Scalable analysis and visualization of high-dimensional astronomical data sets
Ferdosi, Bilkis Jamal

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2011

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 1

Introduction

“Twinkle twinkle little star. How I wonder what you are?” Childhood wonder and adulthood quest of “knowing thyself” make us broaden our eyes through powerful telescopes to get answers to all our unanswered questions. Nowadays, we are well equipped with tens of terabytes of data on millions of stars, galaxies, and other astronomical objects, thanks to astronomical sensors such as optical telescopes, radio telescopes, and space telescopes.

The adage “A picture is worth a thousand words” could not be more true than it is in astronomy. Spectacular images obtained with telescopes are a means to understanding the universe; see for example Figure 1.1. Therefore, extracting information from images is one of the fundamental tasks in astronomy. Astronomers need to identify the stars and galaxies, estimate their positions, photometry and other related attributes from the images to populate their catalogues. The process of populating catalogues from images is depicted in Figure 1.2. Telescopes, such as the VST (VLT Survey Telescope) operated by the European Southern Observatory (ESO), with high resolution CCD cameras (e.g., OmegaCAM) capture images of the sky. Those images later go through various image processing and pattern recognition steps, such as filtering, background estimation, masking, artifact detection, deblending/merging, photometry, and classification (Starck and Murtagh 2002). After these processing steps, one point in the image has become a record with hundreds of attributes in a table of a catalogue. A catalogue consists of a large collection of data and thus acts as a source of scientific information.

The work described in this thesis concerns the information retrieval step using catalogue data.
1.1 Illustrating the Problem

The seventh and final data release (DR7) of SDSS (Sloan Digital Sky Survey) covers 11,663 deg$^2$ of sky and contains 15.7 Tbyte of image data (in 5 bands: u, g, r, i and z, centered at a wavelengths 3551, 4686, 6165, 7481 and 8931 Å), and 44.8 Tbyte of catalogue data with 357 million objects, each with a few hundred attributes describing the object and its observations. Of these objects, spectra (and hence redshifts) are available for about 929,555 galaxies and about 400,000 stars (a full description of SDSS can be found on the SDSS website: www.sdss.org).
The earlier SDSS data releases have already led to a large number of results, varying from studies of the effect of the environment on the properties of galaxies (stellar mass, star formation activity, structure) (Kauffmann et al. 2004) to studies of the distribution and kinematics of stars in the Milky Way galaxy (Sirko et al. 2004b, Sirko et al. 2004a). The study of Kauffmann et al. (2004) shows the power of relating the properties of galaxies (in this case color, star formation activity and stellar mass) to their environment and spatial distribution. They obtained several interesting observations. For example, high density regions are predominantly populated by galaxies with a narrow range in color (0.8 < $g - r$ < 1.0; these are mostly elliptical galaxies) while the low-density areas are populated by galaxies with a wide range of colors (mostly star-forming disk galaxies). This information was obtained by observing scatter plots of $g - r$ color versus stellar mass at different environmental densities. They also observed in the spatial distribution of galaxies in a ‘slice’ at redshift $z = 0.05$ that galaxies with high star-formation rates appear to populate the densest areas, suggesting a link between local galaxy density and star formation activity.

The study of Kauffmann et al. (2004) has made use of known properties of galaxies, so that the multi-dimensional parameter space has been explored using this prior knowledge. For example: elliptical galaxies are known to populate preferentially the densest areas in the universe and obey a well defined color-magnitude (CM) relation (Visvanathan and Sandage 1977, Sandage and Visvanathan 1978b, Sandage and Visvanathan 1978a). But there may be other unexpected properties of galaxies which remain unnoticed as it is impossible to search manually all of parameter space, given that each galaxy in the SDSS catalogue has more than several tens of usable entries (shape in five color bands, position, redshift, spectral characteristics such as line strengths, etc.).

