

Linköping Studies in Science and Technology

Dissertations, No.1400

# **Algorithmically Guided Information Visualization**

Explorative Approaches for High Dimensional,  
Mixed and Categorical Data

**Sara Johansson Fernstad**



**Linköping University**  
**INSTITUTE OF TECHNOLOGY**

Department of Science and Technology  
Linköping University

Norrköping 2011

**Algorithmically Guided Information Visualization:  
Explorative Approaches for High Dimensional, Mixed and Categorical Data**

Copyright © 2011 Sara Johansson Fernstad unless otherwise noted  
sara.johansson@itn.liu.se

Department of Science and Technology, Linköping University  
SE-601 74 Norrköping

ISBN 978-91-7393-056-7  
ISSN 0345-7524

This thesis is available online through Linköping University Electronic Press:  
[www.ep.liu.se](http://www.ep.liu.se)

Printed by LiU-Tryck, Linköping 2011

# ABSTRACT

Facilitated by the technological advances of the last decades, increasing amounts of complex data are being collected within fields such as biology, chemistry and social sciences. The major challenge today is not to gather data, but to extract useful information and gain insights from it. Information visualization provides methods for visual analysis of complex data but, as the amounts of gathered data increase, the challenges of visual analysis become more complex.

This thesis presents work utilizing algorithmically extracted patterns as guidance during interactive data exploration processes, employing information visualization techniques. It provides efficient analysis by taking advantage of fast pattern identification techniques as well as making use of the domain expertise of the analyst. In particular, the presented research is concerned with the issues of analysing categorical data, where the values are names without any inherent order or distance; mixed data, including a combination of categorical and numerical data; and high dimensional data, including hundreds or even thousands of variables.

The contributions of the thesis include a quantification method, assigning numerical values to categorical data, which utilizes an automated method to define category similarities based on underlying data structures, and integrates relationships within numerical variables into the quantification when dealing with mixed data sets. The quantification is incorporated in an interactive analysis pipeline where it provides suggestions for numerical representations, which may interactively be adjusted by the analyst. The interactive quantification enables exploration using commonly available visualization methods for numerical data. Within the context of categorical data analysis, this thesis also contributes the first user study evaluating the performance of what are currently the two main visualization approaches for categorical data analysis.

Furthermore, this thesis contributes two dimensionality reduction approaches, which aim at preserving structure while reducing dimensionality, and provide flexible and user-controlled dimensionality reduction. Through algorithmic quality metric analysis, where each metric represents a structure of interest, potentially interesting variables are extracted from the high dimensional data. The automatically identified structures are visually displayed, using various visualization methods, and act as guidance in the selection of interesting variable subsets for further analysis. The visual representations furthermore provide overview of structures within the high dimensional data set and may, through this, aid in focusing subsequent analysis, as well as enabling interactive exploration of the full high dimensional data set and selected variable subsets. The thesis also contributes the application of algorithmically guided approaches for high dimensional data exploration in the rapidly growing field of microbiology, through the design and development of a quality-guided interactive system in collaboration with microbiologists.



# POPULÄRVETENSKAPLIG SAMMANFATTNING

## ALGORITMISKT VÄGLEDD INFORMATIONSVISUALISERING FÖR HÖGDIMENSIONELL OCH KATEGORISK DATA

Den tekniska utvecklingen under de senaste årtiondena har lett till att mer och mer data samlas in i en rad olika forskningsområden, som exempelvis biologi, kemi, medicin och samhällsvetenskap. Idag är inte den största utmaningen att samla in data, utan att förstå och ta fram användbar information ur komplexa och stora mängder data. Informationsvisualisering är ett forskningsområde som utvecklar metoder för att visuellt analysera komplexa data, men med ökande mängder data blir även den visuella analysen mer komplex och kräver därmed mer avancerade analysmetoder.

I den här avhandlingen presenteras forskning där algoritmiska analysmetoder används tillsammans med informationsvisualiseringsmetoder. Med hjälp av algoritmiska metoder kan mönster och strukturer i data snabbt identifieras. Genom att kombinera detta med interaktiva visualiseringsmetoder kan även analytikerns domänexpertis utnyttjas. Avhandlingen fokuserar framförallt på problem som är kopplade till visuell analys av: 1) kategorisk data, vilket är data där värdena är namn och där det inte finns någon naturlig ordning eller avstånd mellan datavärden; 2) analys av data som består av en kombination av kategoriska och numeriska värden; och 3) högdimensionell data som kan innehålla hundratals, eller till och med tusentals, dimensioner.

Till att börja med presenterar avhandlingen en metod för kvantifiering av kategorisk data, vilket innebär att kategorisk data representeras av numeriska värden. Med hjälp av en algoritmisk analysmetod så definierar kvantifieringsmetoden avstånd mellan kategorier, baserat på underliggande datastrukturer. När datasetet består av en kombination av kategorisk och numerisk data baseras kvantifieringen även på strukturer i den numeriska delen av datan. Kvantifieringsmetoden är en del i en interaktiv analysprocess, där resultaten från kvantifieringen fungerar som förslag på numeriska värden som kan användas istället för de kategoriska värdena. De föreslagna numeriska värdena kan enkelt och interaktivt ändras av analytikern. Genom att använda denna typ av kvantifieringsprocess kan kategorisk data analyseras med visualiseringsmetoder för numerisk data. Dessa är mer generella och mer vanligt förekommande än metoder för kategorisk data. Inom informationsvisualisering finns det idag två huvudmetoder när man analyserar kategorisk data. Den ena metoden är att använda visualiseringsmetoder som är utvecklade speciellt för kategorisk data. Den andra metoden är att representera kategorierna med numeriska värden och

sedan använda visualiseringsmetoder för numerisk data. Kvantifieringsprocessen som presenteras i den här avhandlingen tillhör den senare typen. Avhandlingen presenterar även den första användbarhetsutvärderingen där dessa två metoder jämförs med varandra.

I avhandlingen beskrivs dessutom två metoder för att minska antalet dimensioner i högdimensionella dataset. Metodernas syfte är att behålla flera typer av mönster samtidigt då dimensionaliteten minskas. Syftet är också att metoderna ska vara flexibla och helt och hållet kontrolleras av användaren. Med hjälp av en automatisk analys baserad på kvalitetsmått, där varje kvalitetsmått representerar ett, för analytikern, intressant mönster, identifieras dimensioner som verkar särskilt intressanta. Med hjälp av olika visualiseringsmetoder får analytikern ta del av de mönster som identifierats under analysen. På så vis fungerar den automatiska analysen som vägledning för att välja ut ett mindre antal intressanta dimensioner för mer ingående analys. Visualiseringen ger även en överblick av de mönster som finns i det högdimensionella datasetet och kan genom detta fungera som ett hjälpmedel för att besluta vad den fortsatta analysen ska fokusera på. Dessa metoder gör också att analytikern interaktivt kan utforska både det högdimensionella datasetet och de dimensioner som bedömts som särskilt intressanta. Avhandlingen visar även hur algoritmiskt väglett informationsvisualisering kan användas inom mikrobiologi, som är ett snabbt växande forskningsområde. Detta visas genom att presentera ett interaktivt system för analys av högdimensionell mikrobiell data, vilket utvecklats i nära samarbete med mikrobiologer.

De metoder som utvecklats inom ramen för avhandlingen underlättar effektiv dataanalys genom att använda snabba algoritmiska metoder för att identifiera intressanta mönster. Genom att kombinera dessa metoder med interaktiva visualiseringsmetoder blir det möjligt att även dra nytta av användarens domänexpertis. Detta möjliggör en flexibel analys som kan underlätta identifierandet av användbar information ur komplexa data inom en mängd områden.

# ACKNOWLEDGEMENTS

My first thanks go to my three supervisors. To Mikael Jern for introducing me to the field of information visualization and for making this PhD possible. To Jimmy Johansson for indispensable guidance, feedback and support throughout the years, and for always being a good friend. To Jane Shaw at Unilever R&D, Port Sunlight, UK, for inspiration, support and enthusiasm, and for making our collaboration a fantastic experience. I would also like to thank Matt Cooper for feedback, support and numerous proof-readings throughout, not least during thesis writing.

Furthermore I would like to thank everyone I have met at Unilever R&D, Port Sunlight, UK, for enthusiastic and encouraging feedback and support throughout my PhD studies, and for putting data analysis into interesting and relevant contexts. Particularly I would like to thank Rob Treloar for your enthusiasm and for all interesting questions, Tim Madden for all encouraging feedback and for bringing up relevant questions which have driven the work forward and, not least, David Taylor, Suzi Adams and Sally Grimshaw for all your positive and valuable feedback and for a great collaboration on the microbiomics project. I would also like to thank Phil Helme, Stephen Bennett, Trevor Cox and everyone else that have enthusiastically shown interest in my work throughout these years.

Many thanks also go to my colleagues in the C-Research/Media and Information Technology group, Linköping University, for feedback and support and for making these years a fun experience. Special thanks go to Camilla Forsell for user-evaluation support, to Karljohan Palmerius for providing the L<sup>A</sup>T<sub>E</sub>X template which this thesis is based on and, not least, to Eva Skärblom for indispensable help with all practical issues throughout.

Finally, my special thanks go to my family, for always supporting me and believing in me, and to my loving husband Marcus, for being my best friend and for making me laugh at least once a day.



This work was supported in part by a ‘heavy’ grant under the Swedish Knowledge Foundation’s Visualization Programme.



# CONTENTS

- 1 Introduction 3**
  - 1.1 Multivariate Data . . . . . 4
  - 1.2 Information Visualization . . . . . 6
  - 1.3 Data Mining . . . . . 9
  - 1.4 Algorithmically Guided Visualization . . . . . 11
  - 1.5 Research Challenges . . . . . 13
  - 1.6 Overview of Papers . . . . . 15
  
- 2 Background 17**
  - 2.1 Categorical Data . . . . . 17
    - 2.1.1 Visualization . . . . . 18
    - 2.1.2 Quantification . . . . . 22
  - 2.2 High Dimensional Data . . . . . 27
    - 2.2.1 Visualization . . . . . 27
    - 2.2.2 Dimensionality Reduction and Ordering . . . . . 28
  
- 3 Contributions 35**
  - 3.1 Interactive Quantification of Categorical Data . . . . . 35
    - 3.1.1 Objective . . . . . 36
    - 3.1.2 Result . . . . . 36
    - 3.1.3 Summary of Contributions . . . . . 40
  - 3.2 Evaluation of Categorical Data Approaches . . . . . 41
    - 3.2.1 Objective . . . . . 41
    - 3.2.2 Result . . . . . 42
    - 3.2.3 Summary of Contributions . . . . . 44
  - 3.3 Quality Based Dimensionality Reduction . . . . . 45
    - 3.3.1 Objective . . . . . 46
    - 3.3.2 Result . . . . . 46
    - 3.3.3 Summary of Contributions . . . . . 49
  - 3.4 Visual Exploration of High Dimensional Data . . . . . 49
    - 3.4.1 Objective . . . . . 50
    - 3.4.2 Result . . . . . 51
      - 3.4.2.1 Exploration of Microbial Populations . . . . . 53

3.4.2.2	A Generic Approach for High Dimensional Data Exploration .	56
3.4.3	Summary of Contributions . . . . .	57
<b>4</b>	<b>Conclusions</b>	<b>59</b>
4.1	Summary of Contributions . . . . .	59
4.2	Discussion . . . . .	60
4.3	Future Work . . . . .	61
	<b>Bibliography</b>	<b>65</b>
	<b>Publications Included in the Thesis</b>	<b>73</b>
<b>I</b>	Interactive Quantification of Categorical Variables in Mixed Data Sets	<b>73</b>
<b>II</b>	Visual Exploration of Categorical and Mixed Data Sets	<b>83</b>
<b>III</b>	Visual Analysis of Mixed Data Sets Using Interactive Quantification	<b>95</b>
<b>IV</b>	A Task Based Performance Evaluation of Visualization Approaches for Categorical Data Analysis	<b>107</b>
<b>V</b>	Interactive Dimensionality Reduction Through User-defined Combinations of Quality Metrics	<b>119</b>
<b>VI</b>	Visual Exploration of Microbial Populations	<b>129</b>
<b>VII</b>	Quality Based Guidance for Exploratory Dimensionality Reduction	<b>139</b>

# COMPLETE LIST OF PUBLICATIONS

Papers I through VII are included in the thesis and are cited by their respective number throughout. Papers VIII through XI are in part related to the work presented in the thesis, but are not included.

- I** Sara Johansson, Mikael Jern and Jimmy Johansson. Interactive Quantification of Categorical Variables in Mixed Data Sets. *In proceedings of IEEE International Conference on Information Visualisation*, pages 3–10, London, UK, July 9–11, 2008.
- II** Sara Johansson. Visual Exploration of Categorical and Mixed Data Sets. *In proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Workshop on Visual Analytics and Knowledge Discovery*, pages 21–29, Paris, France, June 28, 2009.
- III** Sara Johansson and Jimmy Johansson. Visual Analysis of Mixed Data Sets Using Interactive Quantification. *ACM SIGKDD Explorations*, 11(2):29–38, Dec. 2009.
- IV** Sara Johansson Fernstad and Jimmy Johansson. A Task Based Performance Evaluation of Visualization Approaches for Categorical Data Analysis. *In proceedings of IEEE International Conference on Information Visualisation*, pages 80–89, London, UK, July 13–15, 2011.
- V** Sara Johansson and Jimmy Johansson. Interactive Dimensionality Reduction Through User-defined Combinations of Quality Metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):993–1000 Nov./Dec. 2009.
- VI** Sara Johansson Fernstad, Jimmy Johansson, Suzi Adams, Jane Shaw and David Taylor. Visual Exploration of Microbial Populations. *In proceedings of 1st IEEE Symposium on Biological Data Visualization*, Providence, RI, USA, October 23–24, 2011.
- VII** Sara Johansson Fernstad, Jane Shaw and Jimmy Johansson. Quality Based Guidance for Exploratory Dimensionality Reduction. *Submitted to Information Visualization*, 2011.
- VIII** Sara Johansson, Kristina Knaving, Amanda Lane, Mikael Jern and Jimmy Johansson. Interactive Exploration of Ingredient Mixtures Using Multiple Coordinated Views. *In proceedings of IEEE International Conference on Information Visualisation*, pages 210–218, Barcelona, Spain, July 15–17, 2009.

- IX** Mikael Jern, Tobias Åström and Sara Johansson. GeoAnalytics Tools Applied to Large Geospatial Datasets. *In proceedings of IEEE International Conference on Information Visualisation*, pages 362–370, London, UK, July 9–11, 2008.
- X** Sara Johansson and Mikael Jern. GeoAnalytics Visual Inquiry and Filtering Tools in Parallel Coordinates Plots. *In proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems*, pages 252–259, Seattle, WA, USA, November 7–9, 2007.
- XI** Mikael Jern, Sara Johansson, Jimmy Johansson and Johan Franzén. The GAV Toolkit for Multiple Linked Views. *In proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 85–97, Zürich, Switzerland, July 2, 2007.

# CHAPTER 1

## INTRODUCTION

The technological advances of the last decades have led to a rapid increase in the amount of data being collected and stored within a variety of fields, such as DNA-sequencing, simulation, financial and climate research. As a result, the major challenge of today is not the collection of data, but the extraction of useful knowledge and information from the vast amounts of complex data that is available. Although technological advances have multiplied our ability to collect and store data, it is still the insights gained from data that are important, and the gaining of insights is greatly constrained by the limitations of the human perceptual system [23, 87]. As a result, efficient techniques for analysis, exploration and identification of interesting structures in data are highly desirable and often necessary to extract useful information and through it gain relevant insights. Visualization and data mining are two fields addressing the issue of extracting and identifying useful information from large and complex data sets, but from different perspectives. Thus, the combination of the two may often be beneficial.

To visualize something means to construct a mental image of it, and thus visualization in its actual meaning has nothing to do with computers but is a purely cognitive activity [79]. Today the term visualization has, however, more and more come to refer to graphical representations of data which are used to support decision making [87] and some define visualization as "*The use of computer-supported, interactive, visual representations of data to amplify cognition*" [14]. Using the latter definition, visualization can be said to act as a link between the human mind and the computer. The use of a visual representation, such as a diagram, facilitates the creation of a mental image of structures within complex data and is, as such, an important part in the process of gaining relevant insights from data. The usability of visual representations is, however, often decreased as the size and complexity of data sets increase, limited both by display size and resolution as well as by the human visual system and the visual metaphors used [23].

Data mining on the other hand can be defined as an automated process of identification and discovery of useful structures in data [25] or as the process of extracting, or 'mining', interesting knowledge from large amounts of data stored in information repositories [40]. Data mining methods may often be successfully combined with visualization approaches, as both focus on the detection of interesting and useful structures and information in data. Due to differences in approaches and methodology their benefits and limitations are, however, different, and as such they may often complement and enhance each other in ways advantageous to the data analyst.

This chapter will provide background on concepts and details of visualization and data mining which are relevant to understanding the contributions of this thesis. It will also describe various aspects of multivariate data and briefly present some of the fundamental issues of visual analysis of multivariate data, as well as providing an introduction to the research challenges motivating the work presented in the thesis.

## 1.1 MULTIVARIATE DATA

A data set can be defined as a collection of data items [81], where an item may, for instance, represent a patient in a medical data base, a municipality in a collection of census data, a test run in a simulation or a sample in a DNA-sequencing study. Various terminologies are used within the literature and data items are sometimes also referred to as objects, records, tuples, points, vectors, observations or samples. Throughout this thesis and the publications included the terms data item and data object are used interchangeably and, in some cases, when referring to biological data, the term sample is also used.

The characteristics of a data item are described by a collection of variables. A variable can be defined as a property, or a characteristic, of a data item that may vary from one item to another or over time [81]. As an example, in a census data set where the data items represent municipalities, the variables may represent various characteristics of the municipalities such as population size, average income, percentage unemployed and the percentage of the population that voted for a specific political party. Variables are sometimes also referred to as attributes, dimensions and features. In this thesis and in the publications included, the terms variable and dimension are used interchangeably.

A multivariate data set is simply a data set including two or more variables. The items of a multivariate data set can be thought of as points in a multidimensional space where each dimension represents a variable. A standard format used for structuring multivariate data is to use an  $m$ -by- $n$  matrix including  $m$  rows, usually representing data items, and  $n$  columns, usually representing variables. Table 1.1 displays a small example of a data matrix including three data items, each representing a film, and four variables, each describing a characteristic of the film.

In terms of data set size, a data set is, in this thesis, referred to as large when it contains either too many variables or too many data items for a chosen visualization method to be usable. Based on this, a moderately sized data set is defined as a data set including a number of variables and data items for which the chosen visualization method is usable. In this context, a visualization method is defined as usable if it, in reasonable time, convey information about data in a clear and interpretable way.

Data variables are often classified into different types, and various classification taxonomies appear in visualization and data mining literature. Tan et al. [81] use the properties of distinctness, order, addition and multiplication to describe and classify variables. They initially separate variables into categorical (qualitative) and numerical (quantitative). Categorical data is then defined as either nominal, where the data values are different names which only provide enough information to distinguish one value from another, or ordinal, where the values provide enough information to order the items. The *Director* variable in table 1.1 is an example of a nominal

Table 1.1: An example of a multidimensional data set including three data items, each representing a film, and four variables of different types, each representing a property of the film.

Director	Rating (UK)	Length (min)	Production year
Ingmar Bergman	A	91	1957
Robert Aldrich	X	134	1962
Alfred Hitchcock	A	128	1958

variable, whereas the *Rating* variable, for which items can take the values of *U* (suitable for children), *A* (children must be accompanied by an adult) and *X* (suitable only for adults), is an ordinal variable. Numerical data are represented by numbers and can be either continuous or integer values. Tan et al. [81] separate numerical data into interval data, for which the differences but not ratios between values are meaningful, and ratio data, where both differences and ratios between values have a meaning. In table 1.1 *Production year* represents an interval variable whereas *Length* is a ratio variable.

Card et al. [14] use a similar but slightly simpler classification by dividing variables into three different types; nominal, which is an unordered set of values, ordinal, which includes an ordered set of values, and quantitative, which represents a numeric range to which arithmetic can be applied. Using this classification the *Director* and *Rating* variables would again be classified as nominal and ordinal respectively, while *Production year* and *Length* would both belong to the same class of quantitative variables. Hand et al. [41] use a comparable classification of quantitative and categorical variables. A slightly different way of classifying data is to define it as either discrete or continuous, where discrete variables include a finite set of values which can be either categorical or numerical (usually integer values), and continuous variables include values that are real numbers [81]. In table 1.1 *Director*, *Rating* and *Production year* are all discrete variables while *Length* can be defined as continuous. Friendly [32] classifies categorical variables into three different types; binary, nominal and ordinal, where binary variables only can take two different values, such as *true* or *false*. The number of distinct values that can be taken by a categorical datum is defined as the cardinality of the variable.

