symposium on Information Visualization* (Washington, DC, USA, 1995), IEEE Computer Society, p. 4.

[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.

[SWH*06] Saul L. K., Weinberger K. Q., Ham J. H., Sha F., Lee D. D.:
Spectral methods for dimensionality reduction.
Semisupervised Learning. MIT Press: Cambridge, MA (2006).

[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.

[TMN03] Tejada E., Minghim R., Nonato L.:
On improved projection techniques to support visual exploration of multidimensional data sets.
Information Visualization 2, 4 (2003), 218–231.

[Tor58] Torgerson W.:
Theory and methods of scaling.
Wiley, 1958.

[Tou80] Toussaint G. T.:
The relative neighbourhood graph of a finite planar set.
Pattern Recognition 12 (1980), 261–268.

[TSL00a] Tenenbaum J., Silva V., Langford J.:
A global geometric framework for nonlinear dimensionality reduction.
Science 290, 5500 (2000), 2319–2323.

[TSL00b] Tenenbaum J., Silva V., Langford J.:
A global geometric framework for nonlinear dimensionality reduction.
Science 290, 5500 (2000), 2319–2323.

[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.

[WFR*10] Waser J., Fuchs R., Ribicic H., Schindler B., Blöschl G., Gröller E.:
World lines.
IEEE Transactions on Visualization and Computer Graphics 16, 6 (2010), 1458 –1467.

[WGK10] Ward M. O., Grinstein G., Keim D. A.:
Interactive Data Visualization: Foundations, Techniques, and Application.
A. K. Peters, Ltd, 2010.

[WJ95] Wand M. P., Jones M. C.:
Kernel Smoothing, vol. 60.
Chapman & Hall/CRC, 1995.

[WLT94] Ward M. O., LeBlanc J., Tipnis R.:
N-land: a graphical tool for exploring n-dimensional data.
In *CG194 Proc: Insight Through Computer Graphics* (1994), pp. 130–41.

[WR02] Walter J., Ritter H.:
On interactive visualization of high-dimensional data using the hyperbolic plane.
In *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD), 2002* (2002), pp. 123–131.

[WR10] Wilson K. W., Raj B.:
Spectrogram dimensionality reduction with independence constraints.
In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (2010), pp. 1938 –1941.

[WS06] Weinberger K. Q., Saul L. K.:
An introduction to nonlinear dimensionality reduction by maximum variance unfolding.
In *AAAI* (2006), AAAI Press.

[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.

[YLMW06] Yang T., Liu J., Mcmillan L., Wang W.:
A fast approximation to multidimensional scaling, by.
In *Proceedings of the ECCV Workshop on Computation Intensive Methods for Computer Vision (CIMCV* (2006).

[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.

[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).

[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.

[ZHP08] Zhao W., Hopke P. K., Prather K. A.:
Comparison of two cluster analysis methods using single particle mass spectra.
Atmospheric Environment 42, 5 (2008), 881–892.

[ZIC*06] Zelenyuk A., Imre D., Cai Y., Mueller K., Han Y., Imrich P.:
Spectraminer, an interactive data mining and visualization software for single particle mass spectroscopy: A laboratory test case.
International Journal of Mass Spectrometry 258, 1–3 (2006), 58 – 73.

[ZIN*08] Zelenyuk A., Imre D., Nam E. J., Han Y., Mueller K.:
Clustersculptor: Software for expert-steered classification of single particle mass spectra.
International Journal of Mass Spectrometry 275, 1-3 (2008), 1–10.

[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.

APPENDIX A

Zusammenfassung

Forscher und Analysten in modernen industriellen und akademischen Umgebungen werden mit einer gewaltigen Menge von multivariaten Daten konfrontiert. Obwohl es in den Bereichen Data Mining und Knowledge Discovery bedeutsame Entwicklungen gegeben hat, besteht immer noch Bedarf an verbesserten Visualisierungen und generischen Lösungen. Der Stand der Forschung in visueller Datenanalyse und explorativer Datenvisualisierung ist es, neue Analysemethoden zu integrieren und Interaktion zu verbessern.

Im Forschungsfeld der explorativen Datenvisualisierung beschreibt diese Dissertation neue Ansätze in der Dimensionsreduktion, die eine Reihe von Probleme beheben, die aktuellen Methoden innewohnen, wie etwa die der Mehrdeutigkeit von Projektionen. Es wird beschrieben, wie Koordinatenwerte innerhalb einer Projektion eingebettet werden können, wodurch Mehrdeutigkeiten behoben werden. Außerdem wird behandelt, wie die Visualisierung der strukturellen Zusammensetzung von hochdimensionalen Daten eine intuitive Darstellung der inhärenten Beziehungen vermittelt, oder wie Eigenschaften und Ausrichtung von mehrdimensionalen Mannigfaltigkeiten in verschiedenen Detailstufen visualisiert werden können. Letzteres wird mittels einer rekursiven Einbettung lokaler Projektionen realisiert, die jeweils mit vollen Freiheitsgraden das zugehörige Detaillevel optimal anzeigen, während sie im globalen Kontext ausgerichtet sind.