Another example to illustrate the need for designing algorithms that work effectively in multi-dimensional parameter spaces concerns research on our own Galaxy. In particular, a great deal of effort has gone into isolating the remnants of satellite galaxies accreted in the past, as predicted by modern galaxy formation theories, using the chemical composition, 3D spatial coordinates and velocities of stars in the Milky Way. Stars harbor unique clues of the assembly history of our Galaxy. Low-mass stars live for much longer than the present age of the universe, and so retain in their atmospheres a record of the chemical elements of the environment in which they were born (Christlieb et al. 2002, Venn et al. 2004).

The last decade has seen a large increase in the size and complexity of data sets containing the relevant physical information required to disentangle the history of the Galaxy. The SDSS survey has played a key role in this respect, with the discovery of streams associated to the most recently accreted satellite, the Sgr dwarf. This discovery (as well as that of an “outer ring”) was made from the distribution of the stars on the sky (a 2D projection of a 6D space). The recent
nature of this event made the structure obvious on the sky. However, many more substructures are predicted by models, and these ought to be buried in the SDSS and similar databases (e.g., RAVE, Steinmetz (2003)).

To explore properly these multi-dimensional (> 10) data volumes, new tools need to be developed to localize and characterize the coherence in parameter space and then to visualize the structures found. Coherence in the data is of course not limited to a few dimensions: finding multidimensional and multiscale patterns must be the focus of new research in this area.

1.2 Goal of the Thesis

In the previous subsection, we discussed the problem of extracting information from large high-dimensional astronomical datasets. Trying to find relations with a priori knowledge may hinder finding unknown relations that are hidden in the haystack of large datasets. Visualization can be an invaluable tool in such cases that utilizes the human power of identifying structures and relations. However, in that case it may be required to observe a huge number of visualizations which is not feasible. Therefore coupling automation with visualization is needed. Hence, the goal of this thesis is to provide new methods/algorithms which:

- are capable of coping with enormous amounts (terabytes) of multi-dimensional parameter data in an automated fashion, thus should be scalable to ensure that processing methods remain usable on data sets which keep growing in size;
- offer automated approaches for analyzing and in particular visualizing structures, patterns etc., in multi-dimensional \( N > 3 \) parameter space;
- are flexible and allow the observer to participate in the analysis by using interactive visualization combined with the human processing/perceptive/analytical power.

In the following section we describe state-of-the-art high-dimensional data visualization techniques, indicating their strengths, limitations and directions of improvement to satisfy our goal of visualizing high-dimensional astronomical data.

1.3 High-Dimensional Data Visualization

Visualization is:

"to form a mental vision, image, or picture of (something not visible or present to sight, or of an abstraction); to make visible to the mind or imagination"


"more than a method of computing. Visualization is the process of transforming information into a visual form, enabling users to observe the information. The resulting visual display enables the scientist or engineer to perceive visually features
which are hidden in the data but nevertheless are needed for data exploration and analysis”

– Gershon (1994)

While the goal of visualization is to enhance understanding, gaining insight, exploration and exploitation and so on, the data to be visualized often come with very high dimensionality. Visualizing such data is not a trivial task because human perception is accustomed to a three-dimensional world. There are several approaches for visualizing high-dimensional data.

Here, these techniques will be discussed according to three major categories:

1. methods that use a Cartesian coordinate system for the first three chosen dimensions, and use colors, glyphs, size and shape attributes for any additional dimension;

2. methods that use a Cartesian coordinate system after applying dimension reduction techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Sammon Algorithm, multidimensional scaling (MDS), projection pursuit, or any other form of transformation of the high-dimensional data.

3. methods that visualize every dimension with equal treatment and without any transformation.

1.3.1 Volume Visualization

Spaces of dimensionality three or higher can be visualized with volume visualization. However, spaces with dimensionality higher than three need to be reduced to 3D space using dimensionality reduction techniques, before they can be visualized with volume visualization. This technique normally is used for visualizing continuous data. Nevertheless, discrete data can be visualized with the same technique as well. In that case, a continuous form of the discrete data such as a density field calculated on a grid can be used.