The definition of variable types for a data set is important since various data types may require different analytic techniques since a single technique is rarely appropriate for all types of data. Specifically, techniques used for numerical variables are often based on a numerical difference or similarity between data items. Categorical data on the other hand does not include any distance measure comparable to a numerical distance, and hence other analytical techniques may need to be used. As a simple example based on the data set in table 1.1, the length of a film may quite effectively be represented using a bar chart where the height of the bar represents the length of the film, as displayed in figure 1.1. Using the same kind of representation for *Director* would, on the other hand, not be as useful since there does not exist any meaningful relationship between a director name and the height of a bar. Furthermore, the similarity of two categories may often depend upon context and is, hence, not as generalizable as numerical similarity. As an example, figure 1.2 displays three colours; dark green, light green and dark purple, which may be the

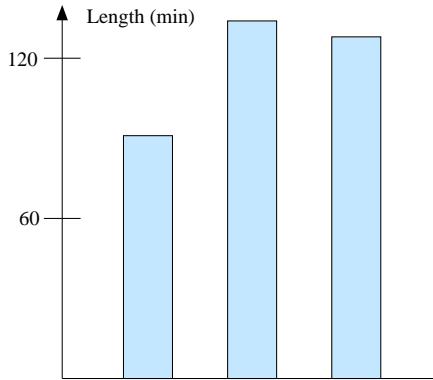


Figure 1.1: A bar chart used to display the length of the small example data set of table 1.1.



Figure 1.2: An example of the context dependency of similarity between categorical values. In the context of colour hue, the dark green and light green colours are most similar, whilst the dark green and dark purple are most similar in the context of lightness.

categories of a categorical variable. In the context of hue the two green colours are more similar to each other than to the purple colour, but in the context of lightness dark green and dark purple are more similar to each other than they are to light green.

Throughout this thesis data type definitions corresponding to the two basic classes of categorical and numerical data are used, where categorical variables may be either nominal or ordinal, and numerical variables are defined similar to the quantitative variables of Card et al. [14]. Additionally the concept of mixed data sets is also used, referring to a data set including both categorical and numerical variables.

## 1.2 INFORMATION VISUALIZATION

Visualization is often separated into scientific visualization and information visualization, the latter being the focus of the visualization research presented in this thesis. There is not always a clear boundary between scientific visualization and information visualization, since methods from both fields may be combined and used concurrently within many areas. Commonly the two fields are separated based on the properties of the data that are analysed and on how the data is represented. Scientific visualization tends to deal with physically based data and the visual representations used usually relate to the physical properties of the data [79]. A typical application of scientific visualization may, for instance, be to use visual representations of molecules, including atoms and chemical bonds, for analysis of molecular data.

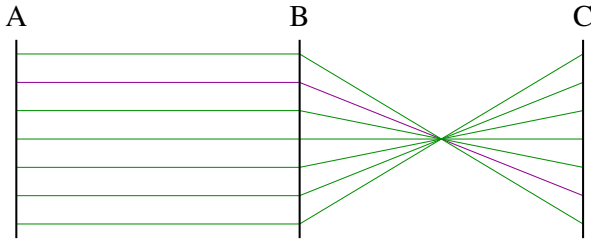


Figure 1.3: Parallel coordinates displaying a data set with three variables, represented by vertical axes, and seven items, represented by polylines.

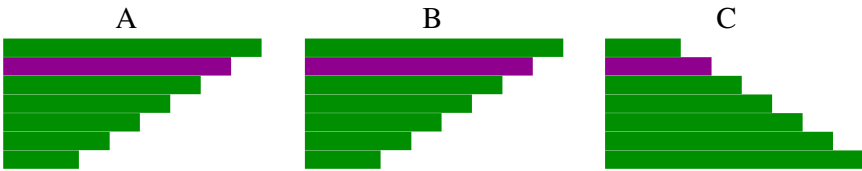


Figure 1.4: A table lens representing a data set with three variables, represented as columns, and seven items, represented as rows.

Information visualization, on the other hand, usually deals with more abstract data without any obvious spatial mapping. Analysis of these data would rarely require or benefit from a visual representation of the actual physical object, and hence information visualization is mostly concerned with abstract concepts and representations. As an example, analysis of a film data set would hardly benefit from visual representations of celluloid films. The principal task of information visualization is to represent abstract data in a way such that information, in terms of useful and interesting structures, may be derived from it. *"The holy grail of information visualization is to make the insights stand out from otherwise chaotic and noisy data."* [16]

Information visualization provides a range of visualization methods for analysis of multivariate data. Of these, parallel coordinates, scatter plot matrix and table lens, are commonly used within this thesis. Using **parallel coordinates** [47, 88] (figure 1.3) the variables of a multivariate data set are mapped to parallel axes. Data items are represented by polylines intersecting the axes at the values of the variable. The **table lens** [68] (figure 1.4) displays each variable as a column and each data item as a row, with values represented by line length. A **scatter plot matrix** [8] (figure 1.5) is a matrix including a set of two-dimensional scatter plots. Often the set of scatter plots includes plots for all variable pairs in a multivariate data set and is then usually structured as a symmetrical matrix. The example data displayed in figures 1.3, 1.4 and 1.5 includes three variables (A, B, C) and seven data items. One data item is highlighted in purple in all figures. Some relationships between variables can be identified from the figures, including a positive correlation between variables A and B (meaning that high values in one variable correspond to high values in the other variable as well), and a negative correlation between variables B and C (meaning that high values in one variable correspond to low values in the other variable). In the

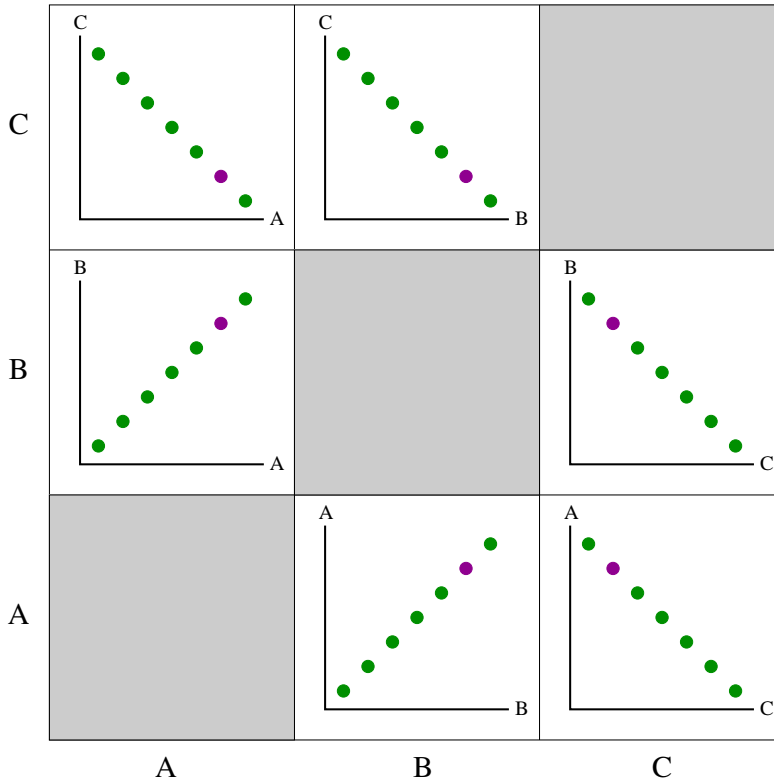


Figure 1.5: A scatter plot matrix representing a data set with three variables and seven items. The cells of the matrix include all two-dimensional scatter plots that can be formed from the data set.

scatter plot matrix, displaying all variable pairs, a negative correlation between variable A and C is also visible.

Today the use of multiple coordinated views is a well established concept in information visualization [6, 13, 69]. Put simply, the concept of multiple views implies that several windows or views are used concurrently for representing data. Usually different visual representations are used within different views displaying different aspects of data. To enable integration of the separate views operations between views are coordinated, meaning that any changes made in one view are reflected in the other views. The use of multiple coordinated views allows the user to analyse and explore data from different perspectives, and since different visual representations have different limitations and enhance different aspects, this may often facilitate the extraction of useful information and the gaining of insights from data. The concept of multiple coordinated views has been a design base for the visualization systems that have been developed as part of this thesis. Various approaches can be taken as to how the views are coordinated. Approaches relevant for the work presented in this thesis include Overview+Detail [14], where one view

presents an overview of the whole data set while other views only display some details of the data, and Master/Slave Relationship [69], where the representation in one view acts as a control over the other views.

Interaction is a core part of multiple coordinated views and of information visualization in general. A large variety of interaction techniques may be used within a system. These can be defined as either indirect, manipulating the data using buttons, menus and sliders which are separated from the visual representations, or direct, which indicates that manipulation of data is performed directly in the visual representation itself. The dynamic querying concept, as introduced by Shneiderman [77], provides an example of indirect manipulation where sliders and buttons are included in a graphical user interface and used as an interactive and visual alternative to standard database queries. A typical example of direct manipulation is the concept of brushing [8] where objects, such as points in a scatter plot, are selected directly in the display using a 'brush'. The selected objects are then manipulated through a brushing operation, such as highlighting or labelling. In the context of multiple coordinated views the brushing operation is usually propagated to all views.

As a natural result of the development and growing maturity of the field of information visualization, visualization methods will become more commonly available and used in data analysis. Based on this an important research aspect in information visualization is to establish the usability of new visualization methods. Through user studies the usability of a visualization may be assessed. User studies can be carried out in different ways and various classifications of studies are available. For example, Plaisant [66] describes the following four main types of evaluation in information visualization:

1. **Controlled experiments comparing design elements**, where specific elements of visual representations, such as sliders or colours, are compared.
2. **Usability evaluation of a tool**, which is used to provide feedback on problems and limitations of a visualization tool, aiming to refine the design of the tool.
3. **Controlled experiments comparing two or more tools**, where the performance of multiple tools is compared.
4. **Case studies of tools in realistic settings**, which is carried out in the users' natural environment doing real tasks, and hence measures the usability of a tool within its intended context.

Furthermore, a user-centered design process, driven by feedback from the intended end-users may ensure the usefulness of the tool within the analytical context of the users.

## 1.3 DATA MINING

Data mining can be defined as the process of discovering or extracting useful information from large amounts of data through the use of automated methods. Another important aspect of data

mining, in addition to extracting information, is to summarize data and identified patterns in ways that are understandable and useful to the data analyst [41]. The field of data mining is closely linked to fields such as statistics, pattern recognition, machine learning and artificial intelligence [25], and it may sometimes be difficult to draw a clear boundary between them. For simplicity all algorithmic and automated methods used will be referred to as data mining methods throughout this thesis, although some methods may originate from other fields.

Some of the most common data mining techniques are clustering, correlation analysis, classification and dimensionality reduction. **Clustering** is the concept of identifying groups, or clusters, of data items that are similar to each other and different to the items of other clusters. **Correlation analysis** is the association of two or more variables, and the correlation coefficient extracted through correlation analysis can be seen as a measure of variable similarity. **Classification** is the concept of assigning data items to a set of predefined classes. It is partly related to discretization or categorization of continuous variables, which means to transform the continuous scale of a variable into a fixed set of categories. **Dimensionality reduction** is the concept of representing a larger number of variables with a smaller number. A large number of different computational methods are available for any of these techniques, some of which will be discussed in more detail later on in the thesis.

The basic tasks of data mining are often classified as either descriptive or predictive, where descriptive tasks aim to describe and summarize the general properties of the data set, whereas the goal of a predictive task is to predict values based on already known values and variables [40, 81]. Hand et al. [41] provide a more detailed classification of data mining tasks, as follows:

- **Exploratory Data Analysis** represents tasks focusing on exploration of data without any clear hypotheses or ideas of what to look for.
- **Descriptive Modelling** represents tasks where the aim is to describe the data set and its properties.
- **Predictive Modelling** represents tasks where the goal is to build a model that can be used for prediction of unknown values from already known values and variables.
- **Discovering Patterns and Rules** represents tasks where the main goal is to detect patterns within data.
- **Retrieval by Content** represents tasks aiming to identify patterns in a data set that are similar to an already known pattern of interest.

Most of the work presented in this thesis focuses on exploration of data for identification of interesting patterns and the forming of hypotheses. In some sense, large parts of the work, especially in terms of using algorithmic methods, can be defined as descriptive analysis, as it is used for describing patterns or relationships identified in the data. However, as a whole the descriptions of identified patterns are used as support or guidance for the analyst in an exploratory analysis process, a concept which will be described in more detail in the following section.

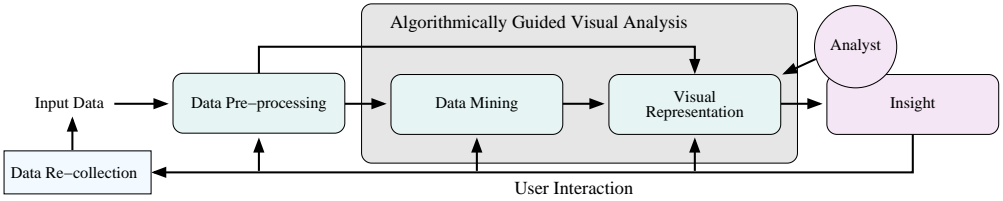


Figure 1.6: Description of the interactive and iterative data analysis process of converting raw data into useful insights.

## 1.4 ALGORITHMICALLY GUIDED VISUALIZATION

An ever increasing amount of data is gathered as new techniques for data collection are developed and as a result many scalability-related issues arise within data analysis. Data sets may typically involve a large number of variables and items and the task of visualizing this data in a usable way may be quite complex. While visual representations and interactive exploration facilitate the process of gaining relevant insights from data, humans are limited in terms of our ability to handle large volumes of data. Data mining can provide fast and automated methods for reducing and representing data, but the use of automated analysis methods distances the data exploration from the analyst. This may limit the analyst’s understanding of structures in the data and of the exploration processes leading to these structures [25]. Furthermore, the prior knowledge and domain expertise of the analyst is rarely taken into account in data mining algorithms [41]. Many of these problems may be overcome by combining automated data mining methods with interactive visualization techniques, which is the main approach of the contributions of this thesis.

The positive impact of applying data mining methods in combination with visualization was established by Wen and Zhou [90] through two user studies. Their results indicate an improved performance, in terms of task completion time and error rate, for both single-step and multiple-step tasks when combining visualization and data mining, the benefit being most significant when difficult tasks are performed or when complex visualization tools are used. Additionally, their study indicated that the interaction context of the analysis has a strong influence on the choice of data mining method to use. This implies that the selection of algorithmic transformation to apply to the data should preferably be decided dynamically within the users analytical context.

The overall process of converting data into useful and relevant insights is sometimes known as knowledge discovery in databases (KDD). Several descriptions of this analysis process, with varying complexity, are available in, for instance, Tan et al. [81], Card et al. [14] and Han and Kamber [40]. Figure 1.6 provides an overview of the process, here referred to as the data analysis process, in a manner relevant to the work described in this thesis. The figure describes an iterative process of converting raw data into insights useful to the analyst, which in turn may be used to refine and focus any of the previous steps in the process, or even lead to re-collection of data. The first step of the analysis process is the step of Data Pre-processing which may comprise activities such as combining several data sources into one data set and data cleaning including removal of noise and inconsistent data. The purpose of data pre-processing is to transform the

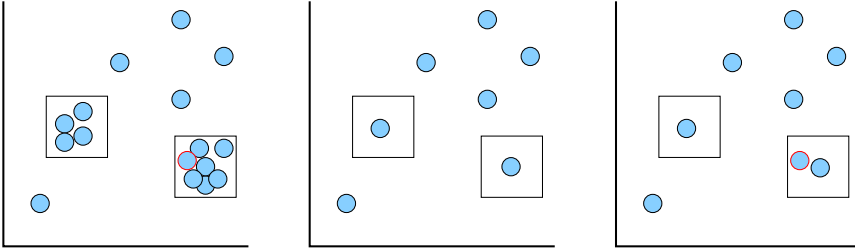


Figure 1.7: Three scatter plots clarifying the concept of algorithmic guidance. Left: The original data set where two groups of similar data items identified by an algorithm is surrounded by black borders. Centre: Result when the identified groups are automatically replaced by a representative data item. Right: The result when the analyst is provided the possibility of excluding from the merging an item known to be of particular interest.

data into a format appropriate for the subsequent analysis steps. This may be followed by the Data Mining step where algorithmic methods are used to identify structures and to reduce the data, for instance by employing dimensionality reduction. The Visual Representation step may follow the data pre-processing step directly or be subsequent to the data mining step. It includes utilizing interactive visual representations to display the data set, either original or manipulated through algorithmic methods, and to highlight potentially interesting structures identified in the data mining step. The analyst is usually very involved in the visualization step by interacting with the visual representations and visually exploring the data to gain insights. The data analysis process, as defined here, is a highly interactive and iterative process where the insights gained by the analyst may be used to further focus the analysis or to refine any of the previous steps, ranging from the collection of new data, to the use of different data mining methods and the employment of a different visual representation or brushing of a different data subset.

This thesis uses the notion of algorithmic guidance. The algorithmically guided visual analysis includes the data mining and visual representation steps of the analysis process. Utilizing data mining followed by visualization does not, however, necessarily assume algorithmically guided visual analysis as the concept is used here. Algorithmic guidance implies that data mining methods are used to aid the analyst in identification of potentially interesting patterns and in decision-making. Patterns identified during the data mining step of the process can be displayed to the analyst as suggestions of structures which may possibly be interesting to investigate further. While the analyst is still provided control regarding which analysis paths to follow and which manipulations should be applied to the data. This is the preferable way of combining data transformation and visualization according to the study of Wen and Zhou [90]. Thus, with this approach the judgement of what is actually interesting is made by the analyst during analysis; based on domain knowledge, analytical task and data structures. The analysis process where algorithmic guidance takes place is typically an explorative process conducted by a domain expert.

A simple example can be used to further highlight the difference between an algorithmically guided analysis process and a process where the exploration is controlled by an automated method. Figure 1.7 displays three scatter plots. The leftmost plot displays the original data set

where the data items, represented by circles, are somewhat cluttered. One way of dealing with cluttered data is to replace groups of similar items with one representative item. By utilizing a clustering technique, two groups of similar items are identified, as surrounded by black borders in the leftmost plot. Using a purely automated approach each of the identified groups is replaced by a representative item, as in the centre plot. However, in this specific example the analyst is aware that one of the data items, represented by a red border in the leftmost plot, is of particular interest due to some domain-relevant property. Using algorithmic guidance the merging of the identified clusters would be suggested to the analyst, and prior to replacing any of the groups the analyst may, for instance, be offered the possibility of excluding individual items from the groups as well as freely deciding whether or not a group should be replaced by a representative item. The rightmost plot in figure 1.7 displays the result when the item of interest was excluded prior to merging the clusters.

The core purpose of algorithmically guided visual analysis is to provide analysis processes which are guided but not constrained by automated methods; utilizing the speed of automated methods as well as making use of the knowledge of an expert analyst. The concept of algorithmically guided visual analysis is a core focus throughout the work described in this thesis and a common denominator of the included publications. The systems and approaches presented in the included publications all employ algorithmic guidance to some extent, although the level of user control varies between the papers.

## 1.5 RESEARCH CHALLENGES

Some of the main research challenges addressed in this thesis approach issues relating to the structure of data. As increasing amounts of data are gathered difficulties involved in analysis of large data sets become more obvious. It is not unusual to deal with data sets including hundreds of variables or hundreds of thousands of data items. Even so, most visualization methods are not designed for representing data of this size and additional methods, such as data mining methods, need to be employed. The scalability of visualization has been defined as one of the top challenges of information visualization today [17]. Papers V, VI and VII of this thesis address the issue of scalability in terms of high dimensionality. High dimensionality is not only an issue of importance in information visualization but is also one of the top challenges in data mining [81]. Most data analysis techniques have been designed for data sets of low or moderate dimensionality and do not work as well for high dimensional data. The issues involved relate not only to the increased computation time of algorithms, but also to visual clutter due to overcrowded displays. Figure 1.8 shows an example of this when parallel coordinates and a scatter plot matrix are used to display a data set including 227 variables. The issue of overcrowding may perhaps partly be solved through larger displays with higher resolution. However, the capacity of human perception is another limiting factor, both in terms of the precision of the eye and of the ability of the human mind to process visual patterns [87]. Large displays may quickly become impossible to overview and small patterns, sometimes as small as a single pixel, are often impossible to perceive in a high resolution display. Hence the issue of high dimensionality is not a purely technical issue which can be solved through faster computers or larger screens, but,

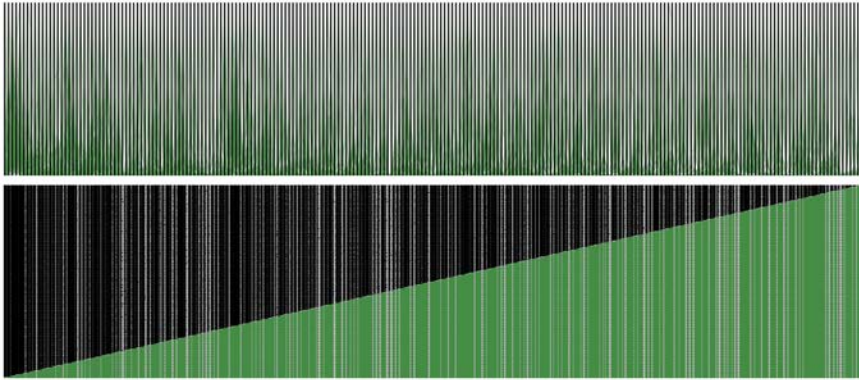


Figure 1.8: Example of overcrowded displays when parallel coordinates (top) and a scatter plot matrix (bottom) are used to represent a high dimensional data set including 227 variables.

perhaps most importantly, is a perceptually based problem which, as with all visualization, has to be considered in the context of the human user.