Im Anwendungsfeld der Luftqualitätsforschung, stellt die Dissertation neue Methoden auf, die der Berechnung der Zusammensetzung von Luftpartikeln aus Massenspektrometrien dienen. Die Forschung an diesem Anwendungsproblem führte zu neuer Grundlagenforschung im Bereich der Visualisierung. Unsere Lösung umfasst eine visuelle und integrative Methodik, bei der Anwender Berechnungen mitverfolgen können, um diese hinsichtlich ihrer Semantik zu überprüfen, und, falls notwendig, steuernd auf den Berechnungsprozess einwirken können, wodurch weitreichende Ergebnisse im Anwendungsfeld erzielt wurden. Sorgfältige Reflektion unserer Lösung hat dazu geführt, allgemeine Design-Entscheidungen aufzustellen, die der konkreten Lösung zu Grunde liegen und auf weitere Anwendungsfelder übertragbar sind. Als Ergebnis beschreiben wir eine allgemeine Methodik zur Berechnung von Parametern, die sich auf Basis eines vordefinierten Analysemodells anhand gegebener multivariater Daten ableiten lassen. Anwendungsgebiete, die von unserer Methodik profitieren, sind Ingenieursdisziplinen und Naturwissenschaften.

APPENDIX B

Lebenslauf

Wissenschaftlicher Werdegang

Studium

04/04 - 09/10	Studium Diplom-Informatik an der Technischen Universität Kaiserslautern

Promotion

10/10 - 09/13	Stipendiat im internationalen Graduiertenkolleg (IRTG) 1131 der DFG, Betreuer: Prof. Dr. Hans Hagen
10/13 -	Wissenschaftlicher Mitarbeiter im regionalen Hochschulrechenzentrum (RHRK) Kaiserslautern

APPENDIX C

Schriftenverzeichnis

Journal publications

1. Engel, D., Rosenbaum, R., Hamann, B., Hagen, H.. Structural Decomposition Trees. Computer Graphics Forum, vol. 30, no. 3, pp. 921–930, 2011.
2. Engel, D., Petsch, S., Guhathakurta, S., Hagen, H.. Neighborhood Relation Diagrams for local comparison of carbon footprints in urban planning. Information Visualization, vol. 11, no. 2, pp. 124-135, 2012.
3. Engel, D., Huettenberger, L. Hamann, B.. A survey of dimension reduction methods for high-dimensional data analysis and visualization. In: Garth, C., Middel, A. and Hagen, H., eds., Visualization of Large and Unstructured Data Sets - Applications in Geospatial Planning, Modeling and Engineering, OpenAccess Series in Informatics (OASIcs), Schloss Dagstuhl, vol. 27, pp. 135-149, 2012.
4. Engel, D., Greff, K., Garth, C., Bein, K., Wexler, A., Hamann, B., Hagen, H.. Visual Steering and Verification of Mass Spectrometry Data Factorization in Air of Quality Research. IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 12, pp. 2275-2284, 2012.
5. Engel, D., Hagen, H., Hamann, B. and Rosenbaum, R.. Structural decomposition trees: Semantic and practical implications. Computer Vision, Imaging and Computer Graphics - Theory and Applications, Communications in Computer and Information Science (CCIS) Series, vol. 359, pp. 193-208, 2013.
6. Engel, D., Hummel, M., Hoepel, F., Bein, K., Wexler, Garth, C., A., Hamann, B., Hagen, H.. Towards High-dimensional Data Analysis in Air Quality Research. Computer Graphics Forum, vol. 32, no. 3, pp. 101-110, 2013.

Conference papers (peer-reviewed)

7. Petsch, S., Guhathakurta, S., Engel, D., Höpel, F., Hagen, H.. Visualizing the relationship between Urban Sprawl and the increasing Carbon Footprint in Maricopa County, AZ. In Proceedings of World Planning School Congress (WPSC), Perth, Australia, 2011.
8. Rosenbaum, R., Engel, D., Mouradian, J., Hagen, H., Hamann, B.. Interpretation, Interaction, and Scalability for Structural Decomposition Trees. in: Kraus, M. and Laramee, eds., Proceedings of International Conference on Information Visualization Theory and Applications (IVAPP), SciTePress Digital Library, pp. 636-647, 2012.
9. Engel, D., Khan, T.. User-aware Design Methodology for Data Visualization. Extended abstract. In: de Greef, T., Flint, T., eds., Workshop: Visualization – Beauty or The Beast, European Conference on Cognitive Ergonomics (ECCE), 2012.
10. Engel, D., Gillmann, C., Fiedler, S., Keller, S., Scheler, I., Hagen, H., Garth, C.. Visual Analysis of Circular Dichroism Spectra in Molecular Biophysics. Extended abstract and poster. In: Jankun-Kelly, T.J., Andrienko, G., Maciejewski, R., Tory, M., Lee, B., Leitte, H., eds., posters at the IEEE conference on visualization (IEEE VIS), 2013.
11. Haeb, K., Schweitzer, S., Fernández, D., Hagen, E., Engel, D., Böttinger, M., Scheler, I.. Visualization of Building Performance Simulation Results: State-of-the-Art and Future Directions. IEEE PacificVis Visualization Notes Short Paper Track, 2014.

Submissions in progress

12. Engel, D., Karer, B., Hagen, H., Scheler, I.. Manifold Learning and Projection of High-Dimensional Data using Graph Abstraction. Submission in progress: IEEE Transactions on Visualization and Computer Graphics (TVCG), 2014.
13. Engel, D., Hamann, B., Hagen, H., Scheler, I.. Generalizing Model-based Visual Analysis of High-dimensional Data, Submission in progress: IEEE Transactions on Visualization and Computer Graphics (TVCG), 2014.