There are two main groups of algorithms to visualize data volumes:

◆ visualization with surface fitting – extracts a surface and then visualizes it,

◆ direct volume visualization – visualizes directly, without extracting a surface, using structured or unstructured grids.

Visualization with Surface Fitting

Surface fitting, also known as isosurface visualization, visualizes the surface that corresponds to points with a particular data value (known as the isovalue) in the volume. This is basically a 3D version of the 2D contouring technique. A widely used technique for visualizing isosurfaces is the marching cubes algorithm (Lorensen and Cline 1987), where points of interest are tessellated with triangles in a regular grid by traversing every cubic cell. In information visualization this technique can be very useful in visualizing structures (such as clusters) of varying
1.3 High-Dimensional Data Visualization

Figure 1.4. Visualization of the olive oil dataset (Forina et al. 1983). (Left) Isosurface visualization of three of the dimensions (palmitic, palmitoleic, linoleic). At isovalue 190 the presence of eight clusters can be observed. (Right) X-ray volume visualization of the same three dimensions.

density, present in a dataset. In the left of Figure 1.4, density image of the three of the dimensions (palmitic, palmitoleic, linoleic) of the olive oil dataset (Forina et al. 1983) is visualized by isosurface visualization with a user-defined isovalue where eight dense clusters become easily visible.

Direct Volume Visualization (DVV)

In surface visualization a particular set of data points (corresponding to a chosen isovalue) is used, whereas in DVV all the data points in the volume can contribute in the visualization process. DVV creates a direct mapping of the volume data onto the image plane where each voxel (“volume element”, analogous to a pixel in a 2D image) is represented by a small cube. The contribution of each cube depends on its distance from the image plane. There are several approaches in DVV, such as X-ray, Maximum Intensity Projection (MIP) volume visualization etc. In X-ray volume visualization, data values along the line of sight are integrated (averaged) whereas in Maximum Intensity Projection volume visualization the maximum data value along the line is used. In information visualization this technique can be used to obtain an overview of the structures present in data space. In the right of Figure 1.4, an X-ray volume visualization of the 3D density field of three of the dimensions (palmitic, palmitoleic, linoleic) of the olive oil dataset is shown.

1.3.2 Scatter Plot Matrix (SPM)

The scatter plot (Chambers et al. 1983) is a very popular way to represent the relationship between two variables. The scatter plot matrix (SPM) is an extension to higher dimensions and shows the pair-wise relationships of a set of variables in a matrix format. The SPM of a
$N$-dimensional dataset comprises $N$ rows and $N$ columns, where the element in row $i$ th and column $j$ th of the ($N \times N$) matrix contains the scatter plot of dimension $i$ versus dimension $j$ of the dataset. In Figure 1.5 the scatter plot matrix of the olive oil data set is visualized.

SPM is very useful to obtain an overview of pair-wise relationships of the dimensions. However, for very high ($\gg 3$) dimensional datasets SPM becomes unreadable due to crowding and visual clutter. Interactive SPM with zooming and panning features can solve the problem for noise-free datasets.Datasets with noise require filtering of the noise and less informative dimensions to make SPM useful and interpretable (Peng et al. 2004).