The structure of data is not only a question of size but also of the variable types included in the data set. Most visualization methods have been designed for numerical data, making use of the straightforward numerical similarity of data items in the layout. Nonetheless categorical data is commonly available and calls for specialized analysis methods, although the number of visualization methods designed for categorical data are far fewer than the number of methods for numerical data. The challenge of visualizing categorical data is hence an important issue within visualization which is addressed in papers I, II, III and IV in the thesis. An additional challenge related to that of categorical data analysis, is the analysis of mixed data sets including a combination of categorical and numerical data. Traditional data analysis methods, both within visualization and data mining, deal with data sets including only one type of variable. Due to the intrinsic differences between categorical and numerical data, data sets including a combination of both require special attention.

Utilizing automated methods for identification of interesting patterns enables fast analysis, but too much automation may distance the analyst from the analysis process and important knowledge of the analyst may not be taken into enough consideration. Interactive visualization systems, on the other hand, make use of the domain knowledge of the analyst, allowing the user to fully control the analysis process and open-endedly explore the data. This process may nevertheless be tedious and time consuming for large and complex data sets, and the task of explorative analysis may be overwhelming at the start. For the purpose of properly involving the analyst in the analysis process it is desirable to find an appropriate balance between employing automation and providing interactive visualization whilst combining the two approaches.

The concept of interestingness is addressed in parts of the thesis, where it refers to how interesting different structures in the data are. This could, for instance, be the interestingness of a cluster, a variable or a data subset, indicating the likelihood of the structure being of interest

to the analyst. For algorithmic guidance the notion of interestingness is of high importance, since a main goal of utilizing algorithms in this context is to attract attention to the structures that are most likely to be of interest. However, what actually is of interest to an analyst depends on domain, task and data; hence a difficulty while designing a data analysis system is to define interestingness appropriately. To summarize, the major research challenges addressed in this thesis are as follows:

- the research of usable methods for exploratory visual analysis of high dimensional data;
- the research of methods for interactive visual analysis of categorical and mixed data sets;
- the definition of interestingness measures appropriate to the analytical context;
- the combination of automated techniques and interactive visualization methods in a well balanced manner.

## 1.6 OVERVIEW OF PAPERS

This section presents a short overview of the publications included in the thesis. A more thorough description of their contributions can be found in chapter 3. The author of this thesis is first author of all papers, has implemented all techniques, systems and applications described within them and has designed the user evaluation described in paper IV. Jane Shaw, Suzi Adams and David Taylor substantially contributed to the data and task sections of paper VI as well as to the use case in that paper. Jane Shaw furthermore contributed a major part of the use case in paper VII. The main focus of the included publications are as follows:

**Paper I, paper II and paper III** address the issues of categorical and mixed data visualization by presenting an approach for interactive quantification of categorical data and an interactive visualization system incorporating this approach. In difference to most preceding quantification methods, the approach presented in these publications utilize relationships within both categorical and numerical variables for quantification.

**Paper IV** includes a formal user evaluation comparing the performance of the two main approaches to categorical data visualization when carrying out two basic data analysis tasks.

**Paper V** addresses the issue of dimensionality reduction through a reduction method combining several quality metrics, where the analyst is guided in the selection of an appropriate number of variables by a representation of loss of structure, and where analysis is facilitated by various variable ordering algorithms.

**Paper VI** presents a system for examination of high dimensional microbial populations, including explorative dimensionality reduction based on combinations of quality metrics and combining information visualization methods with methods commonly used within the microbiology domain.

**Paper VII** continues the work on exploratory dimensionality reduction through a generic system combining flexible and interactive dimensionality reduction with visual overview of structures within the high dimensional data set as well as visual exploration of a subset of variables.

# CHAPTER 2

## BACKGROUND

This chapter aims to provide a more in-depth background to the contributions of this thesis and to provide an overview of the current state of research in the areas of categorical and high dimensional data analysis. The chapter is divided into two sections, each presenting research within one of the areas and each beginning with a brief motivation as to why these areas are important to address. Since the use of algorithmic guidance is the primary theme of the thesis, the combination of algorithmic methods and visualization in the context of categorical and high dimensional data will be discussed throughout the chapter.

### 2.1 CATEGORICAL DATA

Categorical data is data where the values are names, commonly referred to as categories. In the case of ordinal categorical data an inherent order exists between the categories of a variable, but for categorical data, in general, there does not exist any straightforward order or measure of similarity between categories comparable to that of numerical values. When it comes to measuring the similarity of individual data items in a categorical data set, several measures based on category matching are available. The most simple being the overlap similarity measure, where a similarity of 1 is assigned for identical values and a similarity of 0 to not identical values. Thus, in a multivariate data set the items with highest number of overlapping values are considered most similar [11]. The Jaccard coefficient is related to the overlap measure, but focusing on sparse binary data, providing efficient similarity computation by only considering matches of non-zero values [81].

Similarity measures based on category matching only consider whether two values are equal or not and does not take any other aspects of the data into account. This is a limitation which has been addressed through data-driven similarity measures, which also take into consideration the frequency distributions in a data set. Boriah et al. [11] describes and compares a range of similarity measures for categorical data items in the context of outlier detection. Some of the conclusions drawn from this study are that, in many cases, the overlap measure did not perform well and that no single measure was superior or inferior to the others although some measures consistently rendered better performance. Although a range of similarity measures for categorical data are available, the notion of categorical similarity and the use of these measures is

more complex than for corresponding numerical measures. Furthermore, their performance tends to be more dependent on underlying structures in data, as some assign higher significance to rare category matches whereas others assign higher significance to commonly occurring matches [11]. In terms of visualization, the number of methods designed for categorical data are far fewer than the number of methods designed for numerical data [32]. An explanation of this may be the complexity of similarity measures.

### 2.1.1 VISUALIZATION

Most visual representations used for numerical data are in some way based on the numerical values of data items. Considering a scatter plot as a simple example, each item is represented by a glyph positioned according to its numerical value, usually such that higher values are positioned further to the top right part of the plot than lower values. Through this, items that are similar to each other are automatically positioned closer together than items that are less similar to each other, providing a straightforward and relatively easily interpreted representation of relationships and patterns within the data. This agrees well with the Gestalt law of proximity [87], saying that objects positioned close together are perceptually grouped together and, hence, spatial proximity will usually be interpreted as reflecting similarity. As the concept of similarity is more complex for categorical values, visual representations designed for categorical data often utilize other characteristics than similarity. Quite commonly, as will be described later in this section, the focus of the representation lies on category frequency, meaning the relative number of items in the data set that take a certain categorical value or a certain combination of categorical values.

Comparison of category frequencies within a single variable may be successfully carried out using a histogram or a bar chart where height represents frequency and each bar represents a category. However, for multivariate data sets these methods are less effective, requiring one chart for each variable and providing very limited possibilities for analysing inter-variable relationships. For analysing relationships within two binary variables, a fourfold display may be used [32]. The fourfold display is reminiscent of a pie chart and represents each combination of categories by a quarter circle whose area is proportional to the frequency of the category combination. The associations between cells are visually represented through confidence rings along the outer boundaries of the circle segments, where the rings of two adjacent circle segments overlap if the variables are independent of each other. The utility of fourfold displays is highly limited since they are only able to display two variables at a time and only variables that are binary. For association analysis of two variables with higher cardinality a sieve diagram can be used [32]. The sieve diagram is made up of a grid of rectangular cells where the columns represent categories of one variable and the rows represent categories of the other. The width and height of the columns represent the total frequency of corresponding category. Following this, the area of each cell in the grid corresponds to the expected frequency if the two variables were independent. Within each cell a grid is drawn with density corresponding to the observed frequency of the category combination. Variations in density between cells indicate deviation from independence.

The mosaic plot [32] is closely related to the sieve diagram in terms of representing frequencies of category combinations through rectangular cells but, unlike the sieve diagram, the mosaic plot is not limited to displaying only two variables. Furthermore, the cell sizes in a mosaic plot

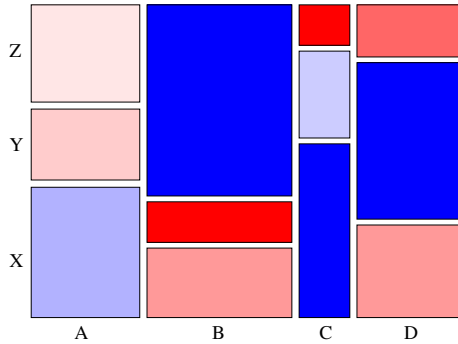


Figure 2.1: A mosaic plot of two variables, including four (A, B, C, D) and three (X, Y, Z) categories respectively. The frequency of category combinations is represented through rectangle size and the frequency's deviation from an independence model is represented by colour.

correspond to the observed frequency of categories and not the expected frequency. In a mosaic plot displaying the categories of two variables, as in figure 2.1, the columns represent categories of one variable and have width according to the total frequency of corresponding category. The height of the cells corresponds to the relative frequency of the categories of the second variable. Similarly to the grid density of the sieve diagram, the cells of the mosaic plot can be coloured according to deviation from a model of independence. Mosaic plots for three or more variables are achieved by recursively subdividing the cells according to frequency [32]. Another way of representing multivariate categorical data using mosaic plots is to use a mosaic matrix [31, 32]. The mosaic matrix, which resembles a scatter plot matrix, is a symmetrical matrix of all two variable mosaic plots. In principle the mosaic matrix is able to display any number of variables but, from a practical point of view, it is limited by display resolution to three or four variables according to Friendly [32]. Additionally one visual representation used for analysing frequencies in multivariate categorical data sets is the cobweb diagram [85]. Visually, a cobweb diagram is reminiscent of a network visualization where each category is represented by a circular node. The nodes are connected by lines whose widths correspond to the fit of the category frequencies to a model of independence, similar to the colouring of a mosaic plot, and where line colours indicate the sign of the deviation. As for mosaic matrices, the cobweb diagram has no theoretical limitation in terms of number of variables to display but, as the number of unique categories in the data set increases as a result of either more variables or higher cardinality, the diagram may easily get cluttered.

A more recent visual representation displaying frequencies of categories is Nested Rings [86], a representation made up of a series of concentric rings. Each ring represents a variable and is divided into a number of slices, representing categories, with size corresponding to category frequency. The Nested Rings representation is interactive and supports dynamic querying in terms of filtering through selection of a category slice, constraining the representation to only display relative frequencies of items belonging to the selected category. As with mosaic matrices a limitation of Nested Rings is its inability to display a larger number of variables in a

usable way; according to the authors five variables seems ideal for display. Shiraishi et al. [76] present a method called granular representation for visual analysis of categorical data, where data items are represented by small circles which are coloured according to some variable. The items can be separated into groups, either manually by dragging category labels which attract items belonging to the category towards the dragging position, or by using an automated clustering approach where items are separated based on their categorical values over multiple variables. This representation is combined with bar charts for examining the relative frequencies of categories. Although the scalability of the granular representation approach was not discussed by the authors, it appears that this representation will also quickly become harder to overview in a multivariate analysis context, as the number of categorical variables and their cardinality, and through that the number of visually represented clusters of data items, increase.

Several of the more recent visualization methods designed specifically for categorical data have a layout based on popular methods for numerical data. One of these is the Hammock plot [71] which utilizes the layout of parallel coordinates. In the Hammock plot the axes of parallel coordinates are replaced by univariate descriptors which textually represent category names. The polylines of parallel coordinates are replaced by polygons connecting categories of adjacent variables, with polygon width proportional to the frequency of corresponding category combination. Parallel sets [58], displayed in figure 2.2, is another method built upon the layout of parallel coordinates, which includes frequency representation in the display. Similar to parallel coordinates, the variables of a multivariate data set are represented by parallel axes in parallel sets. The axes of parallel sets are horizontally laid out and instead of displaying individual items, parallel sets focuses on representing categories and category frequencies. Each category of a variable is represented by a box whose width corresponds to the relative frequency of the category. The categories of two adjacent axes are connected through bands whose width corresponds to the frequency of the combination of categories. Differently from many visualization methods designed for categorical data, parallel sets includes special features for dealing with the numerical variables of a mixed data set. The numerical variables are categorized into a number of equally sized bins, and bands stretching from a categorical axis towards the bin of a numerical axis have a triangular form, which distinguishes them from the bands between two categorical axes. Parallel sets also includes a range of interactive features for exploring the data and utilizes a histogram-like visual representation to display deviation from independence between adjacent axes.

In addition to parallel coordinates, a visual representation which has several times been adapted to suit the purpose of categorical data visualization is the Tree-Map, introduced by Johnson and Shneiderman [51]. A Tree-Map is a space-filling approach for visualizing hierarchically structured data, mapping the hierarchical information to rectangular areas and recursively slicing the rectangles for higher level hierarchies. CatTrees [56] are an extension of Tree-Maps with the ability to create a hierarchical structures from categorical data. The TreemapBar [45] combines Tree-Maps with bar charts and can be described as a bar chart with embedded Tree-Maps. Each bar represents a categorized data subset with height corresponding to a quantifiable value of the subset. For data subsets including hierarchical structures the bars are filled using a Tree-Map layout algorithm. Issues of increasing number of categories and hierarchies are dealt with by providing focus+context through adjusting screen space allocation and widen bars within the

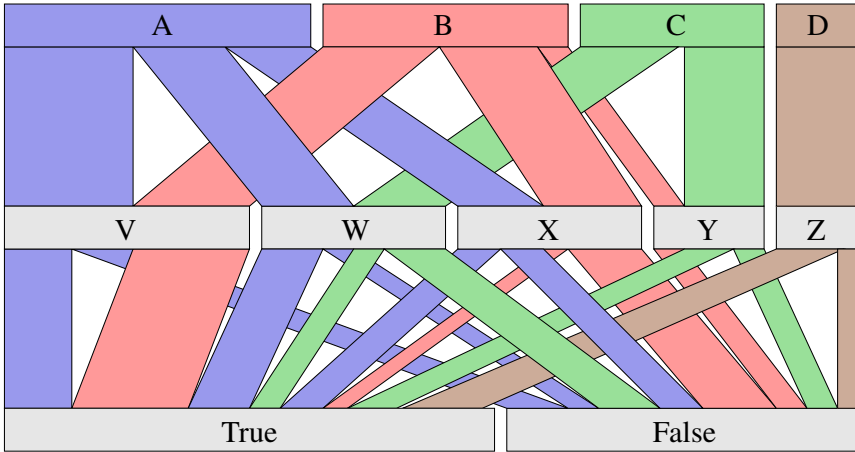


Figure 2.2: Parallel sets displaying a data set with three categorical variables, including four (A, B, C, D), five (V, W, X, Y, Z) and two (True, False) categories each. The width of the axis bars corresponds to category frequency and the width of bands between axes corresponds to frequency of combinations of categories in adjacent axes. Colouring is done according to the categories of the top variable.

focus area. Also related to Tree-Maps, the Attribute Map View [60] utilize a Tree-Map like layout for displaying categorical transaction data. In the Attribute Map View each categorical variable is represented by a horizontal bar. The bars are divided into a number of rectangles, each representing a categorical value of the variable, with size according to category frequency.

Seo and Gordish-Dressman [73] extend the Rank-by-Feature framework [74], which is a tool for exploring multivariate data, with features for analysis of categorical data. These features include ranking criteria for evaluating relationships between categorical variables and possibilities of partitioning the data into groups based on categorical information, by splitting the data set horizontally or vertically. The framework also supports clustering of the categorical partitions for identification of meaningful groups of data and for comparison of partitions.

A data type partly related to categorical data is set-typed data. Set-typed data can be categorical in its nature and exists where a data item can take several different values for a single variable. This may, for instance, occur in customer surveys where a set of answers can be selected by a participant for a single question. Issues related to set-typed data have not been addressed in this thesis. However, several categorical variables of low cardinality may, in some cases, be combined into one set-typed variable to reduce the dimensionality of the data set [28], and as such the visualization of set-typed data is to some extent relevant for this thesis. Freiler et al. [28] presented the set'o'gram representation for visual analysis of set-typed data. The set'o'gram is based on a histogram layout and represents each value, or category, of a variable by a bar, the height of the bar representing the number of items taking the corresponding value. The bars are divided into a number of blocks of varying width, the height of the widest block representing items taking only the value of the corresponding bar, the height of the second widest block representing the number of items taking the value of corresponding bar and taking the value of one

other bar as well, the third widest bar representing items taking the value of the bar and the value of two other bars, and so on. The issue of set-typed data is also addressed by Collins et al. [19] through the Bubble Sets approach. Bubble Sets enclose individual and possibly overlapping sets in a standard visual representation using isocontours. This approach retains the original visual representation by drawing Bubble Sets on top of an existing representation and is, therefore, not limited to any specific data type.

Although a range of various visualization methods exist which are designed and modified to deal with the specific characteristics of categorical data, they are less commonly available than methods designed for numerical data [32]. Furthermore, their usability has not been fully examined since very few evaluations have compared visualization methods for categorical data. The main conclusion drawn from the literature reviews carried out in connection with this thesis is that methods designed for categorical data are more special purpose and often not as generic as methods designed for numerical data. In addition, their usability is often more dependent on the structure and size of the data. Some methods, such as the fourfold display, are strictly limited to displaying only a certain number of variables or categories. Others, such as mosaic plots and matrices, are not theoretically limited but are often not as usable for more than a moderate number of variables and categories. Methods such as parallel sets, which are based on the layout of parallel coordinates, seem to be able to display more variables. None the less, it appears as if they have a tendency to get more easily cluttered than standard methods for numerical data as the number of variables and categories increase. Not least since features such as wide polygon bands may take up large amounts of screen-space. The design of other methods are based on a certain structure, such as hierarchical data, and they are hence rarely useful for data not having this structure. Additionally, according to the Gestalt laws, spatial proximity will usually be interpreted as similarity between objects. Although the categories of a categorical variable generally do not have any inherent order, their order in a visual representation will, hence, be interpreted as indicating similarities and, as shown by Friendly [30], the visibility of patterns in categorical data displays are dependent on the internal order of the categories. With the support of this, the usefulness of visual representations for categorical data may be increased by utilizing an appropriate category ordering algorithm in combination with visualization.

### 2.1.2 QUANTIFICATION

Another approach to visual analysis of categorical data is to employ quantification, which means that each categorical value is represented by a unique numerical value. The quantified data is then treated as if it were numerical and might thus be analysed using algorithms and visual representations originally designed for numerical data. In this way the generality and availability of these methods can be made use of, as well as their ability to more efficiently and effectively display larger data sets. Furthermore it enables analysis of mixed data sets, including both categorical and numerical data, within a single display. When employing this approach it is highly important to select an appropriate method for mapping the categories to numerical values. The assigned numerical values will, due to the nature of numerical data, imply relationships such as similarity between categories. Unless these relationships reflect existing structures, the quantification and subsequent visualization may cause erroneous interpretation of the data. Various

quantification methods are available, which can be separated into roughly three different types; arbitrary, manual and algorithmic.

**Arbitrary quantification** includes all methods where numerical values are assigned to the categories in an arbitrary manner. Random assignment of numerical values is a typical arbitrary quantification, as is quantification based on order of appearance in the data file. Numerical assignment purely based on the names of the categories (alphabetical ordering) may also be defined as an arbitrary quantification, since alphabetical ordering often does not reflect any relevant relationships within the data. Arbitrary quantification may be simple and quickly implemented, but is generally not recommendable since it, more or less by default, generates artificial patterns.

**Manual quantification** means that the analyst, or another domain expert, manually assigns numerical values to the categories based on domain knowledge and knowledge of the task at hand, also including the use of meta-data. A manual quantification ensures that similarities between categories make sense in the context of the data domain, but it is dependent on a knowledgeable analyst. Furthermore, it may be an inefficient and time consuming approach as the number of variables and categories increase.

**Algorithmic quantification** is a time efficient approach where automated methods are used to assign numerical values. These methods often assign numerical values based on the underlying structures of the data, and hence the numerical representations to some extent reflect relationships in the data set. This approach overcomes the manual quantifications dependency of a knowledgeable user, in terms of not requiring any domain knowledge to be employed. This is however also the main drawback of this approach since it does not take into account any domain relevant information that is not part of the data structure, but may be known by an analyst. The algorithmic approach may, hence, fail to reveal patterns which are relevant within the task context.

A main difference between arbitrary quantification and the two other methods is that when utilizing an arbitrary approach the user, or system designer, does not make any active choice regarding the quantification method to use. The remainder of this section will present various quantification methods suggested within a visualization context, most of which can be defined as algorithmic quantification methods.

One of the most commonly suggested methods for quantification of categorical data is Correspondence Analysis (CA) [36, 37]. CA is a statistical method for analysis of data matrices with non-negative values and is often used for analysing tables of category frequencies, known as contingency tables. The algorithm for CA was concurrently developed in several countries using various names [83], such as Optimal Scoring, Dual Scaling and Homogeneity Analysis. Tenenhaus and Young [83] show that these methods all lead to the same equation. For simplicity the approach will be referred to as Correspondence Analysis, or CA, throughout this thesis. The most basic case of CA is known as Simple Correspondence Analysis (SCA) which includes the analysis of associations between the categories of two variables. SCA is applied to a two-way

contingency table where rows represent categories of one variable and columns represent categories of the other variable. To analyse a set of variables Multiple Correspondence Analysis (MCA) is used. MCA is similar to SCA but is applied either to a *Burt matrix* or an *Indicator matrix* [36]. The Burt matrix is a symmetrical matrix of all two-way contingency tables of the data set and can, in terms of structure, be compared to a scatterplot matrix. The indicator matrix is in part similar to the  $m$ -by- $n$  matrix structure of a data set with rows representing items. However, the columns of the indicator matrix represent all categories of the data set, that is, all values that can be taken for all categorical variables. When applying CA to a table or matrix, the rows of the matrix can be thought of as points in a multidimensional space defined by the columns [36]. The CA algorithm employs Singular Value Decomposition (SVD) [34] to define independent dimensions, or principal components, in the table. SVD decomposes a rectangular matrix  $A$  into  $U\Sigma V'$ , where  $U$  and  $V$  are unitary matrices and  $\Sigma$  is a diagonal matrix with diagonal values corresponding to the singular values, or principal values, of  $A$ . Based on a theory called Optimal Scaling, the values of the first independent dimension extracted using SVD, which explains most of the variance within the table, can be used as numerical representations for the categories in the table [37]. Thus, when employing CA for quantification, the numerical representations will be based on relationships in terms of underlying distributions in the data. CA has been used as a quantification method throughout the work described in this thesis, and has been implemented through applying MCA to Burt matrices. Cuadras et al. [22] compare three approaches for quantification of categorical data; CA, an alternative approach using Hellinger distance [67] and a log-ratio approach [1]. Their conclusions are that CA and the Hellinger approach often provide similar results under certain circumstances, such as when rows and columns are almost independent. The log-ratio approach on the other hand may often provide quite different results.

CA has been used in different ways in visualization. Greenacre [37], who defines CA as a generalization of scatter plots, presents a range of methods for visualization of CA results using scatterplot techniques. Some examples are CA maps and CA bi-plots. A CA map displays the categories of a data set in relation to the principal components extracted using CA. Commonly the first two principal components are used as axes in a two-dimensional plane. The map shows the projection of the categories onto this plane. A CA map can be either symmetric, with both row and column categories scaled according to principal components, or asymmetric, with either rows or columns scaled according to principal components and the other scaled in standard coordinates. The CA bi-plot is a technique where each row and column of a data matrix is displayed in a joint asymmetric map, each represented by a point. The points are laid out in the plot such that the scalar product of row vectors and column vectors approximate the values in corresponding cell in the data matrix. The name comes from its ability to display both rows and columns, and a bi-plot can be of any dimensionality, although two-dimensional plots are most common. Friendly [31] uses the first principal component of CA to reorder rows and columns of mosaic displays, to make association patterns more meaningful. In the same way as when employing CA for quantification, Friendly's ordering of categories are based on the principle of Optimal Scaling [37].

Rosario et al. [70] presented the Distance-Quantification-Classing (DQC) approach, which is an approach used for pre-processing of categorical variables prior to visualization using standard methods designed for numerical data. In the distance step of the process, the distance between

categorical values are calculated through identification of independent dimensions. The distances are then used to assign order and spacing to categories in the quantification step. In the final classing step, the quantification is used to identify categories within a variable which are similar enough to be merged into one category, aiming to reduce clutter in high cardinality variables. Various algorithmic techniques could be used for the steps of the DQC process. As an example, Rosario et al. [70] use two versions of CA, MCA and their own focused correspondence analysis, for the distance step, Optimal Scaling for the quantification step and a hierarchical clustering algorithm for classification. Yang et al. [97] describe a quantification approach for mixed data sets, including a combination of categorical and numerical variables, by applying a clustering approach. In addition they present a number of cardinality reduction techniques for more efficient CA computation. These strategies include various clustering approaches applied to reduce the cardinality of the numerical variables, the categorical variables or a combination of both.

Ma and Hellerstein [61] addresses the issue of ordering categorical values through focusing on two tasks; 1) identification of groups of similar categories within a variable, 2) identification of relationships between categories of different variables. Their suggested approach starts with a clustering of data items based on domain semantics, next the clusters are ordered and finally the categories are ordered within the clusters, aiming to reduce ‘holes’ within clusters in scatterplots and to minimize cross-overs in parallel coordinates. A faster algorithm for ordering of categorical variables was later presented by Beygelzimer et al. [10]. In their approach the data set is modelled as a graph, nodes representing categories and edge weight corresponding to the pairwise similarities of nodes. Before ordering the graph, using a spectral method, the graph is recursively coarsened, reducing the size of the graph and hence the complexity of the ordering. The coarsest graph is then ordered and the ordering is propagated back through the less coarse graphs by interpolation. Both these approaches aim at efficient category ordering, but do not directly focus on distances between categories in the same way as for instance CA.

Shen et al. [75] present a quantification approach aiming at clustering and visualization. They quantify based on similarities between data items and a reference set of randomly sampled items. The similarities are composed into a matrix which is then clustered and visualized. Since this quantification aims specifically at clustering a key focus of the approach is preserving the difference between inter-cluster similarities and intra-cluster similarities. Chandola et al. [15] also use a reference data set for quantification of categorical data. They introduce the concept of separability statistics, which are statistics representing the distance between a data item and a labelled reference data set, and utilize the statistics to map the categories to a numerical data space, providing examples of how the data can be visualized using scatterplots and histograms.

A phenomenon which may occur when using visual representations designed for numerical data for the analysis of categorical data is data overlay. In most cases a structural difference between categorical and numerical data is the number of values an item can take. Although categorical variables may be of high cardinality, the number of unique values a data item can take is generally far less for categorical variables than for numerical variables. A natural effect being that more items take exactly the same values for categorical data and, hence, the occurrence of multiple glyphs being plotted on top of each other in scatterplots or multiple lines following the same path across several adjacent axes in parallel coordinates is quite likely. Problems following data overlay include the difficulty of appreciating frequency distributions and of extracting

information regarding which path an individual data item takes across several variables. For visual representations designed for categorical data the former is generally not an issue, since most of them are designed for representing category frequencies. To overcome this issue in visual representations designed for numerical data, various approaches can be employed for enhancing and facilitating the perception of frequency distributions. Although this has not been part of the research presented in this thesis, a brief summary of some of the available methods will follow, as this is a relevant issue in the area of categorical data visualization.

A common method for displaying frequency distributions in visual representations of numerical data is to represent data density through opacity [50, 62, 89], representing each data item by a semi-transparent object (such as a polyline in parallel coordinates or a glyph in a scatterplot) and over-plot objects that coincide. The result being that high density areas have higher opacity than sparse areas and, hence, visual density will correspond to data density. This approach may be usable for displaying clusters and frequency distributions in numerical displays. However, in the case of categorical data the number of items drawn exactly on top of each other is often higher than for numerical data, since perfect overlay, where items take the exact same value, is common for categorical data, whereas numerical data overlay often is only partial, with items taking very close but not exactly the same values. As a result, the number of items overlaying may vary and still appear to have equal density when employing this method for categorical data.

Several other techniques facilitating the perception of frequency distributions in parallel coordinates are available. Hauser et al. [42] suggested the use of histogram representations along the axes, where the width of the histogram corresponds to the underlying frequency. Additionally one method is the use of curved lines, as introduced by Theisel [84] who used curves for revealing correlations between non-adjacent axes. Graham and Kennedy [35] continued this with a method using curved lines which partly overcomes the issue of data overlay. They replace the straight polylines in parallel coordinates with curves whose shape between two adjacent axes is based on the values of the corresponding data items for the two axes, as well as the value it takes for the directly preceding axis. The result being that items with the same value for two adjacent axes, which normally would overlay exactly, diverge slightly if taking different values for the preceding axis, making it easier to follow the paths of individual data items. Additionally one approach is presented by Havre et al. [43], who address the issue of data overlay by inserting one or two additional axes between the original variable axes in parallel coordinates. The data items are spread and ordered along the additional axes, according to categorical values taken in adjacent original axes. This provides a visual appearance similar to the bands between axes in parallel sets [58] when two additional axes are used.

This section has presented a range of methods for representing categorical data with numerical values. Most of the available quantification approaches utilize a purely automated approach, which provides fast quantification based on structures in the data. As discussed previously in this section, this approach is, however, limited in not making use of the context-based knowledge of a domain expert. On the other hand, the approach of manual quantification, which is purely based on the knowledge of a domain expert, may be inefficient and time consuming. Based on this, it appears to be beneficial, in the context of visualization of categorical data, to employ quantification processes where the advantages of manual and algorithmic approaches are combined. This can be achieved by utilizing the concept of algorithmic guidance. By employing automated

methods as guidance and suggestion of similarities between categories, the advantages of algorithmic quantification are made use of. Through combining this guidance with possibilities for modifying the algorithmic result if required, according to domain knowledge and task, information relevant to the analytical context can be made use of.

## 2.2 HIGH DIMENSIONAL DATA

High dimensional data, including a large number of variables, is one of the major challenges in both visualization and data mining. This is to a large extent due to what is known as ‘*the curse of dimensionality*’ [9], which refers to the fact that the amount of data needed to maintain a certain level of accuracy often grows exponentially with the dimensionality [41]. The curse of dimensionality concerns the fact that data become increasingly sparse as the dimensionality increases [81], making many types of data analysis harder. Issues following from this are, for instance, that the definition of density as well as the distance between items become less meaningful, with the potential consequence of poor cluster quality and reduced accuracy in classifications [81]. Due to the increasing sparsity in high dimensional data sets the nearest item may often be far away and, when using a distance measure such as the Euclidean distance, the difference in distance between items will become smaller with an increasing number of variables. Additional issues of high dimensionality include the risk of more variables introducing more noise and redundancy due to a large number of similar variables, as well as an increase in computational cost in terms of both time and memory usage.

In the context of visualization additional issues related to high dimensionality has to be considered. Primarily, the capacity of human perception is limited by the precision of the eye as well as by the ability of the human mind to process visual patterns [23, 87]. The resolution of monitors, which is affected by both pixel resolution and display size is, however, still a bigger issue than the precision of the human eye, and in the context of information visualization an issue more relevant than the number of perceivable pixels is whether patterns can be perceived in the display [23]. The following sections of this chapter will discuss research addressing the issues of high dimensionality in the context of information visualization.

### 2.2.1 VISUALIZATION

Commonly available visual representations for multivariate data, such as parallel coordinates and scatterplot matrices, are usable tools for displaying moderately sized data sets. However, as the number of variables in the data set increases the usability of such representations is drastically reduced due to visual clutter as well as to difficulties in navigation and interaction with the representation. Aiming to overcome these issues, a small number of visual representations focusing on analysis of data sets of higher dimensionality have been developed. A brief overview of these will follow.

Pixel displays, as presented by Keim [54], represent large sets of data through mapping each data value to a coloured pixel. The pixel display can be divided into a number of areas, where each area either represents a variable or a data item. If representing a variable, the colours of

the pixels within represent the data values of all items for the corresponding variable, and if the area represents a data item, the pixels represent the values of all variables for the corresponding item. To facilitate diverse analytical tasks and to enhance perceivability of patterns, various pixel arrangement techniques can be used within the areas. In theory, pixel-oriented techniques are able to display a large number of data values, since only one pixel is used for each value. For a screen resolution of 1600x1200 pixels, pixel displays would be able to represent approximately 1.9 million data values. From a practical point of view it would, however, be difficult to perceive many of the small patterns that may occur in the display.

Yang et al. [93] introduced the Value and Relation (VaR) display, which is a two-dimensional display where each variable of a high-dimensional data set is represented by a glyph. The glyphs are laid out in the display based on relationships between the variables and, through the use of pixel-oriented techniques, the values of data items are mapped to the area of the corresponding variable glyph. In addition to the basic visual representation, the VaR display provides various automated and manual tools for navigation within the display and for selection of variables. The former including for instance overlap reduction techniques and zooming features and the latter providing a user-driven dimensionality reduction.

Both Keim [54] and Yang et al. [93] utilize algorithmic methods to aid the analyst in identifying potentially interesting patterns. Without this the usefulness of the methods would be drastically reduced. For instance, a pixel display where the pixels are randomly laid out may quite possibly appear, to the analyst, to contain only noise since pixels of various colours would be randomly mixed in the display areas, considerably obstructing the analysis of relationships between areas. Comparing the VaR display with pixel displays, the VaR display may be considered to be an interactive pixel display where the pixel areas, representing variables, are laid out based on variable relationships in a two-dimensional plot. An integral part of the VaR approach, and one of its main advantages in relation to standard pixel-oriented approaches, is the algorithmic variable layout process which guides the analyst in exploring relationships between variables.

In addition to visual representations, some systems for exploration of high dimensional data utilizing algorithmic approaches are also available. For instance Nam et al. [64] present the ClusterSculptor system which combines cluster analysis techniques and interactive features for clustering of data sets including hundreds of variables. Barlow et al. [7] utilize partial derivatives for detection of multivariate correlations, using a stepwise exploration approach to reduce clutter. The utility of both ClusterSculptor and the system presented by Barlow et al. are to some extent task dependent, since they focus on the specific tasks of examining clusters and multivariate correlations. Overall, few visual representations are available which are able to effectively display all variables of a data set within one view. A more commonly employed method is to reduce the number of variables prior to visualization, which will be discussed in the following section.

### 2.2.2 DIMENSIONALITY REDUCTION AND ORDERING

The utilization of dimensionality reduction enables usable visualization through standard visual representations, as well as generating better results for many data mining algorithms [81]. Through reduction of dimensionality, issues relating to the curse of dimensionality are alleviated, the time and memory requirements of algorithms and visualization are reduced and, not

least, through the use of appropriate algorithms irrelevant variables may be eliminated and noise may be reduced. The reduction of dimensionality can be divided into two types, according to Tan et al. [81]:

**Dimensionality reduction** includes techniques where the number of variables is reduced by creating a new set of variables, which in some way are a combination of the old attributes. The aim being to create a new and smaller set of variables capturing the important information in the data set more efficiently and effectively.

**Feature subset selection** includes techniques which reduce the dimensionality by representing the data using only a subset of the original variables. Goals of subset selection include the removal of redundant variables, which may occur if several variables are strongly correlated, as well as removal of irrelevant variables containing little or no information useful for the analytical task.

For simplicity all approaches used to reduce the number of variables in a data set are referred to as dimensionality reduction throughout this thesis.

Some of the most commonly used methods for dimensionality reduction are automated methods where the original data are projected into a low dimensional space in which new variables are often linear combinations of the original variables. Several such methods are based on Singular Value Decomposition (SVD) [34]. Principal Components Analysis (PCA) [20, 52] is an SVD based method for dimensionality reduction which is related to Correspondence Analysis (as described in section 2.1.2). PCA transforms the original variables of a data set into a new set of variables, or principal components, using SVD, aiming to retain as much as possible of the variation in the data set. The principal components, which are linear combinations of the original variables, are uncorrelated and ordered such that the first few components explain most of the variation within all of the original variables and, hence, can be used as a lower dimensional representation of the original data set. The contribution of each of the original variables is known as the loading, and the value of a sample in a principal component when projected to the new space is known as the score. Often a data set reduced using PCA is visually presented using a two or three dimensional scatterplot, where the axes represent the first principal components and where items are laid out based on their scores. Jeong et al. [49] present a system for visual and interactive analysis of the result of PCA, aiming to assist the user in understanding PCA, employing a multiple coordinated views approach where the original data space is linked to the space of the principal components.

PCA can be defined as a special case of an analysis method belonging to the group of statistical techniques known as multidimensional scaling (MDS) [21, 59]. The aim of MDS is to define a low dimensional projection of a data set such that the distances between items in the projected space correspond with the dissimilarities between items in the original data set as well as possible. Due to different notions of matching a range of MDS techniques are available. One of these is known as classical scaling, which equates with PCA when the dissimilarities are defined as Euclidean distances [21]. Williams and Munzner [92] present an interactive system for fast computation of MDS which lets the user steer the MDS process during computation by selecting local regions of interest within which to focus the computational effort of the algorithm.

For efficient MDS computation they use hierarchical data structures and a progressive layout approach. The presented system allows for immediate display of the overall data set structures, which aids the user in identifying areas of particular interest. While interactively drilling down into the hierarchical structure, the MDS computation is focused to the current area of interest.

Self-organizing maps (SOM) [55] is another automated method often used in the context of dimensionality reduction and visualization. SOM can be seen as a method for projecting high dimensional data onto a lower dimensional space and is, as such, related to MDS and PCA, although not based on SVD. Additionally, SOM may also be defined as a method for visualization of high dimensional data, since the output map of SOM is often displayed. The SOM algorithm utilizes a set of weight vectors to project the original data items to a low dimensional grid, using unsupervised learning methods. In an iterative process, the weight vector closest to each data item is identified and updated such that it more closely resembles data items related to it, creating an output map where similar data items are positioned in nearby locations. Other dimensionality reduction methods commonly used within data analysis include Factor Analysis [81], which is based on the concept that the original variables can be defined as linear combinations of a set of variables, or characteristics, which cannot be measured; Locally Linear Embedding [81], which analyses local neighbourhoods and reduces dimensionality while aiming to preserve local structures; and Linear Discriminant Analysis (LDA), which projects data to a linear subspace based on cluster information, aiming to separate the clusters [41]. For instance, Choo et al. [18] present an interactive system for classification, performing dimensionality reduction prior to classification using LDA.

Friedman and Tukey [29] presented the projection pursuit algorithm for linear projection of multivariate data sets onto one or two-dimensional spaces. According to Friedman and Tukey a disadvantage of many linear dimensionality reduction methods, such as PCA, is that the projection is based only on a global property, such as the data items variances along various directions in the space. Projection pursuit combines global and local properties, by utilizing both the distances between items in the data set and the variance of the data. Each direction in the data space is associated with a measure of usefulness, indicating how useful it is as a projection axis, and the projection is then varied so as to maximize this measure. For complex data the output of the projection pursuit algorithm may include several different projections. The grand tour, presented by Asimov [5], utilize visualization for exploration of multivariate data by presenting to the analyst a sequence of two-dimensional projections, selected such that it is dense in the set of all possible projections. More recently, Koren and Carmel [57] presented a family of linear transformations for dimensionality reduction. Their approach is able to take several data properties, such as similarities and clusters, into account simultaneously, and is made less sensitive to outliers through pairwise weighting. The reduction methods proposed by Koren and Carmel were later utilized by Engel et al. [24] who used it to create a structural decomposition tree for high dimensional data. The reduction is preceded by a data decomposition using a hierarchical clustering method, and the results are displayed using an interactive star coordinate based method [53]. The clusters of the hierarchical clustering are represented as nodes in the display and the projection path defining the node layout is represented by line segments, aiming to show how data points are projected in the display.

The dimensionality reduction methods described so far are methods where the new set of variables are a combination of the original variables in the data set. In many cases a single original variable may possibly influence, or be part of, several variables in the new data space. Thus, the relationship between the original and the new set of variables is not always intuitive. This is a major disadvantage of these techniques, especially in the context of analytical tasks where the relationships between original variables are of interest, or when the influence of specific variables onto the overall data structures are to be explored. Other approaches which may sometimes be more beneficial include the selection of a subset of variables or the creation of new variables based on grouping of similar variables. An example of the latter is the Principal Components Variable Grouping (PCVG) presented by Ivoisev et al. [48], which utilizes the properties of PCA to group highly correlated variables. As described previously, when applying PCA to a data set the loadings represent the contributions of the original variables to the principal components. Each principal component has a loading vector and variables having the same orientation in this loading space are correlated [48]. Based on this, Ivoisev et al. create groups of highly correlated variables, while ensuring that each variable only appears within one group, and reduce the dimensionality by replacing the groups with a representative variable, which can either be one of the variables within the group or an average of the whole group.

Within the field of information visualization a range of interactive dimensionality reduction systems have been developed, with focus on preserving different structures within the data set. For many of these approaches the visual representation and enhancement of pattern visibility has been an integral part. Differently from many of the automated techniques, the dimensionality reduction methods developed within the information visualization community have often used more of a feature subset selection approach, as defined by Tan et al. [81]. Furthermore, they have often provided interactive and, to some extent, user-controlled selection of variable subsets, aiming to include the analyst in the reduction process, as well as often utilizing algorithmic methods. Hence many of the methods employ an algorithmic guidance approach.

Guo [39] presents a system focusing on cluster structures, where each two-dimensional subspace of the data set is analysed based on a ‘goodness of clustering’ measure. This measure is based on three criterion, the percentage of items within the clusters, the density of the clusters and dependency of the two variables. Guo utilizes a hierarchical clustering approach to sort the variables such that highly correlated variables are positioned close together. A matrix where each cell represents a two-dimensional subspace is used to display the result, using cell colour to represent the ‘goodness of clustering’ measure of the subspace. Through this the analyst may visually identify potentially interesting subspaces with good clustering properties. Subspaces including more than two variables can be formed and explored further by selecting several possibly interesting subspaces, or by utilizing a ‘goodness of clustering’ threshold. A related approach is presented by Seo and Shneiderman [74] introducing the rank-by-feature framework for interactive detection of potentially interesting variables through ranking of one or two-dimensional variable subspaces. The framework provides a set of ranking criteria, such as correlation, linear regression, uniformity, skewness and number of clusters, from which the user may select one for analysis. All possible variable subsets are analysed according to the selected criteria and presented in tabular displays where the cell colour represents the ranking score of the corresponding subset and where the cells are ordered according to score. As in the approach presented by Guo,

the analyst may interactively select interesting subspaces in the table for further analysis, guided by the ranking score.