### 1.3.3 Parallel Coordinate Plot (PCP)

Parallel coordinate plots (Inselberg 2009, Wegman 1990) are constructed by laying out the axes of an $N$-dimensional data set in parallel as opposed to the more familiar orthogonal arrangement of the Cartesian coordinate system. The number of axes in PCP in principle equal the number of dimensions present, and is only limited by the horizontal resolution of the screen. Instances in the data set are represented by a line trace, connecting the case values on each dimension axis, as shown in Figure 1.6. However, PCP suffers from a number of shortcomings. If the number of instances is large, overlapping lines cause clutter and thus hide the structures of interest (see the left of Figure 1.7, created with GGobi\(^1\), a high-dimensional data visualization tool). It also requires reordering of the dimensions to observe certain structures, correlations, etc. To reduce the clutter Fua et al. (1999) proposed hierarchical PCP. They used hierarchical clustering to obtain a multi-resolution cluster display. Variable-width opacity bands were applied to visualize the clusters, where the lowest opacity represents the cluster mean. Depending on the spreading of the clusters in each dimension the opacity of the band dissolves gradually to the edge (see the middle of Figure 1.7, created with XmdvTool\(^2\)). They also provided an interface (as shown in the right of Figure 1.7) to allow structure-based brushing which facilitates the visualization of structures. However, because of the method’s dependency on full-dimensional clustering methods, it may not be possible to recover high-dimensional structures present in the subspaces.

Blaas et al. (2008) proposed a method based on the joint histogram of each dimension pair. Instead of drawing a line for each data point, histogram bins are used to draw the primitives. A more detailed description of the method can be found in Chapter 3. The same dataset as in Figure 1.7 visualized with the method of Blaas et al. can be seen in Figure 1.8. Now it is

\(^1\)http://www.ggobi.org/
\(^2\)http://davis.wpi.edu/xmdv/
1.3 High-Dimensional Data Visualization

**Figure 1.6.** Basics of the parallel coordinate plot.

**Figure 1.7.** (Left) PCP for a dataset with 5500 data points, created with GGobi. (Middle) Hierarchical PCP created with XmdvTool. (Right) Interface for structure-based brushing of XmdvTool.

**Figure 1.8.** PCP of the same dataset as in Figure 1.7, visualized with the method of Blaas et al.
possible to see the structures more clearly. Nevertheless, it is still not possible to perceive the high-dimensional clusters present. To visualize such structures it is necessary to reorder the dimensions properly. There are several methods for this purpose known in the literature (Ankerst et al. 1998, Peng et al. 2004, Yang et al. 2003). A detailed description of these methods can be found in Chapter 3.

1.3.4 Radial visualization (RadViz)

The RadViz technique for high-dimensional data visualization was first proposed by Hoffman et al. (1997). Similar to PCP, it is capable of visualizing a large number of dimensions at a time. However, unlike PCP, data points in RadViz are visualized by means of a non-linear transformation using a force-based model. The dimensions of the dataset are visualized along the perimeter of a circle. Each \( N \)-dimensional data point is assumed to be connected with each dimension axis by a spring, where the spring constant of each spring is equal to the value of the data point in the respective dimension. The data point is visualized at a position inside the circle where the spring forces are null (see Figure 1.9 for an example of RadViz obtained by the GeoViz Toolkit\(^3\)).

RadViz is capable of visualizing multidimensional clusters. However, it can lose details of local structures. Data points with quite different values can get the same position in the display and end up as clumps of points. In addition, the visualization of RadViz depends heavily on the ordering of the dimensions. Therefore, it also requires methods to find the optimal dimension ordering for visualizing the structures.

1.3.5 Tours

Tours is an animation method of low-dimensional projections of high-dimensional data obtained with linear transformations, where the sequence of projections is created by varying the projection matrix (Asimov 1985, Buja and Asimov 1986, Cook and Swayne 2007). Projections in the animation can be chosen in three different ways: grand tour, projection pursuit guided tour, manual manipulation. In grand tour mode, projections are selected randomly from the space of all possible projections. To have a smooth transition between two selected projections, intermediate projections are created along a geodesic path. The projection pursuit guided tour (Cook et al. 1995) selects the projections by optimizing a user-specified function intended to find some patterns of structures in the data. Manual manipulation of tours is obtained by user interaction such as mouse movement that changes the projection matrix of a constrained dimension. Tours

\(^3\)http://www.geovista.psu.edu/geoviztoolkit/
produce nice visual effects of the low-dimensional projections of high-dimensional data. Nevertheless, it is not always easy to interpret the interesting projections found in tours due to the fact that a transformation of the original data takes place.