Yang et al. presented two different dimensionality reduction approaches, the Visual Hierarchical Dimension Reduction (VHDR) approach [95] and the Dimension Ordering, Spacing and Filtering Approach (DOSFA) [94], both utilizing a similarity based hierarchical structure. The hierarchical structure is generated using a variable clustering algorithm based on similarities between variable pairs. A similarity threshold is used to identify clusters of similar variables and, by iteratively re-performing the variable clustering with increasing thresholds, a hierarchical structure is formed. The hierarchical structure is displayed using InterRing [96], a circular interactive representation for hierarchical trees, allowing manual modification of the structure. In VHDR [95] the analyst can explore the hierarchical tree structure and select interesting variable clusters from it. Dimensionality reduction is then performed by creating new variables representing the selected clusters, generating a new lower dimensional data set which can be analysed using standard visual representations.

The order of variables in a visual display has a large impact on our ability to perceive structures [3]. This is addressed by Yang et al. in DOSFA [94], which is not only an approach for reducing the dimensionality but also for finding a visual layout of variables to facilitate pattern detection. As the name indicates the approach includes three steps; variable ordering, variable spacing and variable filtering. DOSFA provides both a similarity based ordering, where similar variables are laid out close to each other, and an importance based ordering using an importance measure, such as, for instance, variance, where variables are ordered according to how much they contribute to the measure. Subsequent to ordering a non-uniform variable spacing approach is employed, where highly similar adjacent variables are positioned closer together than less similar adjacent variables. The final step is dimension filtering where a lower dimensional data set is generated by removal of variables, based on a combination of similarity and the importance measure used for ordering. For groups of significantly similar variables all but one are removed, as are variables of low importance. Throughout the three steps the user may interactively modify the ordering, spacing and filtering. Another approach for dimensionality reduction and ordering based on similarity is presented by Artero et al. [4], aiming to reduce clutter when displaying high dimensional data. The basis of their approach is a matrix containing similarity values for all pairs of variables in the high dimensional data set. Similar to the DOSFA approach, the algorithms described by Artero et al. position similar variables close together. Reduction of variables is achieved through removal of those variables most similar to their neighbours after ordering.

Peng et al [65] presents further approaches to variable ordering for a range of visual representations using various clutter measures, aiming to reduce the visual clutter of the displays. Related to the work of Peng et al., in terms of defining different metrics for different visual representations, aiming to find informative visual displays, are the projection quality metrics presented by Sips et al. [78], Tatu et al. [82] and Albuquerque et al. [2]. Differently from many of the dimensionality reduction approaches introduced within the field of information visualization these methods are more or less purely automated but, unlike methods such as MDS and PCA, they measure the quality of a projection in a visual space, instead of measuring quality in data space. Sips et al. [78] measure the quality of projections into two-dimensional scatterplots utilizing class consistency measures. Assuming that the data items are classified, two measures are used: one

based on the distance to the centre of the classes and the other based on the spatial distribution of the classes, considering a projection where the classes are clearly separated to be of high quality. Based on these metrics a threshold is used to select a good subset of all possible scatterplots. Tatu et al. [82] measured the quality of scatterplot and parallel coordinates projections using various measures. Utilizing an automated approach to identify views with either good cluster properties or significant correlation, focusing on both classified and unclassified data. Albuquerque et al. [2] present a similar approach but focus on the visual quality of RadViz [44], pixel-oriented displays and table lenses. Ferdosi and Roerdink [26] present methods for variable ordering in parallel coordinates and scatterplot matrices. Their reordering approach focusses on multivariate structures. It aims to find relevant subspaces for clustering and uses a bottom up approach where first all one-dimensional subspaces are ranked according to a quality value, secondly all two-dimensional subspaces including the highest ranked one-dimensional space are considered, selecting the highest ranked of these spaces and iteratively continuing in a similar manner with higher-dimensional spaces.

The DimStiller system, presented by Ingram et al. [46], is a dimensionality reduction system guiding the analyst through a chain of stepwise data transformations, including a range of techniques for dimensionality reduction and analysis. Focus is put on providing both global and local guidance throughout the reduction process, aiding the user in selecting a useful chain of transformations as well as providing visual feedback facilitating parameter tuning and identification of the most informative settings for a single transformation. Relating to the approach of utilizing algorithmically guided visualization for analysis of high dimensional data, Bremm et al. [12] presents a system where visualization and algorithmic methods are combined to find useful descriptors of a data set, where the data set consists of several multivariate descriptors which may have different dimensionality. They use scatterplots and grid based visual representations both to provide overview of the initial data set as well as for enabling comparison of pairs of descriptors, facilitating the removal of redundant descriptors.

Since dimensionality reduction reduces the amount of data, it often includes some loss of information. However with appropriate dimensionality reduction, removing only redundant and irrelevant data, the importance of the lost information is minimized. With information loss as focus, Schreck et al. [72] presented a projection precision measure which is used to evaluate the information loss of point-based projections by comparing distances between points in the original and projected data. The projection precision is then visually incorporated into the visualization of the projected data.

To summarize, a large number of dimensionality reduction methods are available and, as an effect of the increasing amount of gathered data, the improvement of existing methods and development of new ones is an active research area within data mining as well as visualization. Through removing similar variables redundant information is removed and by removing irrelevant variables the data is made less noisy. Furthermore, by removing variables from a display, visual clutter may be reduced and patterns may be more easily perceived. This section has provided an overview of some of the most commonly used automated methods for dimensionality reduction within information visualization, including MDS, PCA and SOM, and it has provided an overview of interactive and algorithmically guided dimensionality reduction approaches developed within the information visualization community. Most of the automated methods can

efficiently project data sets including a large number of variables to a lower dimensional space. A common drawback of these methods, as of automated methods in general, is that it distances the analyst from the analysis process and does not take prior and potentially important knowledge of the analyst into account. Utilizing the approach of algorithmic guidance these limitations can be overcome. Many of the available dimensionality reduction methods which utilize a combination of interactive and algorithmic methods are, however, focused on preserving only one or a few specific structures when reducing the data. Some may be focused on retaining the overall cluster structure of the data while others remove variables that are very similar. The reduction result when applying them to a data set may thus significantly differ. Nonetheless, in the case of exploratory analysis, the analyst usually has no clear hypotheses and does not necessarily know prior to analysis which structures may be of interest and, hence, which dimensionality reduction method may be most appropriate to employ. Furthermore, the removal of variables always involves a removal of structure and the selection of an appropriate number of variables to retain, to minimize the loss of important structures, is in itself an important challenge of dimensionality reduction. A matter of importance in connection with this, is the ability to provide an overview of the structures within the whole data set, to facilitate decision making during the dimensionality reduction process. These are issues addressed in papers V, VI and VII of this thesis.

# CHAPTER 3

## CONTRIBUTIONS

This chapter will present an overview of the main contributions of the publications included in this thesis. The chapter includes four sections. Section 3.1 describes the approach of interactive quantification for visual exploration of categorical and mixed data sets, including the contributions of papers I, II and III. This is followed by section 3.2 describing a performance evaluation comparing the quantification approach with a visualization method designed for categorical data, as presented in paper IV. The third and fourth sections focus on issues related to visualization of high dimensional data. Section 3.3 describes an interactive dimensionality reduction approach combining several quality metrics, presented in paper V. Finally section 3.4 presents an approach for algorithmically guided and explorative dimensionality reduction, incorporated in a system for exploration of microbial populations in paper VI and in a generic system for exploration of high dimensional data in paper VII.

### 3.1 INTERACTIVE QUANTIFICATION OF CATEGORICAL DATA

Categorical data is common in many areas such as the social sciences, biology, chemistry and medicine, and, as with other data types, the amount of categorical data gathered is constantly increasing, raising a demand for efficient and effective analysis methods. While numerical data may often be analysed using generic methods based on the numerical similarity of data items, such as scatterplots and parallel coordinates, there exists no generalized similarity measure for categorical data comparable to numerical similarity. Hence, specialized analysis methods are needed, which take the specific features of categorical data into consideration. In terms of visual analysis of categorical data, one of the two main approaches is usually employed. The first approach being to use visualization methods designed based on the specific features of categorical data. The second approach is to utilize a quantification method and then analyse as if the data would have been purely numerical. The second approach has been the focus of papers I, II and III. This section will describe the main contributions of these publications.

As concluded in section 2.1.1, the usability of visualization methods designed for categorical data analysis seems to depend on the data structure to a higher degree than methods designed for numerical data. This makes the approach of utilizing visualization methods for numerical data attractive when analysing multivariate categorical data sets. On the other hand, while assigning

numerical values to categories there is a substantial risk of introducing artificial patterns and, thus, the selection of appropriate quantification methods is an essential part of the analysis. Both manual quantification, where a domain expert manually assigns numerical values, and algorithmic quantification, where automated methods are used to assign numerical values based on data structure and distributions, have advantages and disadvantages. However, through properly combining the two it may be possible to provide quantification that is both fast and makes use of the domain and context based knowledge of the analyst.

Another issue closely related to categorical data visualization is that of mixed data analysis. In most areas where categorical data are present it is also common for data sets to include a combination of categorical and numerical variables. These data sets include diverse similarity types and hence different data subsets will require different analysis methods. Only a few visualization methods, such as parallel sets [58], take this into consideration whereas most overlook it and focus on only one data type.

### 3.1.1 OBJECTIVE

The objective of this part of the research was to design and implement a generic system for explorative analysis of categorical and mixed data sets. The primary goal, as presented in paper I, was to make use of the usability and generality of analysis methods designed for numerical data through utilizing a quantification approach and to provide an efficient as well as user-controlled quantification method which takes relationships within both categorical and numerical variables into consideration. The second goal, as presented in papers II and III, was to extend the quantification system with an interactive environment for combined visual and algorithmic analysis, to enable analysis of various data sets using the suggested methods and through this provide a level of confidence in their usefulness. To summarize, the objective of this work was to provide:

- efficient and user-controlled quantification;
- quantification taking relationships of all variables in a mixed data set into consideration;
- an interactive system where the quantification method is utilized for visual and algorithmic analysis.

### 3.1.2 RESULT

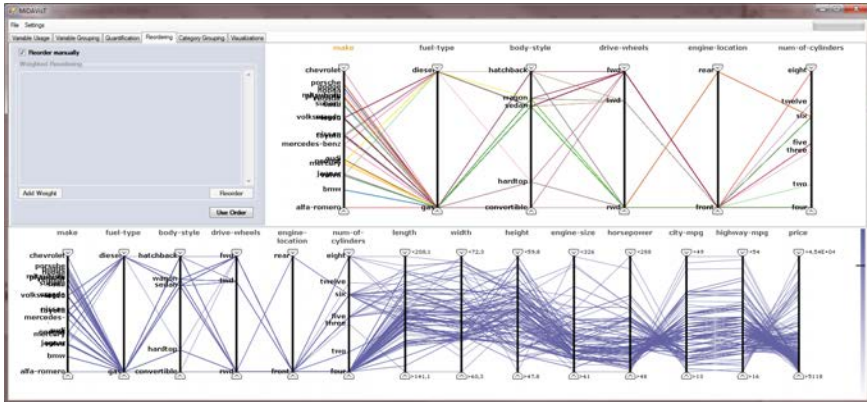
The aim of providing an efficient as well as user controlled quantification method was addressed in paper I, where a method combining algorithmic quantification and interactive features in a visual environment was presented. The proposed method also addresses the issue of mixed data sets through the incorporation of similarity information from the numerical variables into the quantification process. In the initial step of the presented method an algorithmic quantification is employed, deriving values of association between the categories in the data set. These values then act as suggestions of similarities between categories through which the analyst is guided in subsequent manual modification of the quantification result.

As an example, the method employs Correspondence Analysis (CA) [37] for the initial automated quantification, since CA is a commonly used and generic method not designed for any specific task and, hence, is suitable within an environment aiming at explorative data analysis. This part of the quantification is in many ways closely related to the method of Rosario et al. [70]. However, any other method able to derive similarity measures from a multivariate categorical data set could be used instead.

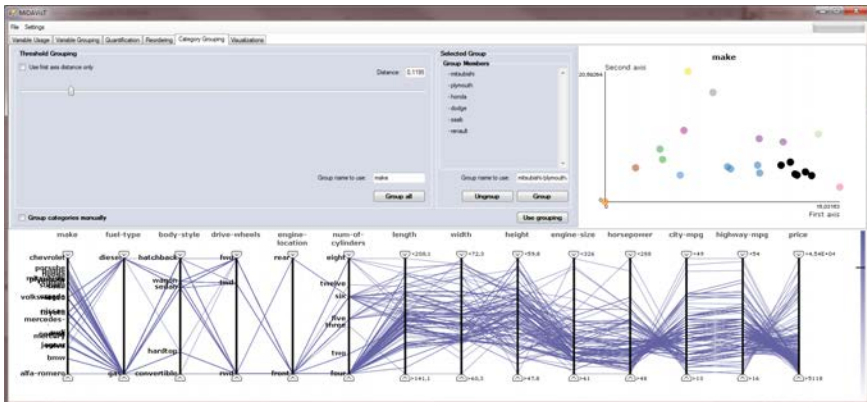
CA is normally applied to frequency tables representing distributions within purely categorical data sets. To perform quantification based on the underlying structures in a mixed data set, relationships and distributions among numerical variables need to be included in the frequency table. A straightforward method for doing this, as employed in the quantification method presented in this thesis, is to categorize the numerical variables and treat them as categorical while applying CA. Through this the numerical representations used for categories will be based not only on structures within the subset of categorical variables but on structures within all variables of the mixed data set. The most important aspect of categorization in this context is to categorize the numerical data such that it reflects relationships among them. This is approached in the work presented in this thesis by employing a  $K$ -means clustering algorithm [63] on the subset of numerical variables. The derived clusters are then treated as categories of a categorical variable and their frequency information is included in the contingency table prior to applying CA. This approach was later also employed by Yang et al. [97]. The presented method furthermore provides an alternative interactive categorization controlled by the analyst. Further details on clustering and categorization are available in paper I.

As previously mentioned the algorithmic quantification is the initial part of the proposed quantification method and acts as guidance and suggestion of similarities between categories. A range of interactive methods allowing manual modification of the quantification succeeds the automated analysis. Within an interactive environment, utilizing parallel coordinates and scatterplots, the analyst can modify the numerical representations and merge groups of categories using intuitive interactive methods, such as dragging and dropping categories along the axes of parallel coordinates and rectangle selection in a scatterplot for selecting category groups to merge. The interface for manual modification is displayed in figure 3.1(a). The top part of the window includes the interface for manual modification of numerical representations where the categorical variables are displayed using parallel coordinates along with suggested quantifications as represented by category names. Within this view the analyst can interactively modify the quantification based on preferences and domain knowledge. Figure 3.1(b) includes the interface for merging of groups of categories. Through a scatterplot where categories of a selected variable are displayed and laid out based on the similarities identified through quantification, the user can interactively select groups to merge. The interface also provides automated extraction of similar groups based on a similarity threshold.

Papers II and III present a full interactive system, called MiDAVisT (Mixed Data Analysis Visualization Tool), which utilizes the interactive quantification method presented in paper I and extends it with an interactive environment for analysis of the quantified data, displayed in figure 3.2. Within the interactive environment of MiDAVisT, common methods for numerical data analysis are employed, including visualization methods such as scatterplot matrix, table lens and parallel coordinates, and algorithmic analysis methods such as clustering and correlation anal-



(a) The interface for interactive modification of suggested quantifications. In the top view, which displays the categorical variables and suggested quantifications, the quantification may be interactively modified by for instance dragging and dropping categories along the axes.



(b) The interface for merging of groups of categories. A group of similar categories is selected and highlighted in black.

Figure 3.1: Interfaces for interactive modification of quantification results. The bottom view of the interface displays the full data set with suggestions of numerical representations for categories, as represented by category names along the categorical axes.

ysis. The usefulness of MiDAVisT is demonstrated, in the papers, through two case scenarios. In paper II a horse colic data set from the UCI Machine Learning Repository [27] is analysed. The data set contains both categorical and numerical variables including information on different symptoms, treatment and outcome for horses that have been treated for colic. The case scenario demonstrates how the presented quantification aids in identifying categories of horses having more similar or less similar symptoms, and how explorative visual analysis following quantifi-

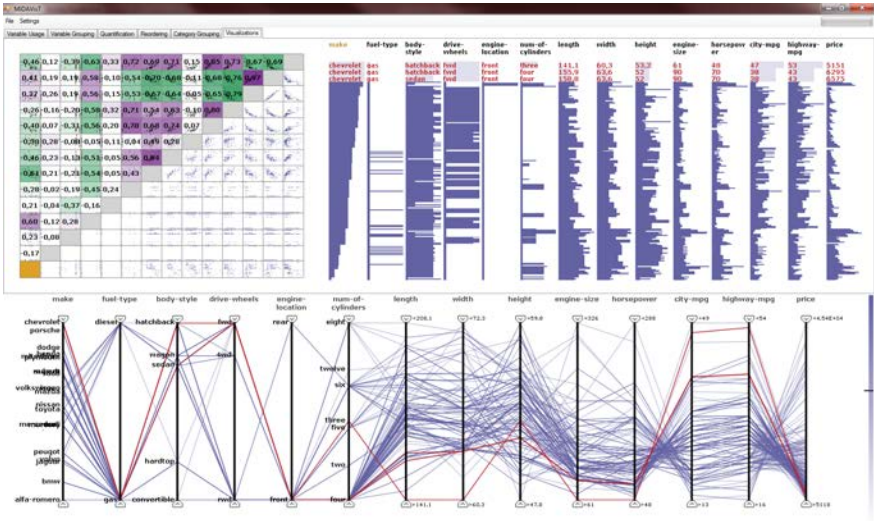


Figure 3.2: The interactive environment for explorative analysis of the quantified data set, including common visual representations and analysis methods designed for categorical data. Here displaying an automobile data set including six categorical and eight numerical variables. Three cars are selected and highlighted in red in the table lens and parallel coordinates, and in the top left part of the scatterplot matrix pairwise correlations are represented by coloured cells.

cation makes it possible to identify symptoms relating to the outcome of the treatment. In paper III a geospatial data set from the VAST 2008 *Migrant Boat* Mini-Challenge [38] is analysed using MiDAViT. This data set also includes a combination of categorical and numerical variables, and contains information on migrant boats departing from a fictional island during three years. Through quantification and subsequent visual analysis, changes over the years in terms of encounter coordinates and interdiction are identified, as well as a relationship between interdiction and encounter coordinates. The case scenario also demonstrates how category frequency, which is a common feature of interest in categorical data analysis, can be explored within the interactive environment of MiDAViT through sorting of rows in the table lens.

Subsequent to publication of papers I through III the MiDAViT tool has been extended with additional functionality in terms of selection of variable subsets for data sets including a variable classification. Through an interactive interface, as displayed in figure 3.3, the variables of the data set can be subdivided into a tree-structure based on one or several variable classifications. The analyst may then select a subset of variables to quantify and analyse through a simple tree-visualization, facilitating exploration of data sets including a larger number of variables, which may be hard to explore using traditional visual representations.



## 3.2 EVALUATION OF CATEGORICAL DATA APPROACHES

Methods for visual analysis of categorical data can be divided into two main approaches; the first being to utilize visualization methods designed for categorical data, hereinafter referred to as CatViz, and the second to employ a quantification approach and visualize the data using methods designed for numerical data analysis, hereinafter referred to as QuantViz. While working with the quantification method and MiDAVisT system, described in the previous section, the lack of comparison and evaluation of the two approaches became obvious. As an initial attempt to address this deficiency a performance evaluation comparing two visualization methods for categorical data is presented in paper IV of this thesis. Most CatViz methods are designed based on category frequency, meaning the number of data items belonging to one or a combination of categories, often represented by size. Visualization methods for numerical data are, on the other hand, commonly based on numerical similarity, or distance, between data items. Hence, the two main approaches to visual analysis of categorical data are to some extent intrinsically different since they represent very different aspects of the data. We cannot assume that all methods are equally useful within the same task contexts, and by comparing the two approaches we may obtain guidance as to which method to use and when to use it. Evaluations of CatViz methods are, however, relatively unusual. Stasko et al. [80] compare the performance of two visualization methods usable for categorical data but only within the context of hierarchically structured data. Whilst some evaluations comparing quantification methods have been carried out most are not performed within a visualization context. The following sections will describe the objectives and results of the evaluation presented in paper IV which compares the performance of the CatViz and QuantViz approaches.

### 3.2.1 OBJECTIVE

The evaluation carried out and presented in paper IV was intended as an initial attempt to compare various aspects and methods of the two main approaches to visual analysis of categorical data. One of the objectives of the paper was to put focus on the lack of evaluations within the area of categorical data analysis, especially in terms of comparing the performance of CatViz and QuantViz approaches, and to provide an example of how the two approaches may be compared despite their differences. Another objective was to carry out a user study to compare the performance of two visualization methods, each representing one of the main visualization approaches to categorical data analysis, within the context of typical data analysis tasks. The aim with the user study was to provide an initial guidance as to which method renders the highest performance within a specific task context. To summarize, the main goals of paper IV were:

- to compare the performance of methods representing the main approaches of categorical data visualization;
- to provide user guidance by comparing performance within a task context;
- to put focus on the lack of evaluations in the area.

### 3.2.2 RESULT

This section will briefly describe the evaluation and performance results. A more detailed description of it can be found in paper IV, including full motivation of design decisions, methods used and analysis of results. The study presented in the paper can be defined as a controlled experiment comparing two approaches to visual analysis and is, hence, related to Plaisant's [66] evaluation areas of controlled experiments comparing design elements and controlled experiments comparing two or more tools, as described in section 1.2.

Prior to designing the study a range of CatViz and QuantViz methods were reviewed with the purpose of selecting good representatives for each. In terms of CatViz methods, no formal comparison has been made between the majority of methods and it was outside the scope of the paper to perform such a comparison. Thus, the choice of method to use was based on a subjective opinion regarding which method would be least affected by data set size. The choice fell on Parallel Sets [58] which theoretically should be able to display approximately as many variables as parallel coordinates. In terms of quantification, Correspondence Analysis was used to obtain numerical representations, based on the fact that it has been suggested several times as a quantification method, that it is a generic method not limited to specific task types and that, through a previous evaluation [22], it has been picked out as a good quantification method.

A major concern while preparing the study was to design it such that it enables fair comparison between the two methods, while not removing basic differences or limiting the benefits of any method. Hence, a number of design decisions were made, aiming to remove differences that are not fundamental properties of the visual representations, while retaining differences in terms of fundamental properties which may be the source of one method's advantage over the other. This was addressed for instance through the decision to remove all interactivity by using screen shots of the visualization methods, since interactive features, such as the possibility of filtering or brushing in parallel coordinates, are not fundamental properties of the visual representation, although commonly occurring. Furthermore, if comparing interactive methods it would be difficult to distinguish between performance results due to interactivity and performance results due to visualization approach.

A second decision was to use a single view for CatViz whilst using a multiple view setup of three common multivariate visualization methods (scatterplot matrix, table lens and parallel coordinates) for QuantViz. This may appear as unfair but is based on the aim of retaining fundamental differences and not limiting the benefit of any method. Using multiple views increases the perceptual burden since information from several sources need to be coordinated, it may however be beneficial if the views complement each other. In terms of CatViz, most methods are based on category frequency and generally have the same strengths and weaknesses. Thus, using a single view was considered most beneficial for CatViz. Visualization methods designed for numerical data are, on the other hand, more diverse in their representations and, thus, have different strengths and weaknesses which may complement each other. Based on this, using multiple views was considered most beneficial for QuantViz. The selection of visual representations to use for QuantViz was based on the commonness of the methods and their ability to represent various aspects of the data. For instance a table lens complements parallel coordinates in its ability to display category frequency, while parallel coordinates straightforwardly displays

multivariate relationships and similarities. Additionally the scatterplot matrix is able to show all pairwise correlations concurrently, which is not possible using parallel coordinates or table lens. One alternative solution to comparing a single view setup with a multiple view setup could be to extend one of the visual representations for numerical data with additional features which would make the use of other representations redundant. One example would be to extend parallel coordinates with features for displaying category frequency, such as adding histograms along the axes or spreading polylines according to frequency. This would however not agree with the original idea of comparing a CatViz method with a QuantViz method, since it is questionable whether the extended visual representation would actually be a method for numerical data or whether, due to the extensions, it has been converted into a CatViz method.

Similarly to utilizing basic visual representations, the tasks to be carried out were selected to represent basic elements of typical data analysis tasks. Two main task types were selected, including frequency related tasks and similarity related tasks. Frequency related tasks are typical for categorical data and were, in the study, represented by two tasks; identification of the most common category within a specific variable and identification of the combination of categories that is most common in two adjacent variables. For data analysis in general, a common task is to identify patterns, and most patterns can be defined in terms of similarity. Hence identification of similarity patterns was considered a highly relevant task to focus on. In the same way as with frequency tasks, the similarity related tasks were represented by two tasks; identification of the categories within a specific variable that are most similar to each other and identification of the variable which is most similar to a specific variable.

The evaluation was carried out through a user study with fifteen participants. The study included four phases, making sure that all participants performed both kinds of tasks with both kinds of visualization approaches. Within the study, performance was defined as a combination of response time and accuracy when performing the tasks, which were recorded for all participants. Additionally, questionnaires were used to receive the subjective opinion of the participants in terms of preference of method and perceived difficulty of performing tasks.

Figure 3.4 presents an overview of the performance results for the four phases. General conclusions drawn from analysing the results were that the QuantViz approach was the preferred method for performing similarity related tasks, and was also the method yielding best performance for this type of task, as shown in figure 3.4. Correspondingly the CatViz approach was most preferred and gave best performance for the frequency related task. However, when separating the tasks into questions including one variable and questions including two variables, as displayed in figure 3.5, there was a noticeable performance difference for frequency tasks using the QuantViz approach. The performance when answering questions regarding the most common category within a specific variable was almost as high using the QuantViz approach as when using the CatViz approach, whereas the performance when answering questions regarding the most common combination of categories within two variables was noticeably worse using the QuantViz method, a difference which was also mentioned by some of the participants. Full details of the results and analysis can be found in paper IV.

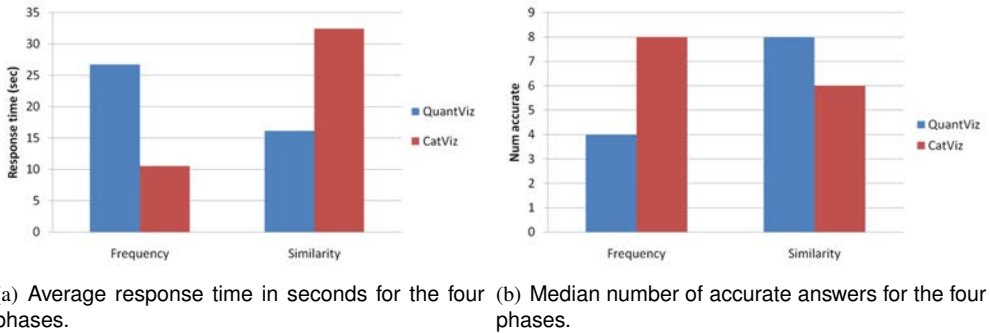


Figure 3.4: Performance results for the QuantVis and CatViz approaches when performing frequency and similarity related tasks.

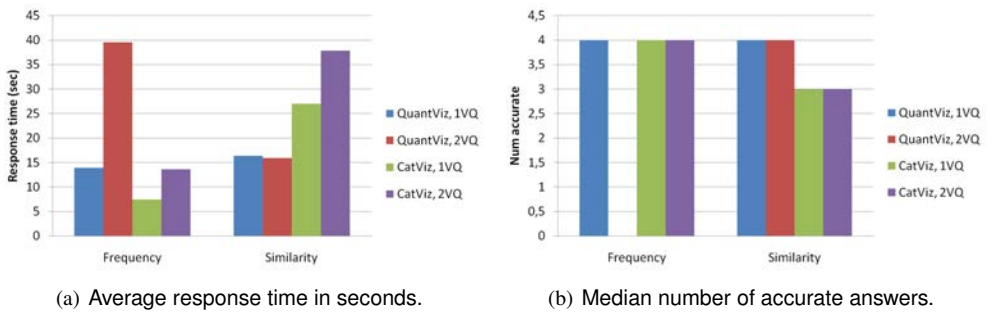


Figure 3.5: Performance results for the QuantVis and CatViz approaches when performing frequency and similarity related tasks, when the tasks are separated into questions including one variable (1VQ) and questions including two variables (2VQ).

### 3.2.3 SUMMARY OF CONTRIBUTIONS

The paper described in this section contributes a formal evaluation of two methods representing the main approaches to visual analysis of categorical data. The evaluation is designed as an experimental study comparing the two methods in the context of two basic data analysis tasks and through this provides guidance as to which method may be most useful for which task. By approaching the subject of comparing CatViz and QuantViz methods the paper puts focus on a deficiency in the area of categorical data analysis, and on the importance of addressing this issue. The study is not intended to find the ultimate solution as to how categorical data may be best analysed but as an initial attempt to compare the two approaches and as an encouragement for further evaluations in the area.

### 3.3 QUALITY BASED DIMENSIONALITY REDUCTION

High dimensional data, sometimes including hundreds or thousands of variables, are a major challenge for data analysis in general and visual analysis in particular, since both screen size and resolution as well as the human perceptual system limits the usability of traditional visual representations as the number of variables increases [23, 87]. This is often dealt with by employing dimensionality reduction prior to visualization. Paper V addresses the issue of high dimensionality and presents an interactive dimensionality reduction system where automated algorithms are combined with user interaction to provide efficient and user-controlled dimensionality reduction.

A large number of dimensionality reduction methods have been proposed, many of which are fully automated and efficiently reduce hundreds or thousands of variables into low-dimensional projections. Although many of these algorithms provide fast dimensionality reduction, enabling efficient analysis in various contexts, automated methods are limited in not letting the user influence the reduction during the process. The utilization of interactive visual representations in the reduction process may increase the insights gained, facilitate decision making and improve the effectiveness of analysis considerably. As an example, in the context of visualization the appropriate number of variables to keep in a reduced data set depends not only on the size and resolution of the display but also on the analysis task and the structure of the data. In a situation where patterns and relationships are unknown to the analyst prior to analysis, presentation of patterns identified by algorithmic methods during the reduction process may facilitate selection of an appropriate number of variables to retain in the reduced data set.

Different dimensionality reduction methods focus on different aspects and features of data, with the general aim of preserving structures within the data set while reducing it to a smaller representative subset of variables. As the data set is reduced a certain loss of information is unavoidable since variables are removed, here referring to information in terms of structures and details within data. The amount of information lost depends not only on the size of reduced and original data set but also on the structures within both the removed and retained variables. Which structures are of interest to preserve depends on the analysis task and for some tasks, especially in the context of exploratory analysis, several structures may be of interest at the same time. Most dimensionality reduction methods, however, focus on preserving only one structure at the time. Although applying several reduction methods separately may enable analysis based on various structures, it would not enable extraction of a single subset of the most interesting variables based on more than one structure. This may be of interest in, for instance, the context of explorative analysis where structures and relationships within the data are unknown to the analyst. Additionally, even though several structures may be of interest, their relative importance, or level of interestingness, may vary depending on the task as well as the patterns and relationships present within the high dimensional data set.

The order of variables in a multivariate data set has a large impact on our ability to perceive structures [3]. As the number of variables displayed in a visual representation increases the variable ordering becomes more important and the benefit of employing automated variable ordering grows as manual re-ordering becomes more and more inefficient. Hence the investigation of structures within a high dimensional data set may often benefit from applying variable ordering algorithms subsequent to dimensionality reduction.

### 3.3.1 OBJECTIVE

The primary objective of the work presented in paper V was to develop a straightforward method for reducing the dimensionality of a high dimensional data set based on several structures, aiming to extract variable subsets in a way such that multiple patterns are simultaneously preserved. The dimensionality reduction method was to be implemented in a system where the analyst is able to influence the reduction and make decisions based on structures identified during analysis. The method should also be incorporated in an interactive environment where visual representations and algorithmic aids, such as variable ordering, are provided. Aiming to facilitate the gaining of insights, regarding importance of individual structures, and decision making in terms of subset size, as well as providing features for basic exploration and identification of structures. To summarize, the main objectives of paper V were to develop a dimensionality reduction method and system which

- reduces dimensionality based on several structures;
- includes possibilities for the analyst to interactively influence the reduction;
- utilizes visualization methods to aid decision-making and pattern identification.

### 3.3.2 RESULT

The aim of performing dimensionality reduction based on several structures at once was addressed in paper V using a weighted sum of quality metrics, where each quality metric represents a specific structure or pattern. In the paper correlation, cluster and outliers were used as example metrics. A quality value was extracted for each metric and each variable, representing the variables involvement in the corresponding structure. The more involved in the structure, the higher the quality value. For the three example metrics high values were assigned to variables strongly correlated with several other variables; to variables that were part of multivariate clusters with high density covering a relatively large part of the data set; and to variables including multivariate outliers with few neighbouring data items. An overall measure of interestingness was obtained for each variable by summarizing their quality values, using weight values to define the relative importance of each metric. Dimensionality reduction was then performed by ordering the variables according to level of interestingness and retaining only the top  $K$  variables with highest overall interestingness, where  $K$  is a number selected by the analyst.

The automated algorithm, as described above, was included as part of an algorithmically guided interactive system for user-controlled dimensionality reduction, where the first step was the extraction of quality values. This was followed by a visual representation of the loss of quality metrics, or information, against the number of variables to retain in the data set after reduction, displayed in figure 3.6. The loss of each quality metric, which is defined as the ratio between the summed quality values of the removed variables and the sum of the quality values of all variables, is represented by a line in a line graph, where the horizontal axis represents number of variables to retain and the vertical axis represents the percent of the overall metric structures that is lost. An additional line represents the total loss of all quality metrics and a red vertical line is positioned

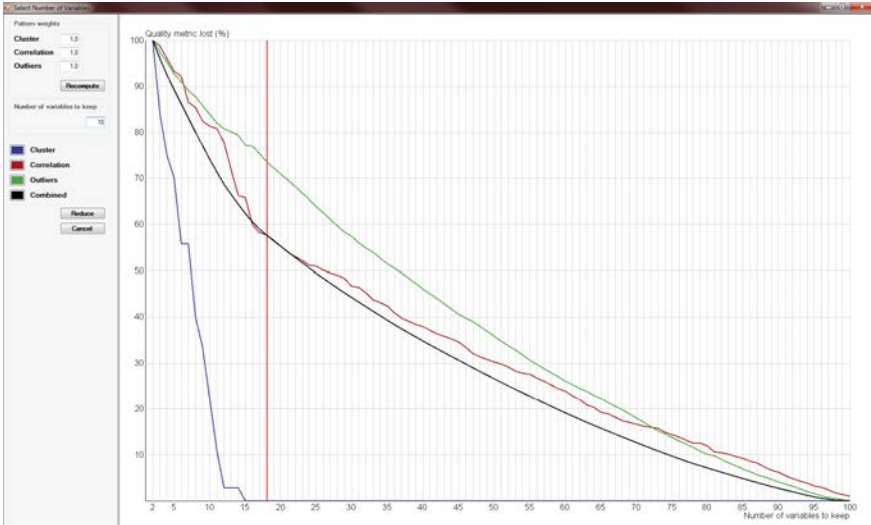


Figure 3.6: An interactive display presenting the trade-off between loss of quality metric structures (represented by the vertical axis) and number of variables to retain in the reduced data set (represented by the horizontal axis). The quality metrics, as well as a measure of overall variable interestingness, are represented by lines in the graph and the currently selected number of variables to retain is represented by a red vertical line.

at the currently selected number of variables to retain. The quality metric loss display enables investigation of the trade-off between number of variables to retain and loss of quality metrics. Through this it acts as a guidance in the selection of an appropriate number of variables to retain and aids the analyst in making an initial estimation of the relative importance of the metrics. For instance, in figure 3.6, where the blue line represents the cluster metric, it is apparent that only a relatively small number of variables are involved in cluster structures, since the slope of the line is steep and the loss of cluster structure quickly approaches zero as the number of variables to retain increases above ten. While the green line, representing outlier structures, displays a close to linear relationship between loss of outlier structures and the number of variables to retain, indicating that the outlier structures within the data set are more evenly distributed among the variables.

In the interface of figure 3.6 the analyst is able to interactively assign weights indicating the relative importance of individual metrics to the overall measure of interestingness. The interestingness of the variables is immediately re-computed as the weights are modified. The variables are re-ordered according to their new level of interestingness and the selected subset of the  $K$  most interesting variables is updated accordingly. Figure 3.7 displays the top nine variables as different metric weights are used for the same synthetic data set, originally including 100 variables. In the top view the cluster metric is assigned five times as high a weight as the other metrics, whereas correlation is assigned a five times higher weight in the bottom view.

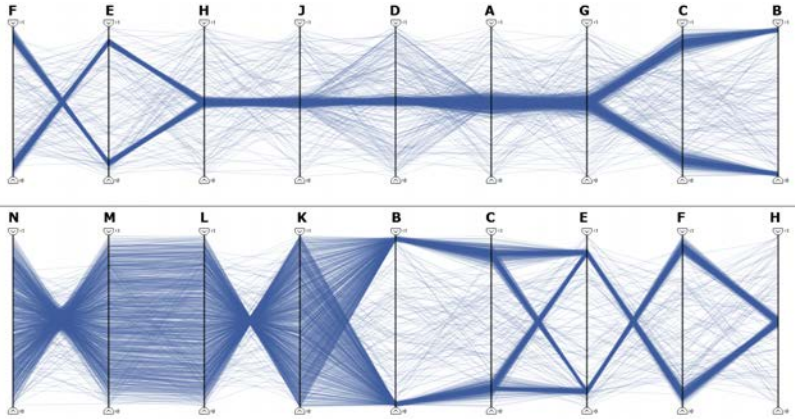


Figure 3.7: A synthetic data set including 100 variables reduced to the nine variables with highest level of interestingness when the cluster metric is assigned larger weight (top) and when the correlation metric is assigned larger weight (bottom).

As a final step in providing a dimensionality reduction environment suitable for exploratory analysis, the system presented in paper V includes an interface for initial visual analysis of the reduced data set. The interface provides various visual representations and automated variable ordering, to support the diverse range of possible analysis tasks following a dimensionality reduction and to facilitate identification of various patterns. The variable ordering algorithms aim to enhance structures in the data to make patterns, in terms of available quality metrics, more easily perceived. Two different variable ordering algorithms are employed, details of which can be found in paper V. The first algorithm was inspired by the variable ordering suggested by Artero et al. [4] and is a variable ordering based on relationships between pairs of variables. It is therefore suitable for enhancing, for example, correlation patterns. The second algorithm aims at enhancing multivariate patterns, taking patterns within subgroups of variables into consideration while ordering variables, aiming to keep together groups of variables that are involved in the same multivariate structures. It is therefore suitable for enhancing structures such as clusters and outliers. Figure 3.8 displays examples of the variable orderings as presented in the case-scenario described in paper V. The figure displays a subset of 15 variables ordered according to clusters in the top view and according to correlation in the bottom view. The different variable orderings enables identification of varying patterns, such as the low-dimensional cluster structures visible within the variables marked B, C, D, E and F in the top view. Similarly, the correlation patterns between the variables marked G, H and I in the bottom view are hard to distinguish in the top view where the variable marked A corresponds to the variable marked I in the bottom view, and where the variable marked F corresponds to the G variable.

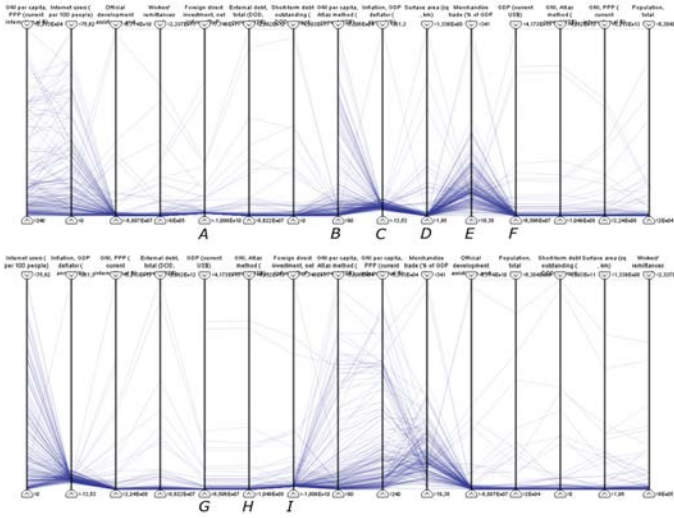


Figure 3.8: A variable subset of 15 variables ordered to enhance multivariate cluster structures (top) and correlation patterns (bottom).

### 3.3.3 SUMMARY OF CONTRIBUTIONS

The paper presented in this section contributes a dimensionality reduction method which preserves several types of structures simultaneously by combining a set of quality metrics into one overall measure of variable interestingness. The overall measure is used as a filter for selecting a subset of the most interesting variables. The reduction method is incorporated into an interactive system where the relative impact of each metric upon the overall interestingness is controlled by the user through weight values, and where dimensionality reduction is visually guided by the loss of quality metric patterns in the selection of an appropriate number of variables to retain. The system also contributes two variable ordering algorithms, based on pairwise relationships and multivariate patterns, which enhance perception of specific patterns and, through that, facilitate pattern identification.

## 3.4 VISUAL EXPLORATION OF HIGH DIMENSIONAL DATA

The work described in section 3.3 presented a method for combining several quality metrics into an overall measure of variable interestingness, usable as a threshold for selecting a subset of the potentially most interesting variables in a high dimensional data set. Feedback has indicated that the simplicity of utilizing a single measure of interestingness is found advantageous by end-users. However, although a weighted sum of quality metrics may be useful for dimensionality reduction, the method described in section 3.3 is subject to some limitations which are addressed in papers VI and VII.