In this subsection, we discussed some of the high-dimensional data visualization techniques that can facilitate understanding of the methods described in later chapters of the thesis. However, there exists several other high-dimensional data visualization techniques, such as Dimension Stacking, Worlds within Worlds, Circle Segment, Andrews Curve, Table Lens, Chernoff Faces, query-dependent Spiral and Axes Techniques, etc. Brief descriptions of these methods can be found in different surveys of high-dimensional data visualization techniques (Hoffman and Grinstein 2002, Oliveira and Levkowitz 2003).

1.3.6 Interaction

While visualization concerns the formation of images from the data that helps the user to derive insights, interaction is a tool to supplement visualization in this endeavor. Most of the time, visualization and interaction go hand in hand. Useful and easy-to-use interaction methods add value to the visualization. Several interaction techniques exist. Most of these techniques are inspired by the visual information seeking mantra:

*Overview first, zoom and filter, then details-on-demand*  
— Shneiderman (1996)

A detailed survey of interaction taxonomy and techniques can be found in Kosara et al. (2003). These techniques are in an information visualization context. However, they are applicable to any kind of visualization. Here, we will discuss four of such techniques.

**Focus+Context** is one of the most common interaction techniques. It enables visualization of a large amount of data in a relatively small screen space. It can emphasize regions of interest without losing the overall context. There are several methods that can achieve such interaction, such as distortion oriented fish-eye views (Furnas 1986), the overview-based information mural (Jerding and Stasko 1998), the filtering-based magic lens (Bier et al. 1994) etc.

**Brushing and Linking** is an interaction technique that becomes useful when both of these tasks, i.e., brushing and linking of the multiple data view, are done together. Brushing is the process where the user selects data points using the mouse or direct touch (in touch sensitive displays). Brushed (selected) points can then be investigated using multiple views.

**Interaction with Dimensions** such as rotation around the x, y, or z axis, provides a new view of the data points without direct interaction with these points.

**Rearranging Dimensions** is specifically useful for visualizations such as the parallel coordinate plot, where reordering of the dimensions can reveal new insights of the data. It can be done with drag and drop operations using the mouse or direct touch.
1.4 Astronomical Tools in Use

There are several tools like TOPCAT, Aladin, VisIVO, VOSpec, Tipsy, VisIt, etc. that are used in astronomy. Most of these tools are built to fulfill the specialized needs of different astronomical applications. There are tools to handle catalogue and tabular data (e.g., TOPCAT), image data (e.g., Aladin), simulation data (e.g., VisIt, Tipsy), 3D visualization (e.g., VisIVO), or visualizing spectral data (e.g., VOSpec).

Tool for OPerations on Catalogues And Tables (TOPCAT) was developed within the UK-based Starlink and AstroGrid project. It offers fast access, viewing and editing of large tabular datasets. Tables can be viewed in a scrollable browser where other manipulation such as reordering, hiding existing columns, adding new synthetic columns or making subsets of rows is possible. Rows in multiple tables can be concatenated based on astronomical features of the objects.

The tool also provides visualization of the data through histograms, 2D/3D scatter plots, 3D spherical polar plots, stacked line plots, and 2D density maps (histograms on a 2-dimensional grid). Along with the visualization it also allows interaction with the plots, such as visual selection of rows. A snapshot of TOPCAT’s user interface along with its table browser and 4D scatter plot is shown in Figure 1.10.

TOPCAT serves the astronomers quite well in the way its name indicates. However, although it provides powerful table viewing and editing features, its visualization can only fulfill the requirement of 1D/2D/3D and at most 4D (using color axes) visualization requirements. Dimensions higher than four can not be visualized using TOPCAT. In addition, if the number of columns is very high, the number of plots (1D/2D/3D) to observe can be quite high as well. TOPCAT provides no guidance to explore such large number of subspaces and the user needs to explore these manually.

Aladin is an interactive sky atlas developed by the Centre de Données astronomiques de Strasbourg (CDS). It facilitates the visualization of part of the sky extracted from an image database (Bonnarel et al. 2000). This tool basically is used for identifying and cross-validating astronomical sources (stars, galaxies, etc.). It allows the creation of scatter plots and contour plots using catalogue data which are overlayed on the image of interest. In Figure 1.11 a snapshot of Aladin is shown.

Visualization Interface to the Virtual Observatory (VisIVO) is a tool for visualizing and analyzing astrophysical data from virtual observatory (Becciani et al. 2006). It offers 2D/3D scatter plots for point data, volume and isosurface visualization, and orthogonal slice visualization for mesh-based data (see Figure 1.12). It enables higher dimensional (> 3) visualization through colors, glyphs and their size and shape attributes. It also provides a set of operations such as

---

4The Starlink Project was a long running UK project supporting astronomical data processing; http://starlink.jach.hawaii.edu/starlink.
5Virtual observatory software for astronomers; http://www.astrogrid.org/.
6http://aladin.u-strasbg.fr/
7http://visivo.oact.inaf.it/
1.4 Astronomical Tools in Use

Figure 1.10. Interface of TOPCAT with table viewer and 3D scatter plot.

Figure 1.11. Interface of Aladin
Figure 1.12. Interface of VisIVO with 2D/3D scatter plot, volume rendering and orthogonal slice visualization.

Figure 1.13. Interface of VisIt with volume rendering, vector visualization, 3D scatter plot and parallel coordinate plot.
correlation filtering, random subset generation, scalar histograms, mathematical operations, etc., for data analysis and modification.

VisIVO produces nice 2D and 3D visualizations. However, its higher dimensional visualization through color, glyphs, size and shape attributes in 3D Cartesian coordinates is not so useful for very large and high-dimensional datasets, as usage of such techniques produces visual crowding and clutter for large datasets, thus prohibiting perception of information. VisIVO does not provide any guidance to explore very high-dimensional datasets where the number of plots to observe can be quite large, especially when the task is to discover new or unexpected phenomena.

VisIt\(^8\) is a visualization tool designed to accommodate very large datasets through distributed and parallel processing. Because of its distributed architecture it is capable of visualizing simulation data in the place where generated without moving the data to a visualization server. Astronomers who work with such data often use the tool. It offers several visualization techniques such as scatter plots, contour plots, pseudo color plots, volume plots, vector plots, parallel coordinate plots (PCP), etc. It also provides a set of operators such as cone, clip, onion peel, and threshold. The cone operator slices 3D data with a cone, creating a surface in the shape of a cone. The clip operator can clip box- or sphere-shaped regions from the dataset before the dataset is plotted. The onion peel operator creates layers starting from a seed. At first only the seed cell is visualized. When the user increases the number of layers, cells that are connected with the seed cell are visualized. The threshold operator extracts and

VisIt provides a quite large set of options for visualization. However, the interface of VisIt is not very intuitive. The choice of the variables to be plotted has to be made entirely by the user. For example, to visualize data with parallel coordinates the user needs to choose the variables and their position in the plot. In the case of very high-dimensional data complete manual exploration is time consuming, tedious and may also prohibit extraction of unexpected phenomena. Another difficulty in the case of the parallel coordinate plot is to find a suitable ordering of the axes. Without a proper ordering of the axes it is very difficult to perceive high-dimensional structures using PCP. visualizes the cells within a specified range of values.

VOSpec\(^9\) is a virtual observatory spectral analysis tool. It provides several analysis tools such as measures of central tendency, dispersion, wavelet analysis, tuning of spectra and fitting utilities such as TSAP (Theoretical Spectral Access Protocol), best fit, polynomial, Gaussian, etc. It provides simple visualizations to support these operations on spectra.