Firstly, when using a weighted sum of metrics it may be difficult to estimate an appropriate set of weights, since the analyst has to consider the relative importance of a set of statistical concepts. This may specifically be true in the context of explorative analysis where the structures of interest may not be known to the analyst prior to analysis. Secondly, by solely using a single measure of interestingness as a threshold for dimensionality reduction, information regarding the values of the original metrics is lost. Although the loss of quality metric display (figure 3.6) provides some information on the existence of structures within the high dimensional data and selected subset, the user is unable to explicitly examine the contribution of each metric to the selected variable subset, and also does not provide an overview of relationships between structures in the full high-dimensional data set. Thirdly, while weight values and display of trade-off between quality metric loss and number of variables retained enables user influence as well as guidance in the dimensionality reduction process, the method still applies a semi-automated approach as it automatically retains a subset of the most interesting variables, which limits the involvement of the user in the reduction process. Moreover, it is quite possible that a variable assigned a lower value of interestingness may still be of interest, for instance due to significant involvement in specific structures or due to some property known by the analyst. Furthermore, additional reduction methods, such as merging of similar variables or thresholding on a subset of metrics, may be valuable in an explorative analysis context. Overall, a more flexible and interactive dimensionality reduction may be desirable, allowing for more user influence and more flexibility during the dimensionality reduction process.

The work presented in papers VI and VII originates from the idea of quality guided dimensionality reduction using a combination of quality metrics, as described in section 3.3, and focusses on overcoming the limitations of the previous system and on providing a more flexible and interactive environment for exploratory dimensionality reduction. The methods presented in papers VI and VII are closely related in terms of being based on the same ideas. However, paper VI presents a system designed for exploration of high dimensional microbial data, providing a range of features relevant within the field of microbiology, whereas paper VII presents a generic system for exploration of high dimensional data utilizing measures of general statistical properties.

### 3.4.1 OBJECTIVE

The main objective of the work presented in papers VI and VII was to develop a flexible and interactive technique for explorative dimensionality reduction aiming to continue the work on combining several quality metrics into an overall measure of interestingness but, this time, using a method that does not require input from the analyst in terms of weight values. Moreover, a primary goal was to enable a dimensionality reduction which was fully controlled by the user, using quality metrics and an overall measure of interestingness solely as a means for guidance and including functionality for manually including or excluding variables. To provide guidance, the relationships between the individual quality metrics and the overall measure of interestingness were to be visually displayed. Through the display providing overview of structures within the high dimensional data set and enabling examination of the impact of individual metrics upon the overall interestingness. Another goal was to enable flexible dimensionality reduction by incor-

porating the dimensionality reduction method into interactive systems where various reduction options should be provided. The systems were also to allow interactive examination of both the full high-dimensional data set and the reduced variable subsets.

In addition to this, the main objective of paper VI was to provide an interactive system for visual analysis of microbial populations, which by nature are high-dimensional. The aim being to combine information visualization methods with metrics and features commonly used within microbiology and, through this, provide an interactive analysis environment which complements existing tools and acts as a means for developing new and domain relevant insights. The objective of the work presented in paper VII was, on the other hand, to develop a generic system for exploration of high dimensional data sets, aiming to provide a system flexible enough to support diverse data analysis tasks and analysis of data from various domains. To summarize, the objectives of the work presented in papers VI and VII was to develop a dimensionality reduction technique and interactive systems which:

- contribute a single measure of interestingness based on several quality metrics, not requiring user input;
- enable interactive dimensionality reduction fully controlled by the user;
- provide visual overview of structures and relationships within the data;
- enable flexible dimensionality reduction through various reduction possibilities;
- supply interactive features and visual representations for examination of selected variable subsets.

### 3.4.2 RESULT

Papers VI and VII present two systems for dimensionality reduction and explorative analysis of high-dimensional data sets, based on an interactive dimensionality reduction method aiming to achieve the goals described in section 3.4.1. The objective of providing a single measure of variable interestingness based on several structures, overcoming the limitations of utilizing a weighted sum, was addressed by employing a ranking algorithm based on the non-domination principle [33]. Similarly to the approach of weighted sums, a set of quality metrics is used and a quality value is extracted for each quality metric and each variable in the data set. The quality values are then used as input vectors to the ranking algorithm which assigns rank to each variable such that variables with high quality values are assigned high ranks and ensuring that all variables with the same rank have a level of equivalence in their metric profiles. Further details on computation of rank are found in the papers.

Differently from the initial approach of paper V, the algorithmic analysis of the reduction method presented in papers VI and VII, including extraction of quality metrics and assignment of rank, was meant to act as guidance rather than to solely provide a measure for automated dimensionality reduction. This was approached by visually presenting overviews of quality values and ranks, in paper VI through the *Rank and Quality* (RaQ) view using parallel coordinates, as

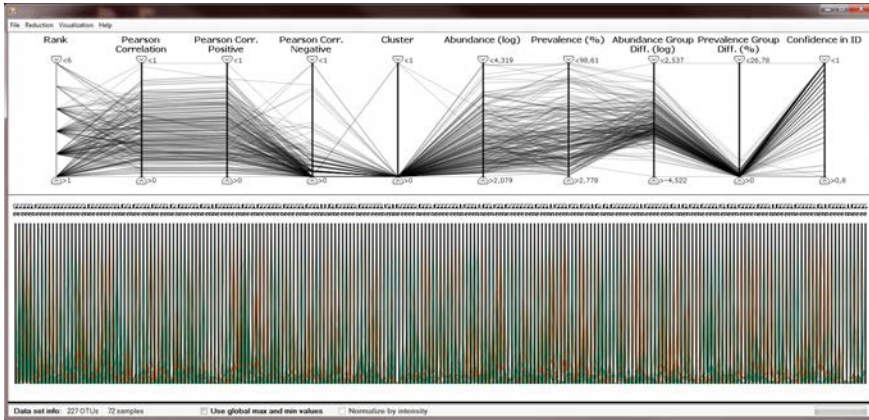


Figure 3.9: The visual environment of MicrobiVis, the system for explorative dimensionality reduction presented in paper VI, displaying a data set including 227 variables. The top view displays variable rank and quality metrics as axes and variables as polylines in parallel coordinates, and the bottom view displays the full high dimensional data set.

displayed in the top view in figure 3.9, and through a combination of the RaQ view and a PCA plot in paper VII, as displayed in the top part of figure 3.10(a). The axes of the parallel coordinates in the RaQ view represent rank and quality metrics and the polylines represent the variables of the high dimensional data set. Through this an overview of structures within the full data set is presented to the analyst, as well as providing possibilities of exploring relationships between individual quality metrics and the overall measure of variable interestingness. The RaQ view not only provides an overview of structures, it also enables interactive and fully user-controlled dimensionality reduction through filtering along the axes, immediately removing from the data all variables outside of the filter ranges. Thus enabling dimensionality reduction based on an overall measure of interestingness as well as on individual quality metric axes. The filtering in the RaQ view is linked to a visual representation of the high dimensional data set in the bottom view, which displays the selected set of variables and is interactively updated while filtering in the RaQ view. The appropriate number of variables to retain when reducing dimensionality may vary from case to case, depending on both screen size and structures within the data set. While filtering in the RaQ view the analyst will be able to perceive when structures become visible and can be clearly viewed with the currently used visual representation and available screen size. The analyst is hence guided in the selection of the number of variables to retain by the visual display of the selected variable subset.

Additionally, one objective to address was the issue of specific variables being more interesting than indicated by their rank. This may, for instance, be due to certain properties known to the analyst but unidentifiable by an algorithm, or due to involvement in specifically interesting or unusual patterns. This was dealt with by supplying various features for interactively selecting individual variables to be unaffected by dimensionality reduction and, hence, retained despite being outside of the filtering thresholds of the RaQ view. Correspondingly, individual variables

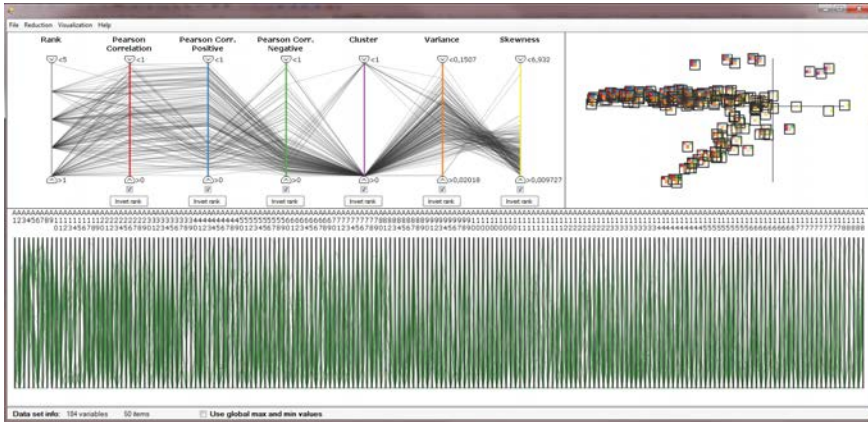
may be less interesting than indicated by their rank which was addressed through similar features for removing variables even though they lie within the reduction thresholds.

#### 3.4.2.1 EXPLORATION OF MICROBIAL POPULATIONS

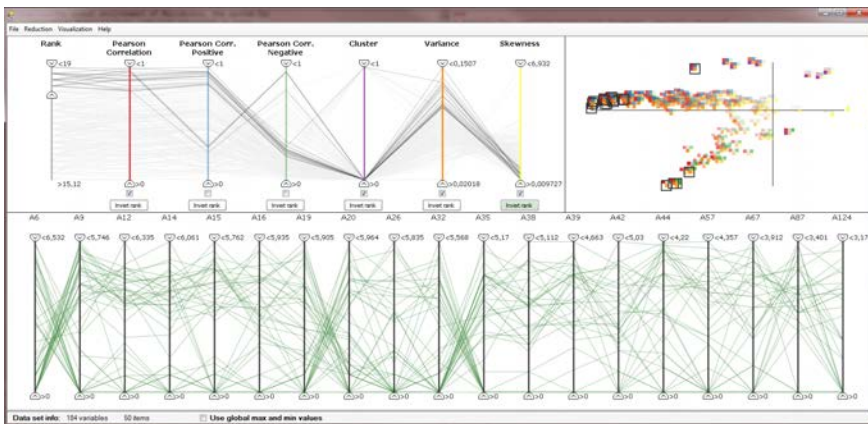
In addition to the common objectives described above, the two papers focus on exploration of high dimensional data within different application areas. Paper VI presents MicrobiVis, a system for visual analysis and interactive exploration of microbial populations, developed in collaboration with industrial data scientists and microbiologists. In the context of data analysis where data are gathered through studies of sampled microbial populations, samples may be considered as data items and microbial species as variables. Such studies generate high-dimensional data where a single sample may include hundreds of different microbial species. Hence, usable analysis methods for high dimensional data are important for gaining insights and generating hypotheses. Furthermore, to acquire domain relevant insights focus needs to be put on tasks, structures and features of specific interest in the field of microbiomics, rather than solely utilizing features useful within a generic domain. On the other hand, by employing analysis techniques not traditionally used within the domain, new and different insights may be reached. Based on this, the MicrobiVis system includes a combination of generic information visualization methods and domain specific features, making it possible for the users to gain new insights while thinking in a biological context.

More specifically, MicrobiVis utilizes a range of domain specific quality metrics, as described in paper VI, along with generic statistical metrics which were considered to be of relevance within the domain context. An example being correlation which may indicate cohabitation of microbial species since a strong positive correlation implies that a high counts of one species will also mean a high counts of the other species. MicrobiVis also includes a domain relevant visual representation in the form of a phylogenetic tree [91], which is a hierarchical structure based on similarities and differences between various biological species, as displayed in the bottom view of figure 3.11. In the phylogenetic tree the selected microbial species, as visible in the lower parallel coordinates display, are highlighted in red. Aided by this the analyst may link variable subsets, selected due to specific structures and identified through quality metric analysis, to phylogenetic relationships as represented in the tree. Additionally, to provide domain relevant and flexible dimensionality reduction MicrobiVis includes reduction utilizing a hierarchy of organisms based on the biological classification system [91], where species is the basic unit having an associated taxonomy of genus, family, order and so on. As an example of biological classification, humans are classified as *Homo Sapiens* at species level, *Homo* at genus level, *Hominidae* at family level and *Primates* at order level. The family of Hominidae also includes chimpanzees, gorillas and orangutans.

For illustration in the paper three taxonomic levels of interest to the microbiologists were utilized: phylum, genus and species level, where species and genus are the two lowest levels and phylum provides a higher level classification. Thus, in the context of data representation in MicrobiVis, stepping up one level in the taxonomic hierarchy from species to genus level would correspond to representing a group of species by a single representative genera, hence automatically providing a biologically relevant dimensionality reduction. Through this the analyst may



(a) Displaying the full high-dimensional data set.



(b) Data set reduced to a smaller subset through filtering on rank (leftmost axis) in the top parallel coordinates display.

Figure 3.10: The visual environment of the system for explorative dimensionality reduction presented in paper VII, here displaying a data set originally including 184 variables. The top left view displays variable rank and quality metrics as axes, and variables as polylines, in parallel coordinates. In the top right view the variables are represented by glyphs in a PCA plot and in the bottom view the data set is displayed.

interactively step between different taxonomic levels, facilitating exploration at different levels of detail and enabling identification of patterns that may be more visible at a certain level. Examination of a subset of species belonging to one or a set of specific genera may also be of interest, for instance to explore whether a pattern identified at genus level exists for all species belonging to that genera. To support this, MicrobiVis includes features for rapid selection of groups of species through selection of one or a set of genera in the application menu. All species

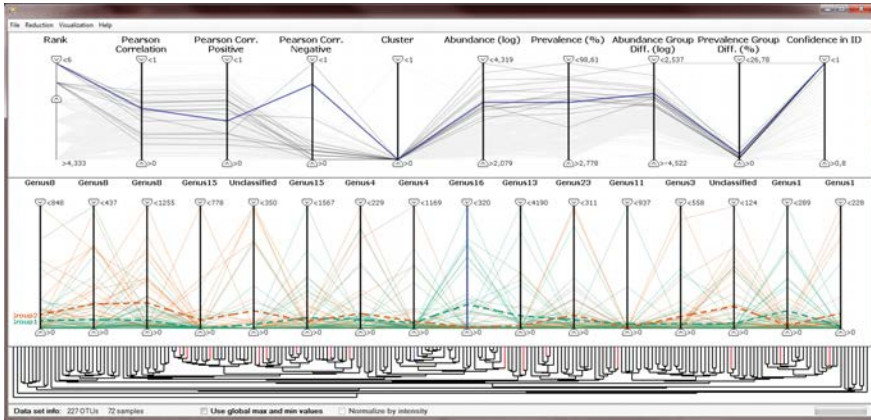


Figure 3.11: The visual environment of MicrobiVis including a visual representation of a phylogenetic tree in the lowest of the three views. Variables retained in the reduced data set, as displayed in the bottom parallel coordinates, are highlighted in the tree. In the lower of the parallel coordinates displays, samples are coloured according to a sample classification and additional polylines representing class averages are displayed.

not belonging to the selected genera are removed from the visual display, hence providing an additional dimensionality reduction method based on taxonomic classification.

Sampled microbial population data often include sample classifications, for instance in terms of varying sample sites or subject groups. Thus, the comparison of sample classes may be highly relevant. An example being the comparison of the microbial ecology of the intestines of subjects being treated with different drugs. This is addressed in paper VI by including quality metrics measuring class difference and through visual separation of samples belonging to different classes using colouring and visual representation of class averages within the selected subset of variables, as displayed in figure 3.11 where the samples are separated into two classes.

Feedback from end-users, as presented in more detail in paper VI, indicates that the MicrobiVis system complements available analysis methods well in terms of providing an environment where hypotheses may easily be created through initial data exploration and compared against output of other microbiomics tools. Through the interactive dimensionality reduction using the filter sliders of parallel coordinates, an initial interesting view can quickly be generated from which the user can interactively focus the analysis, which was much appreciated. Furthermore, due to the structure of MicrobiVis, microbiologists found it easy to think in a biological context during exploration, rather than having to adapt to a less familiar model. The system also enabled for the analyst to interactively step their view between different taxonomic levels, which seems to be less easily achieved in other tools.

### 3.4.2.2 A GENERIC APPROACH FOR HIGH DIMENSIONAL DATA EXPLORATION

The MicrobiVis system was specifically designed for exploration of microbial populations. However, many of its features may also be usable in a more generic analysis context. Originating from this, a second system was developed, as presented in paper VII, based on the same fundamental principles of algorithmically guided and interactive dimensionality reduction. Since the second system aims at a more generic analysis domain it does not include any features specific to the field of microbiology, but has been extended with several features to provide a flexible environment for explorative dimensionality reduction. Furthermore, the quality metrics utilized in paper VII are based on general statistical properties. In terms of ranking, the influence of a specific metric on the overall variable rank may be interactively modified by inverting the metric or by removing it from the rank computation. For an inverted metric, high rank is assigned to variables with low values for that metric.

In addition to the RaQ view, as available in both systems, the generic system also includes a second view, called the *glyph view*, for providing overview of patterns in the full high-dimensional data set. In the glyph view, displayed in the top right part of figure 3.10(a), each variable in the high-dimensional data set is represented by a glyph. The glyphs are made up of a set of coloured rectangles where each rectangle represents a quality metric and the opacity of the rectangles represents the corresponding quality value. The glyphs are laid out in a PCA plot where the quality metrics are the input vectors to the PCA. Through colours and layout the glyph view enables identification of variables which can be considered to be outliers, or groups of variables that are similar, in terms of quality metric profiles. Through this, the glyph view provides features for gaining insights which can guide the analyst and facilitate decision-making during the dimensionality reduction process.

In the context of generic high-dimensional data analysis, different methods for reducing the data may be useful, due to the wide variety of possible tasks. As in MicrobiVis, the primary dimensionality reduction method used in the generic system is performed by filtering along the axes in the RaQ view. Through this variables of less interest are removed from the high dimensional data set. However, in some cases it is quite possible that groups of highly similar variables, representing closely related attributes, may be better represented by a single variable. The system presented in paper VII addresses this by providing an additional window for exploration of groups of highly similar variables, as displayed in figure 3.12. Within this window groups of strongly correlated variables are automatically extracted, based on a correlation threshold set by the user. The relationships within the groups may then be visually explored through two visual representations, where a selected group of variables is displayed alongside a representative variable which replaces the group if it is merged into a single variable. Based on relationships identified by the analyst and domain knowledge regarding the variables' relationship in terms of attribute meaning, variables may be interactively removed from the group to merge. The analyst then decides which groups to merge and the variables of those groups are immediately replaced by their representative variable in all views of the system. As an example, the selected group of variables displayed in figure 3.12 includes three variables of a social science data set, from the left representing the percentage of divorced males, the percentage of divorced females and the total percentage divorced in different communities. The rightmost variable is the represen-

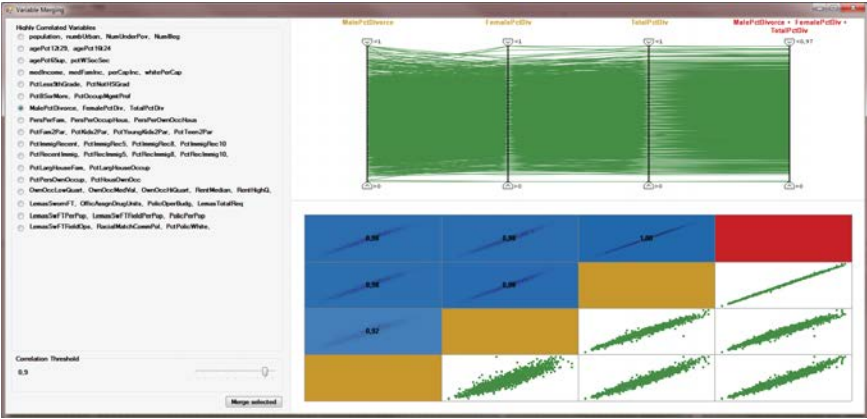


Figure 3.12: The window for merging of groups of highly similar variables. Variable groups are extracted based on pairwise correlation using a correlation threshold. A selected group is then visually displayed, together with its representative variable, using parallel coordinates and a scatterplot matrix where pairwise correlation values are displayed in the top left half of the matrix.

tative variable to replace them. Since the three variables are strongly correlated, with pairwise correlation values of 0.92 and 0.98, as visible from the cells in the upper left half of the scatter plot matrix, and since they are also closely related in terms of meaning, it may be useful to replace them with a single variable. Through the combination of automatically extracting groups of similar variables and displaying them visually, enabling examination of relationships within the variable groups, both the speed of algorithms as well as the domain knowledge of the analyst are made use of. Hence, the previously described interactive and explorative dimensionality reduction method is, in paper VII, complemented by an additional reduction method focusing on a different aspect of variable relationships, providing a flexible dimensionality reduction usable for various analysis tasks and suitable for data analysis within various domains.

### 3.4.3 SUMMARY OF CONTRIBUTIONS

Papers VI and VII contribute a method for explorative dimensionality reduction and examination of data sets including hundreds of variables, utilizing algorithmic and automated analysis methods for efficiency and as means for guiding the analysis. Concurrently the method supplies dimensionality reduction making use of the knowledge of a domain expert, for instance in terms of providing features for selecting variables to be excluded from the reduction process.

Through visual representation of quality metrics and rank, along with linked representations of variable subsets interactively selected through filtering in the RaQ view, the proposed method contributes a user-controlled dimensionality reduction visually guided by the quality metric profiles of variables. The visual representations also provide an overview of structures in the data set and facilitate decision-making during the analysis process by enabling interactive examination of both the high dimensional data set and selected variable subsets. They furthermore facilitate

selection of an appropriate number of variables to retain as patterns appear in the interactively updated display of the selected variable subset. The method also contributes flexible dimensionality reduction based either on a single measure of overall variable interestingness, on several quality metrics, or on individual metrics, as filtering in the RaQ view may be performed on various combinations of metric and rank axes. This enables examination of the impact of individual metrics, both on the overall interestingness and on the selected variable subset.

The method is incorporated into two systems designed for different task contexts. The systems provide additional flexibility by enabling dimensionality reduction based on biologically relevant classification in paper VI and through merging of groups of similar variables in paper VII. Furthermore they include various possibilities for interactively examining selected subsets of variables. Paper VI moreover contributes a system developed through an iterative design process driven by the requirements of the intended end-users.

# CHAPTER 4

## CONCLUSIONS

The research presented in this thesis has focused on the combination of algorithmic methods and interactive information visualization to facilitate explorative data analysis processes. More specifically, the primary focus of the work has been on the utilization of automated methods as a means of guidance. Aiding the analyst in decision-making and in the identification of potentially interesting patterns while, through visual representations and interactive features, providing the analyst control over which analysis paths to follow and abilities to modify the algorithmic result. In particular, this thesis has presented research where algorithmically guided visualization has been employed in the areas of categorical and high-dimensional data analysis. This section will provide a summary of the contributions, followed by discussions around conclusions drawn from the work and suggestions for future research directions.

### 4.1 SUMMARY OF CONTRIBUTIONS

The contributions of this thesis have been highlighted in chapter 3 and are described in detail in the included publications. Papers I through IV present methods focusing on categorical and mixed data sets, while papers V through VII present methods for analysis of high-dimensional data. The main research contributions of this work include:

- an automated method for quantification of categorical data which incorporates relationships from numerical variables in mixed data sets into the quantification process;
- a framework for including the automated quantification method in an interactive visual analysis pipeline, utilizing the approach of algorithmic guidance, enabling exploration and interactive modification of the automatically generated quantification;
- a user study evaluating the performance of the two main approaches for categorical data analysis within a task context;
- dimensionality reduction approaches aiming to preserve multiple structures by extracting a single value of variable interestingness by combining a set of quality metrics;

- systems providing algorithmically guided dimensionality reduction, enabling user-control over the reduction process through various features;
- the application of algorithmically guided information visualization in the field of microbiology.

In the context of categorical data analysis, automated quantification approaches are commonly used. They provide efficient and effective quantification based on underlying data structures and enable analysis using methods designed for numerical data. Few of the methods presented prior to papers I, II and III of this thesis did, however, incorporate information on relationships within the numerical variables into the quantification process. Neither did they combine algorithmic and interactive methods to make use both of the speed of automated methods and the domain knowledge of an expert user. Through an evaluation of the two main approaches to categorical data visualization, an initial step was provided in comparing the two approaches and in providing guidance as to which approach may be the most useful within a specific task context. Paper IV contributes the first evaluation with this focus in the field of information visualization.

Dimensionality reduction is usually performed such that it aims to preserve one specific structure of interest, but for many analytical tasks several structures may be of interest. This thesis has contributed two methods for extracting overall measures of variable interestingness based on combinations of several quality metrics. The first method provides overall interestingness through a sum of quality metrics where weight values are used to control the relative importance of each metric. The second method utilizes a ranking approach to supply an overall measure of interestingness, not requiring knowledge regarding the relative importance of the set of quality metrics. These methods have been implemented in interactive visualization pipelines, providing algorithmically guided and user-controlled dimensionality reduction and exploration of high-dimensional data sets. The approach of exploratory and interactive dimensionality reduction was also applied to the field of microbiology through a system for visual exploration of microbial populations, contributing the combination of algorithmic methods, information visualization techniques and methods based on analytic tasks common in microbiology.

## 4.2 DISCUSSION

This thesis has primarily concerned the use of algorithmically guided information visualization in the areas of categorical data analysis and exploration of high-dimensional data sets. As previously discussed in the thesis, automated algorithmic analysis has many benefits in terms of efficiency and time consumption compared to manual data examination. This is particularly true in the context of high-dimensional data exploration since the amount of data gathered within various domains is constantly growing, with data sets including hundred or thousands of variables becoming increasingly common. Very few visualization methods, if any, are able to effectively display such high-dimensional data sets without aid from data mining approaches such as dimensionality reduction. Furthermore, data analysis may often be explorative, meaning that the analyst has no clear hypothesis of what to look for in the data. Algorithmic approaches may

facilitate the generation of hypotheses, by providing methods for automatic identification of potentially interesting structures which the analyst may be interested in examining further.

Automated methods are, on the other hand, often unable to take the prior knowledge and domain expertise of an analyst into account. The automation distances the exploration process from the user and may, as a result, limit the analyst's understanding of structures and relationships in the data. Through the combination of automated and interactive methods, the advantages of both algorithmic and manual exploration approaches may be made use of. Some of the benefits possible through employing algorithmically guided information visualization include:

- making use of the speed of automated methods when identifying patterns and relationships within data;
- aiding the analyst in focusing subsequent analysis through presentation of potentially interesting patterns;
- utilizing the domain expertise and task based knowledge of the analyst.

Through utilizing interactive methods, the analyst can be significantly involved in the analytical process and may, through the exploration process itself, gain important insights regarding patterns and relationships within the data, as well as of the processes driving these structures.

For categorical data analysis, as discussed in this thesis, the combination of algorithmic and interactive methods enables the use of methods designed for numerical data in such a way that the domain knowledge of the analyst is made use of. This may be beneficial in many analytical tasks, since methods designed for numerical data are more commonly available and often more generic than methods designed for categorical data. Furthermore their performance is often less dependent on data structures. Many methods have been suggested for visual analysis of categorical data. Their usability has not, however, been thoroughly established since few evaluations have been performed. Through formal comparisons, such as the evaluation presented in paper IV, their usability can be confirmed and, perhaps even more importantly, guidance can be provided to analysts regarding which methods may be most useful for a range of analytical tasks.

Considering high-dimensional data, the utilization of automated algorithms may often be necessary, since very few visualization methods are able to deal with truly high-dimensional data sets including hundreds of variables. However, when employing dimensionality reduction, the utilization of user influence may be beneficial for the analysis since users may be aware of variables of certain interest and importance due to the task context, which an automated method may not be able to identify. Moreover, the path of analysis may quite possibly change as identified patterns and relationships are presented to the analyst. Thus, the ability to interactively change the dimensionality reduction process may, in the end, provide a more useful analysis than purely automated methods.

## 4.3 FUTURE WORK

While new and more efficient techniques for the collection of data are developed the size and complexity of gathered data will continue to increase. Consequently, the requirement for new

and effective methods for data analysis will also continue to grow. Both information visualization and data mining provide methods for efficient and effective analysis of complex multivariate data and the combination of algorithms and visualization methods may often be beneficial. In the context of growing amounts of data this appears to continue to be an important research focus for the future.

In connection with the development of new visualization techniques it will be increasingly important to establish their usability. Although there has been a growing interest in evaluations within the visualization field during the last years, this is still an area requiring more effort from the visualization community, not least in terms of focusing on the usability of systems and techniques within a task and domain based context. Regarding the focus areas of this thesis, more studies are required to establish the benefit of algorithmic guidance in relation to the possible risk of making visualization systems more complex by introducing algorithms of which the user may not have a full understanding. Most importantly, methods for visualization of categorical data need to be evaluated. Paper IV puts focus on this by presenting the first study comparing methods representing the two main approaches to categorical data visualization. However, this study was only a first attempt at establishing the usability of the two approaches and more studies, comparing different methods within different task contexts, are required to establish usability and provide sufficient guidance as to when methods may be most appropriately used. Additionally, available visualization methods designed for categorical data need to be properly evaluated to fully appreciate their benefits and limitations.

Another interesting possible direction of future research is to design and develop methods for emphasizing the use, and possible distortion, of algorithmically modified data. This may include visual representations of uncertainty when employing a quantification method, where uncertainty can be defined as the variability in quantification result when employing different quantification algorithms or when modifying algorithm parameters. Along the same line, considering dimensionality reduction, visual representations of information loss may considerably facilitate decision making and apprehension of how the reduction remove information from the data. This has been addressed by Schreck et al. [72], using a single measure of information loss, but when aiming to preserve a variety of structures, multiple information loss metrics may need to be considered and require concurrent visual representation.

As information visualization systems and tools grow more complex, due to more complex data and analytical tasks, there is a considerable risk of interfaces becoming less intuitive. An interesting future research direction may, hence, be to extend the concept of guidance to also include visual guidance aiding the user in analysis by suggesting possible analytical paths and interactions with the data.

The scalability of visualization has been defined as one of the top research challenges in information visualization [17] and will probably continue to be so for quite some time since the amount of gathered data still continues to grow. Along the same line, the issues of high dimensionality will continue to require focus. The concept of algorithmically guided information visualization can be successfully applied within more or less any domain and for any task of data analysis, thus having a great potential for future research. More specifically, high-dimensional, categorical and mixed data sets are common within a wide range of domains, requiring usable and flexible methods for explorative analysis. One of these areas is bioinformatics, where a recently

growing interest has been shown in the visualization of biological data. Due to the recent technological advances in, for example, DNA-sequencing, the application of algorithmically guided information visualization in the area of bioinformatics is an interesting and promising future research direction.



# BIBLIOGRAPHY

- [1] J. Aitchison and M. Greenacre. Biplots of compositional data. *Applied Statistics*, 51(4):375–392, 2002.
- [2] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor. Improving the visual analysis of high-dimensional datasets using quality measures. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 19–26, 2010.
- [3] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering for an enhanced visualization of multidimensional data. In *Proceedings of IEEE Symposium on Information Visualization*, pages 52–60, 1998.
- [4] A. O. Artero, M. C. F. de Olivera, and H. Levkowitz. Enhanced high dimensional data visualization through dimension reduction and attribute arrangement. In *Proceedings of the conference on Information Visualization*, pages 707–712, 2006.
- [5] D. Asimov. The grand tour: A tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing*, 6(1):128–143, 1985.
- [6] M. Q. W. Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the Workshop on Advanced Visual Interfaces*, pages 110–119, 2000.
- [7] S. Barlowe, T. Zhang, Y. Liu, J. Yang, and D. Jacobs. Multivariate visual explanation for high dimensional datasets. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 147–154, 2008.
- [8] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.
- [9] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1966.
- [10] A. Beygelzimer, C.-S. Perng, and S. Ma. Fast ordering of large categorical datasets for better visualization. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 239–244, 2001.

- [11] S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of Siam International Conference on Data Mining*, pages 243–254, 2008.
- [12] S. Bremm, T. von Landesberger, J. Bernard, and T. Schreck. Assisted descriptor selection based on visual comparative data analysis. *Computer Graphics Forum*, 30(3):891–900, 2011.
- [13] A. Buja, J. A. McDonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. In *Proceedings of IEEE Visualization*, pages 156–153, 1991.
- [14] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.
- [15] V. Chandola, S. Boriah, and V. Kumar. A framework for exploring categorical data. In *Proceedings of the SIAM International Conference on Data Mining*, pages 187–198, 2009.
- [16] C. Chen. *Information Visualization: Beyond the Horizon*. Springer-Verlag, second edition, 2004.
- [17] C. Chen. Top 10 unsolved information visualization problems. *IEEE Computer Graphics and Applications*, 25(4):12–16, 2005.
- [18] J. Choo, H. Lee, J. Kihm, and H. Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 27–34, 2010.
- [19] C. Collins, G. Penn, and S. Carpendale. Bubble sets: Revealing set relations with isocontours over existing visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1009–1016, 2009.
- [20] T. F. Cox. *Introduction to Multivariate Analysis*. Hodder Arnold Publication, 2005.
- [21] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, second edition, 2001.
- [22] C. M. Cuadras, D. Cuadras, and M. J. Greenacre. A comparison of different methods for representing categorical data. *Communications in Statistics – Simulation and Computation*, 35:447–459, 2006.
- [23] S. G. Eick and A. F. Carr. Visual scalability. *Journal of Computational and Graphical Statistics*, 11(1):22–43, 2002.
- [24] D. Engel, R. Rosenbaum, B. Hamann, and H. Hagen. Structural decomposition trees. *Computer Graphics Forum*, 30(3):921–930, 2011.
- [25] U. Fayyad, G. G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, 2002.

- [26] B. J. Ferdosi and J. B. Roerdink. Visualizing high-dimensional structures by dimension ordering and filtering using subspace analysis. *Computer Graphics Forum*, 30(3):1121–1130, 2011.
- [27] A. Frank and A. Asuncion. UCI machine learning repository. <http://archive.ics.uci.edu/ml>, 2010.
- [28] W. Freiler, K. Matković, and H. Hauser. Interactive visual analysis of set-typed data. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1340–1347, 2008.
- [29] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23(9):881–890, 1974.
- [30] M. Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, 1994.
- [31] M. Friendly. Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3):373–395, 1999.
- [32] M. Friendly. *Visualizing Categorical Data*. SAS Institute Inc., 2000.
- [33] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, 1989.
- [34] G. H. Golub and W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *SIAM Journal on Numerical Analysis*, 2(B):205–224, 1965.
- [35] M. Graham and J. Kennedy. Using curves to enhance parallel coordinate visualisations. In *Proceedings of Seventh International Conference on Information Visualization*, pages 10–16, 2003.
- [36] M. Greenacre. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall, 2006.
- [37] M. Greenacre. *Correspondence Analysis in Practice*. Chapman & Hall, second edition, 2007.
- [38] G. Grinstein, C. Plaisant, S. Laskowski, T. O’Connel, J. Scholtz, and M. Whiting. VAST 2008 challenge: Introducing mini-challenges. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 195–196, 2008.
- [39] D. Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2(4):232–246, 2003.
- [40] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2001.
- [41] D. Hand, H. Mannilla, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

- [42] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 127–130, 2002.
- [43] S. L. Havre, A. Shah, C. Posse, and B.-J. Webb-Robertson. Diverse information integration and visualization. In *Proceedings of SPIE - The International Society for Optical Engineering*, pages 201–211, 2006.
- [44] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. DNA visual and analytic data mining. In *Proceedings of the 8th conference on Visualization*, pages 437–441, 1997.
- [45] M. L. Huang, T.-H. Huang, and J. Zhang. TreemapBar: Visualizing additional dimensions of data in bar chart. In *Proceedings of the 13th International Conference Information Visualization*, pages 98–103, 2009.
- [46] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for dimensional analysis and reduction. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 3–10, 2010.
- [47] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(4):69–91, 1985.
- [48] G. Ivosev, L. Burton, and R. Bonner. Dimensionality reduction and visualization in principal component analysis. *Analytical Chemistry*, 80(13):4933–4944, 2008.
- [49] D. H. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang. iPCA: An interactive system for PCA-based visual analytics. *Computer Graphics Forum*, 28(3):767–774, 2009.
- [50] J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing structure within clustered parallel coordinates displays. In *Proceedings of the 11th IEEE Symposium on Information Visualization*, pages 125–132, 2005.
- [51] B. Johnson and B. Shneiderman. Tree-maps: A space filling approach to the visualization of hierarchical information structures. In *Proceedings of IEEE Conference on Visualization 1991*, pages 284–291, 1991.
- [52] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, second edition, 2002.
- [53] E. Kandogan. 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*, pages 107–116, 2001.
- [54] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.
- [55] T. Kohonen. The self-organizing map. *Neurocomputing*, 21(1–3):1–6, 1998.

- [56] E. Kolatch and B. Weinstein. Cattrees: Dynamic visualization of categorical data using treemaps. [http://www.cs.umd.edu/class/spring2001/cmsc838b/Project/Kolatch\\_Weinstein/index.html](http://www.cs.umd.edu/class/spring2001/cmsc838b/Project/Kolatch_Weinstein/index.html), May 2001.
- [57] Y. Koren and L. Carmel. Robust linear dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 10(4):459–470, 2004.
- [58] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, 2006.
- [59] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978.
- [60] Z. Liu, J. Stasko, and T. Sullivan. SellTrend: Inter-attribute visual analysis of temporal transaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1025–1032, 2009.
- [61] S. Ma and J. L. Hellerstein. Ordering categorical data to improve visualization. In *Proceedings of IEEE Symposium on Information Visualization*, pages 15–18, 1999.
- [62] J. J. Miller and E. J. Wegman. Construction of line densities for parallel coordinate plots. *Computing and Graphics in Statistics*, pages 107–123, 1991.
- [63] B. Mirkin. *Clustering for Data Mining: A Data Recovery Approach*. Chapman & Hall, 2005.
- [64] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre. ClusterSculptor: A visual analytics tool for high-dimensional data. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 75–82, 2007.
- [65] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of IEEE Symposium on Information Visualization*, pages 89–96, 2004.
- [66] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pages 109–116, 2004.
- [67] C. R. Rao. A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. *Questiio*, 19(1–3):23–63, 1995.
- [68] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 318–322, 1994.
- [69] J. C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 61–71, 2007.

- [70] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang. Mapping nominal values to numbers for effective visualization. *Information Visualization*, 3(2):80–95, 2004.
- [71] M. Schonlau. Visualizing categorical data arising in the health sciences using Hammock plots. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, 2003.
- [72] T. Schreck, T. von Landesberger, and S. Bremm. Techniques for precision-based visual analysis of projected data. *Information Visualization*, 9:181–193, 2010.
- [73] J. Seo and H. Gordish-Dressman. Exploratory data analysis with categorical variables: An improved rank-by-feature framework and a case study. *International Journal of Human-Computer Interaction*, 23(3):287–314, 2007.
- [74] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proceedings of IEEE Symposium on Information Visualization 2004, INFOVIS 2004*, pages 65–72, 2004.
- [75] Z.-Y. Shen, J. Sun, Y.-D. Shen, and M. Li. R-map: Mapping categorical data for clustering and visualization based on reference sets. In *Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 992–998, 2008.
- [76] K. Shiraishi, K. Misue, and J. Tanaka. A tool for analyzing categorical data visually with granular representation. In *Proceedings of 13th International Conference on Human-Computer Interaction*, pages 342–351, 2009.
- [77] B. Shneiderman. Dynamic queries for visual information seeking. *IEEE Software*, 11(6):70–77, 1994.
- [78] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.
- [79] R. Spence. *Information Visualization: Design for Interaction*. Pearson Education, second edition, 2007.
- [80] J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human-Computer Studies*, 53(5):663–694, 2000.
- [81] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.
- [82] A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. Magnor, and D. Keim. Automated analytical methods to support visual exploration of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 17(5):584–597, 2011.

- [83] M. Tenenhaus and F. W. Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50(1):91–119, 1985.
- [84] H. Theisel. Higher order parallel coordinates. In *Proceedings of the 5th Fall Workshop on Vision, Modeling, and Visualization*, pages 415–420, 2000.
- [85] G. J. G. Upton. Cobweb diagrams for multiway contingency tables. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 49(1):79–85, 2000.
- [86] A. S. Vivacqua and A. C. B. Garcia. NRV: Using nested rings to interact with categorical data. In *Proceedings of the IADIS International Conference on Interfaces and Human Computer Interaction*, pages 85–92, 2008.
- [87] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, second edition, 2004.
- [88] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of American Statistics Association*, 85(411):664–675, 1990.
- [89] E. J. Wegman and Q. Luo. High dimensional clustering using parallel coordinates and the grand tour. *Computing Science and Statistics*, 28:352–360, 1997.
- [90] Z. Wen and M. X. Zhou. Evaluating the use of data transformation for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1309–1316, 2008.
- [91] E. O. Wiley and B. S. Lieberman. *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. John Wiley & Sons, Inc., second edition, 2005.
- [92] M. Williams and T. Munzner. Steerable, progressive multidimensional scaling. In *Proceedings of IEEE Symposium on Information Visualization*, pages 57–64, 2004.
- [93] J. Yang, A. Patro, S. Huang, N. Mehta, M. O. Ward, and E. A. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. In *Proceedings of IEEE Symposium on Information Visualization*, pages 73–80, 2004.
- [94] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proceedings of IEEE Symposium on Information Visualization*, pages 105–112, 2003.
- [95] J. Yang, M. O. Ward, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of Eurographics/IEEE TCVG Symposium on Visualization*, pages 19–28, 2003.
- [96] J. Yang, M. O. Ward, and E. A. Rundensteiner. InterRing: An interactive tool for visually navigating and manipulating hierarchical structures. In *Proceedings of IEEE Symposium on Information Visualization*, pages 77–84, 2002.

- [97] S. Yang, Z. Xiang, T. Daquan, and X. Weidong. A memory-saving and efficient data transformation technique for mixed data sets visualization. In *Proceedings of 13th International Conference Information Visualization*, pages 260–265, 2009.