Besides these tools, astronomers also use high-level programming languages such as IDL\(^10\), MATLAB\(^11\), Python\(^12\) etc., to produce custom visualizations of the data.

---

\(^8\) Developed by the US Department of Energy (DOE) Advanced Simulation and Computing Initiative (ASCI); https://wci.llnl.gov/codes/visit/

\(^9\) Developed by ESA (European Space Agency) VO (Virtual Observatory) team; http://esavo.esac.esa.int/vospec/

\(^10\) http://www.ittvis.com/

\(^11\) http://www.mathworks.com/products/matlab/

\(^12\) http://www.python.org/
1.5 Thesis Contribution and Organization

This thesis describes methods and algorithms for high-dimensional astronomical data exploration and visualization. In Chapter 2 (Ferdosi et al. 2011) we study the performance of four density estimation techniques: k-nearest neighbors (kNN), adaptive Gaussian kernel density estimation (DEDICA), a special case of adaptive Epanechnikov kernel density estimation (MBE), and the Delaunay tessellation field estimator (DTFE). The adaptive kernel based methods, especially MBE, perform better than the other methods in terms of calculating the density properly and have stronger predictive power in astronomical use cases. Moreover, the computation time of these methods is lower than for other methods and it computes the density on grids that can facilitate visualization (as an image in 2D and a volume in 3D) and analysis of the data. Using the MBE method we can also achieve scalability with respect to the number of data points. After the original feature space has been transformed into image space, further computation can be done in image space that has constant size, although the size of the dataset can grow very large.

The next step is to extract useful information from such spaces. Clustering is one of the techniques that can help discovering structures in a dataset. However, full-dimensional clustering is not so useful since structures may exist in different subspaces that may indicate different relations among particular subsets of dimensions. Subspace clustering is an approach that can be applied for this purpose. Subspace clustering is the process of finding clusters in subspaces of the full feature space, either directly (Agrawal et al. 1998) or by identifying relevant subspaces for
(later) clustering based on some quality criteria (Baumgartner et al. 2004). In Chapter 3 (Ferdosi et al. 2010) we propose an interactive approach to find relevant subspaces which is strongly tied to morphological properties of object distributions. We used connected morphological operators implemented using the Max-tree data structure to identify the clusters (high-intensity regions in the density image). A “quality” of the subspaces is defined depending on their clustering property. We recover various known relations directly from the data with little or no a priori assumptions. Therefore, our method can act as a starting point in analyzing large datasets and help users to find new relations as well.

Using the method described in Chapter 3 we can obtain interesting subspaces of any dimension. However, visualizing high-dimensional structures in a meaningful and user-interpretable way is far from straightforward. Traditionally, low-dimensional representations of high-dimensional spaces, obtained by methods such as Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), etc., are used to perform visualization in a Cartesian coordinate system. However, they pose the problem of interpretation of the visualization, because of the transformation of the original feature space to a new coordinate system. Two widely used methods to visualize high-dimensional data without transformation are the scatter plot matrix (SPM) and the parallel coordinate plot (PCP). For effective visualization of high-dimensional structures, they also require a proper ordering of the dimensions. In Chapter 4 (Ferdosi and Roerdink 2011) we propose algorithms for reordering dimensions in PCP and SPM in such a way that high-dimensional structures (if present) become easier to perceive. We use the quality criterion and the cluster indication capability of the method described in Chapter 3 to present three algorithms: two for finding suitable dimension ordering for PCP and one for SPM.

In Chapter 5, we discuss several design issues of a visual analytic tool for astronomical data using a large touch sensitive display and present a prototype for such a tool. Large touch-displays provide more screen space, support more intuitive and natural interactions with touch sensitive inputs, and can make sharing of the analysis process and collaboration with others possible. Thus, they can facilitate analysis of astronomical data which are not only large in size and dimension but also complex in nature.

Finally, in Chapter 6 we provide a summary, draw conclusions, and present an outlook for future work.